

Analisis Histori Pembelian Laptop untuk Menentukan Faktor Utama Pembelian Menggunakan Regresi Logistik

<http://dx.doi.org/10.28932/jutisi.v10i2.8897>

Riwayat Artikel

Received: 21 Mei 2024 | Final Revision: 10 Agustus 2024 | Accepted: 10 Agustus 2024

Creative Commons License 4.0 (CC BY – NC)



Feliks Victor Parningotan Samosir[✉]#1, Hery^{#2}

^{#1} Informatika, Universitas Pelita Harapan
MH Thamrin Boulevard 1100 Kelapa Dua, Tangerang, 15811, Indonesia

¹ feliks.parningotan@uph.edu

^{#2} Sistem Informasi, Universitas Pelita Harapan
MH Thamrin Boulevard 1100 Kelapa Dua, Tangerang, 15811, Indonesia

² hery.fik@uph.edu

[✉]Corresponding author: feliks.parningotan@uph.edu

Abstrak — E-Commerce merupakan konsep baru yang biasa digambarkan sebagai proses jual beli barang atau jasa melalui World Wide Web atau pertukaran produk, jasa, dan informasi melalui jaringan informasi, termasuk internet. Tokopedia, salah satu platform e-commerce terbesar di Indonesia, memasarkan banyak produk, termasuk barang elektronik. Studi ini menganalisis data laptop dari Tokopedia untuk menentukan faktor yang mempengaruhi keputusan pembelian menggunakan model regresi logistik. Data pembelian dan spesifikasi laptop dianalisis untuk menentukan fitur-fitur signifikan yang mempengaruhi keputusan pembelian. Metode yang digunakan meliputi seleksi fitur dengan ID3, Random Forest, dan K-Means, dilanjutkan dengan regresi logistik untuk mengevaluasi pengaruh masing-masing fitur. Hasil regresi logistik pada fitur `price_cat`, `memory_size`, dan `merk_cat` menunjukkan bahwa ketiganya dapat digunakan untuk membuat persamaan regresi logistik guna menentukan probabilitas pembelian laptop. Namun, hasil perhitungan regresi menunjukkan bahwa nilai r^2 untuk `price_cat` adalah yang tertinggi, yaitu 0,82, yang menunjukkan bahwa `price_cat` memiliki pengaruh signifikan dalam menentukan probabilitas pembelian laptop.

Kata kunci— e-commerce; laptop; regresi logistik; spesifikasi; tokopedia

Analysis of Laptop Purchase History to Determine Key Factors Using Logistic Regression

Abstract — E-Commerce is a new concept typically described as the process of buying and selling goods or services through the World Wide Web or exchanging products, services, and information through information networks, including the internet. Tokopedia, one of the largest e-commerce platforms in Indonesia, markets many products, including electronic goods. This study analyzes laptop data from Tokopedia to determine the factors influencing purchase decisions using a logistic regression model. Purchase history and laptop specifications data are analyzed to identify significant features affecting purchase decisions. The methods used include feature selection with ID3, Random Forest, and K-Means, followed by logistic regression to evaluate the impact of each feature. The logistic regression results on the features `price_cat`, `memory_size`, and `merk_cat` show that all three can be used to create a logistic regression equation to determine the probability of laptop purchases. However, the regression results indicate that the r^2 value for `price_cat` is the highest, at 0.82, indicating that `price_cat` has a significant influence in determining the probability of laptop purchases.

Keywords— e-commerce; laptop; logistic regression; specifications; tokopedia

I. PENDAHULUAN

Perkembangan teknologi informasi telah membawa perubahan signifikan dalam berbagai aspek kehidupan, termasuk dalam bidang perdagangan. E-Commerce, sebagai salah satu hasil dari perkembangan tersebut, telah menjadi fenomena global yang mengubah cara manusia melakukan jual beli [1]. Di Indonesia, Tokopedia merupakan salah satu platform e-commerce terbesar di Indonesia [2] yang menyediakan berbagai macam produk, termasuk barang elektronik seperti laptop.

Laptop menjadi salah satu barang elektronik yang banyak diminati karena kebutuhan akan perangkat komputasi yang fleksibel dan *portable*. Dengan semakin banyaknya pilihan laptop yang tersedia, konsumen seringkali dihadapkan pada tantangan dalam menentukan pilihan yang tepat. Oleh karena itu, memahami faktor-faktor yang mempengaruhi keputusan pembelian laptop menjadi penting, baik bagi konsumen untuk membuat keputusan yang lebih baik maupun bagi penjual untuk mengembangkan strategi pemasaran yang efektif [3].

Data riwayat pembelian dan data customer yang ada dalam database Tokopedia sangat banyak, memberikan berbagai cara untuk mengolah data tersebut. Mahardini dkk [2] dalam penelitiannya menganalisis pengaruh kualitas produk, harga, dan kualitas layanan terhadap keputusan pembelian di Tokopedia wilayah DKI Jakarta menggunakan pendekatan korelasional dan metode SEM-PLS versi 3.0 dengan 100 responden yang dipilih secara random sampling. Hasil penelitian menunjukkan bahwa kualitas produk dan kualitas layanan berpengaruh signifikan terhadap keputusan pembelian, sementara harga tidak berpengaruh signifikan. Responden merasa sangat puas dengan kualitas produk dan layanan Tokopedia, tetapi kurang puas dengan harga yang dianggap mahal. Penelitian ini memiliki keterbatasan pada jumlah variabel independen dan jumlah responden, serta kejujuran dan pemahaman responden dalam mengisi kuesioner.

Dalam penelitian yang dilakukan oleh Himmah dkk [3], regresi logistik multinomial digunakan untuk menganalisis pengaruh strategi bauran pemasaran (produk, harga, tempat, dan promosi) terhadap pemilihan merek laptop oleh mahasiswa Universitas Diponegoro. Populasi penelitian ini adalah semua mahasiswa Universitas Diponegoro yang memiliki laptop, dengan jumlah responden sebanyak 150 orang yang dipilih menggunakan teknik *purposive sampling* dan *accidental sampling*, mengingat kerangka sampel yang tidak tersedia. Data dikumpulkan melalui pengisian kuesioner dari mahasiswa yang kebetulan berada di area kampus. Variabel dependen (merek laptop) dikodekan menjadi kategori numerik, sementara variabel independen mencakup elemen-elemen bauran pemasaran. Model regresi logistik multinomial mengestimasi probabilitas pemilihan merek berdasarkan nilai variabel independen. Hasil analisis menunjukkan koefisien regresi yang mengindikasikan pengaruh signifikan dari setiap elemen bauran pemasaran terhadap probabilitas pemilihan merek tertentu. Estimasi probabilitas terbesar untuk merek Acer terjadi pada kelompok dengan produk sedang, harga tinggi, dan promosi tinggi yaitu sebesar 77.461%. Estimasi probabilitas terbesar untuk merek Toshiba terjadi pada kelompok dengan produk tinggi, harga tinggi, dan promosi sedang yaitu sebesar 49.239%. Estimasi probabilitas terbesar untuk merek HP terjadi pada kelompok dengan produk sedang, harga sedang, dan promosi sedang yaitu sebesar 46.074%. Estimasi probabilitas terbesar untuk merek Axioo-Dell terjadi pada kelompok dengan produk sedang, harga sedang, dan promosi sedang yaitu sebesar 14.764%. Estimasi probabilitas terbesar untuk merek lain terjadi pada kelompok dengan produk sedang, harga tinggi, dan promosi sedang yaitu sebesar 22.134%. Kesimpulan dari penelitian ini memberikan wawasan penting bagi produsen laptop untuk merumuskan strategi pemasaran yang lebih efektif untuk segmen pasar mahasiswa. Penelitian lainnya oleh Sunita [4] mengeksplorasi pengaruh kecerdasan buatan (AI) dalam keputusan pembelian di e-commerce melalui analisis regresi logistik. Studi ini menganalisis data dari 120 responden di Chennai untuk menyelidiki pengaruh rekomendasi produk, chatbot dan asisten virtual, ulasan dan penilaian pelanggan, opsi kustomisasi produk, serta promosi dan diskon. Hasil penelitian menunjukkan bahwa rekomendasi produk, chatbot dan asisten virtual, serta promosi dan diskon memiliki pengaruh positif yang signifikan terhadap keputusan pembelian, menyoroti pentingnya rekomendasi yang dipersonalisasi, bantuan pelanggan yang efisien, dan kampanye promosi yang baik. Namun, ulasan dan penilaian pelanggan serta opsi kustomisasi produk tidak menunjukkan dampak signifikan pada keputusan pembelian. Temuan ini memberikan implikasi praktis bagi bisnis yang ingin meningkatkan keterlibatan pelanggan dan meningkatkan penjualan. Dengan menerapkan strategi efektif terkait faktor-faktor yang berpengaruh, organisasi dapat meningkatkan kepuasan pelanggan dan meningkatkan tingkat konversi di pasar e-commerce yang kompetitif.

ID3 (Iterative Dichotomiser 3) adalah algoritma untuk membuat pohon keputusan berdasarkan entropi dan gain informasi, membagi data secara rekursif hingga semua data terklasifikasi dengan baik [5]. Penelitian Kissinger Sunday et al. [6] di Usmanu Danfodiyo University menggunakan algoritma ID3 dan J48 untuk menganalisis data kinerja mahasiswa dalam pendidikan pemrograman, menunjukkan bahwa J48 lebih akurat. Random Forest (RF) adalah algoritma ensemble yang menggunakan bootstrap aggregation dan random subspace untuk klasifikasi [7] dan regresi [8] [9]. Penelitian Sudev G et al. [10] menggunakan dataset Black Friday Sales untuk membangun model prediksi berbasis regresi, menemukan bahwa RF unggul dengan kinerja terbaik. Lilhore et al. [11] mengembangkan Hybrid Weighted RF (HWRf) untuk prediksi dan klasifikasi perilaku pembelian online, menunjukkan efisiensi RF dalam mengklasifikasikan faktor perilaku konsumen. Prasetyo et al. [12] mengeksplorasi seleksi fitur untuk mengurangi jumlah fitur dalam menentukan centroid awal algoritma K-Means [13] yang dioptimalkan dengan Dynamic Artificial Chromosome Genetic Algorithm (DAC GA), menemukan bahwa seleksi fitur mempercepat waktu komputasi tetapi mengurangi kualitas hasil klusterisasi.

Penelitian ini bertujuan untuk mengidentifikasi spesifikasi laptop yang paling signifikan mempengaruhi keputusan pembelian di Tokopedia. Fitur-fitur yang dianalisis meliputi merk, ukuran memori, prosesor, ukuran disk, harga, dan jumlah terjualnya. Pendekatan utama yang digunakan adalah regresi logistik, yang menghubungkan variabel independen dengan variabel dependen kategori [4]. Regresi logistik digunakan untuk mengidentifikasi faktor-faktor yang mempengaruhi keputusan pembelian, memprediksi perilaku konsumen, dan memberikan rekomendasi produk yang lebih akurat dalam konteks e-commerce.

Sebelum melakukan regresi logistik, dilakukan seleksi fitur untuk mengidentifikasi fitur paling signifikan sambil mengeliminasi fitur yang tidak relevan, *redundant*, dan mengandung *noise* [7]. Seleksi fitur meningkatkan interpretabilitas model, mengurangi waktu pembelajaran, meningkatkan akurasi, dan mengatasi masalah dimensionalitas tinggi [14]. Metode seleksi fitur yang digunakan meliputi klasifikasi pohon keputusan ID3, Random Forest (RF), dan pengelompokan K-Means.

Eksperimen klasifikasi menggunakan pohon keputusan ID3 dan Random Forest untuk mengevaluasi performa model berdasarkan akurasi klasifikasi. Hasil klasifikasi membantu mengidentifikasi fitur-fitur paling berpengaruh untuk digunakan dalam model regresi logistik. Untuk pengelompokan data dengan K-Means, fitur-fitur dengan nilai korelasi diatas 0.5 digunakan. Pengelompokan ini membantu menemukan spesifikasi laptop yang paling banyak dibeli dan menentukan dua klaster dengan nilai pembelian terbanyak.

Regresi logistik kemudian digunakan untuk melihat nilai r^2 dari setiap fitur, dengan fitur yang memiliki nilai r^2 tertinggi dianggap paling signifikan dalam memberikan rekomendasi pembelian laptop. Dengan menggunakan model regresi logistik, penelitian ini diharapkan dapat menghasilkan model yang efisien dalam memprediksi pembelian laptop di Tokopedia. Model ini tidak hanya akan meningkatkan akurasi rekomendasi tetapi juga membantu platform dalam mengelola inventaris dengan lebih efektif, yang pada akhirnya meningkatkan kepuasan pelanggan dan efisiensi operasional.

II. METODOLOGI PENELITIAN



Gambar 1 Diagram Alir Metodologi Penelitian

Gambar 1 memberikan gambaran besar tentang metodologi yang digunakan dalam penelitian ini. Tahapan Pengumpulan Data sampai pada Pembersihan dan Transformasi Dataset akan dipaparkan pada Bagian Metodologi Penelitian ini. Sementara itu, Tahapan Seleksi Fitur sampai Analisis Hasil akan dijelaskan dengan rinci pada Bagian Hasil dan Pembahasan. Lalu terakhir Tahapan Kesimpulan dan Rekomendasi akan dituliskan pada Bagian Simpulan.

A. Ulasan Dataset

Dataset diambil dari `E-commerce Laptop Price Scraping - Laptop Price from Indonesia E-Commerce` oleh Arta Kusuma [15]. Dataset ini berisi harga laptop dari beberapa aplikasi e-commerce seperti Shopee dan Tokopedia. Harga laptop yang dihimpun berasal dari Oktober 2020-Januari 2021.

	name	price	location	sold	link	date
0	ACER ASPIRE 3 A314 - INTEL CELERON N4000 - 14 ...	3.590.000	Tangerang	4	https://ta.tokopedia.com/promo/v1/clicks/8a-xg...	02-10-2020
1	Laptop Acer Aspire 3 Celeron N4120 256GB SSD 4...	4.499.000	Jakarta Barat	1	https://ta.tokopedia.com/promo/v1/clicks/8a-xg...	02-10-2020
2	LAPTOP ACER ASPIRE 3 A314-21 A4 9120e RAM 4GB ...	3.883.450	Surabaya	3	https://ta.tokopedia.com/promo/v1/clicks/8a-xg...	02-10-2020
3	Acer Aspire 3 A314 A4 9120 4GB 1TB W10 14.0 DV...	3.599.000	Jakarta Pusat	2	https://ta.tokopedia.com/promo/v1/clicks/8a-xg...	02-10-2020
4	PC All In One acer Aspire C24-960 8GB 1TB Core...	7.499.900	Jakarta Pusat	3	https://ta.tokopedia.com/promo/v1/clicks/8a-xg...	02-10-2020
...

Gambar 2 Cuplikan Dataset Laptop Price from Indonesia E-Commerce

Gambar 2 menunjukkan cuplikan dari dataset ini. Dataset memiliki enam kolom, yaitu *name*, *price*, *location*, *sold*, *link*, dan *date*. Total jumlah baris pada dataset adalah 150688 baris. Dataset terdiri dari :

- 100881 baris data Oktober 2020
- 7209 baris data November 2020
- 42598 baris data Januari 2021

B. Pembersihan dan Transformasi Data

Tahapan ini akan membersihkan dan merapikan data. Semua kolom masih dalam tipe data object sehingga perlu dikonversi sesuai dengan tipe datanya. Ada tiga tahap pembersihan:

- Pembersihan I terkait dengan karakter yang tidak diinginkan pada dataset untuk kebutuhan konversi tipe data.
- Pembersihan II adalah membuang dataset yang tidak ada hubungannya dengan laptop
- Pembersihan III adalah membuang data yang tidak memiliki informasi spesifikasi yang diinginkan.

Pembersihan I: Pada pembersihan pertama, dilakukan pemeriksaan terlebih dahulu pada kolom yang seharusnya bertipe data numerik seperti karakter '.' pada kolom *price*, karakter '-' pada kolom *date*, dan karakter 'rb' pada kolom *sold*. Setelah karakter tersebut dihilangkan, data siap dikonversi.

Pembersihan II: Untuk membuat data menjadi lebih spesifik, maka dataset perlu diuraikan menjadi beberapa kolom berdasarkan spesifikasi laptop. Informasi tentang spesifikasi laptop ini dapat dilihat dari kolom *name*. Kolom baru yang akan dibentuk adalah *type*, *merk*, *prosesor*, *memory_size*, dan *disk_size*.

Sebelum menguraikan dataset, data perlu dibersihkan sekali lagi. Data yang tidak ada hubungannya dengan laptop perlu dibuang dengan kata kunci tertentu. Contohnya adalah : ['*pc design*', '*pc gaming*', '*panasonic*', '*compresor*', '*ac daikin*', '*ac gree*', '*phone*', '*asus rog 2*', '*asus rog 3*', '*rog phone*', '*hand pallet*', '*kyoritsu*', '*power supply*', '*daikin*', '*yoke mpi*', '*isolasi pipa*', '*mini power*']. Selanjutnya, data siap untuk diurai.

	name	price	location	sold	date	type
0	ACER ASPIRE 3 A314 - INTEL CELERON N4000 - 14 ...	3590000	Tangerang	4	2020-10-02	Aspire
1	Laptop Acer Aspire 3 Celeron N4120 256GB SSD 4...	4499000	Jakarta Barat	1	2020-10-02	Aspire
2	LAPTOP ACER ASPIRE 3 A314-21 A4 9120e RAM 4GB ...	3883450	Surabaya	3	2020-10-02	Aspire
3	Acer Aspire 3 A314 A4 9120 4GB 1TB W10 14.0 DV...	3599000	Jakarta Pusat	2	2020-10-02	Aspire
4	PC All In One acer Aspire C24-960 8GB 1TB Core...	7499900	Jakarta Pusat	3	2020-10-02	Aspire
...

Gambar 3 Hasil penguraian dataset untuk kolom *type*

Penguraian dimulai dari *type* terlebih dahulu berdasarkan kolom *name*. Karena penguraian untuk kolom *merk* nantinya cukup menggunakan hasil yang ditampilkan pada kolom *type*. Gambar 3 menunjukkan hasil penguraian dataset untuk kolom *type*.

Pembersihan III

Data yang tidak mencantumkan *type* pada kolom *name* akan dibuang. Lalu dilanjutkan dengan penguraian kolom *merk* berdasarkan kolom *type*. Selanjutnya penguraian untuk kolom *prosesor*, *memory_size*, dan *disk_size*. Setiap kali selesai

menguraikan, pengecekan dilakukan apakah ada data yang tidak mencantumkan nama prosesor atau ukuran memori atau ukuran disk. Jika ada, maka data tersebut perlu dibuang.

	name	price	location	sold	date	type	merk	prosesor	memory_size	disk_size
0	Laptop Acer Aspire 3 Celeron N4120 256GB SSD 4...	4499000	Jakarta Barat	1	2020-10-02	Aspire	Acer	Intel Celeron	6 GB	256 GB
1	LAPTOP ACER ASPIRE 3 A314-21 A4 9120e RAM 4GB ...	3883450	Surabaya	3	2020-10-02	Aspire	Acer	AMD A4	4 GB	1 TB
2	Acer Aspire 3 A314 A4 9120 4GB 1TB W10 14.0 DV...	3599000	Jakarta Pusat	2	2020-10-02	Aspire	Acer	AMD A4	4 GB	1 TB
3	PC All In One acer Aspire C24-960 8GB 1TB Core...	7499900	Jakarta Pusat	3	2020-10-02	Aspire	Acer	Intel Core i3	8 GB	1 TB
4	Laptop Acer Aspire 3 - A314-32 - INTEL N4000 D...	3799000	Jakarta Pusat	118	2020-10-02	Aspire	Acer	Intel Celeron	4 GB	1 TB
...
123224	ASUS VIVOBOOK FLIP TM420IA RYZEN 5-4500U 8GB 5...	11999000	Jakarta Barat	1	2021-01-26	Vivobook	Asus	AMD Ryzen 5	12 GB	512 GB
123225	Laptop Asus Vivobook K413FQ i5 8gb 512Gb 14fhd...	12400000	Jakarta Utara	2	2021-01-26	Vivobook	Asus	Intel Core i5	12 GB	512 GB
123226	Asus vivobook A416ma intel N4020 4gb 256gb 14 ...	5799000	Bandung	1	2021-01-26	Vivobook	Asus	Intel Celeron	6 GB	256 GB
123227	ASUS VIVOBOOK A407UF i5 8250U 4GB 256GB MX130 ...	9800000	Kab. Bekasi	7	2021-01-26	Vivobook	Asus	Intel Core i5	6 GB	256 GB
123228	LAPTOP ASUS VIVOBOOK K413EQ I7-1165G7 8GB 512G...	13899000	Jakarta Barat	1	2021-01-26	Vivobook	Asus	Intel Core i7	12 GB	512 GB

119059 rows x 10 columns

Gambar 4 Hasil akhir penguraian dataset

Gambar 4 menunjukkan hasil akhir dari penguraian. Perhatikan jumlah barisnya. Sebelumnya barisnya ada 150688 dan setelah hasil pembersihan dan penguraian tersisa 119059 baris. Dengan begini, data siap untuk digunakan dalam eksperimen.

C. Eksperimen: Klasifikasi, Klusterisasi, Seleksi Fitur, dan Regresi Logistik

Penelitian ini menggunakan tiga metode utama untuk seleksi fitur: ID3 (Iterative Dichotomiser 3), Random Forest (RF), dan K-Means. Algoritma ID3 digunakan untuk membuat pohon keputusan berdasarkan entropi dan gain informasi, membantu mengidentifikasi fitur yang paling berpengaruh. Random Forest, sebagai algoritma pembelajaran *ensemble*, menggunakan banyak pohon keputusan untuk analisis klasifikasi dan regresi. K-Means, sebagai algoritma pengelompokan, membagi data ke dalam beberapa klaster berdasarkan kemiripan fitur. Fitur-fitur dengan nilai korelasi diatas 0.5 digunakan dalam pengelompokan ini untuk menentukan dua klaster dengan nilai pembelian terbanyak.

Proses seleksi fitur dimulai dengan eksperimen klasifikasi menggunakan ID3 dan Random Forest untuk mengevaluasi performa model berdasarkan akurasi klasifikasi, membantu mengidentifikasi fitur-fitur paling berpengaruh. Selanjutnya, pengelompokan data dengan K-Means digunakan untuk menemukan spesifikasi laptop yang paling banyak dibeli. Fitur-fitur yang terpilih dari ketiga metode seleksi fitur ini kemudian diuji dengan regresi logistik untuk menentukan pengaruhnya terhadap keputusan pembelian. Regresi logistik digunakan untuk melihat nilai r^2 dari setiap fitur, dengan fitur yang memiliki nilai r^2 tertinggi dianggap paling signifikan dalam memberikan rekomendasi pembelian laptop.

III. HASIL DAN PEMBAHASAN

A. Seleksi Fitur dan Klasifikasi dengan Pohon Keputusan

Proses klasifikasi dilakukan dengan pembuatan pohon keputusan ID3 dan Random Forest. Data yang diinput adalah data yang sudah dibersihkan dan sudah di-*encode* untuk fitur-fitur yang akan digunakan. Dalam eksperimen klasifikasi dengan pohon keputusan ini digunakan fitur *sold*, *prosesor*, *memory size*, *disk_size*, dan *prise* yang digunakan untuk mengklasifikasikan *merk* laptop dari fitur-fitur tersebut.

```
[ ] dtree.score(X_test, y_test)
⇨ 0.6426171678145473

[ ] sklearn.metrics.accuracy_score(y_test, y_pred)
⇨ 0.6426171678145473
```

Gambar 5 Hasil Prediksi Klasifikasi Decision Tree ID3

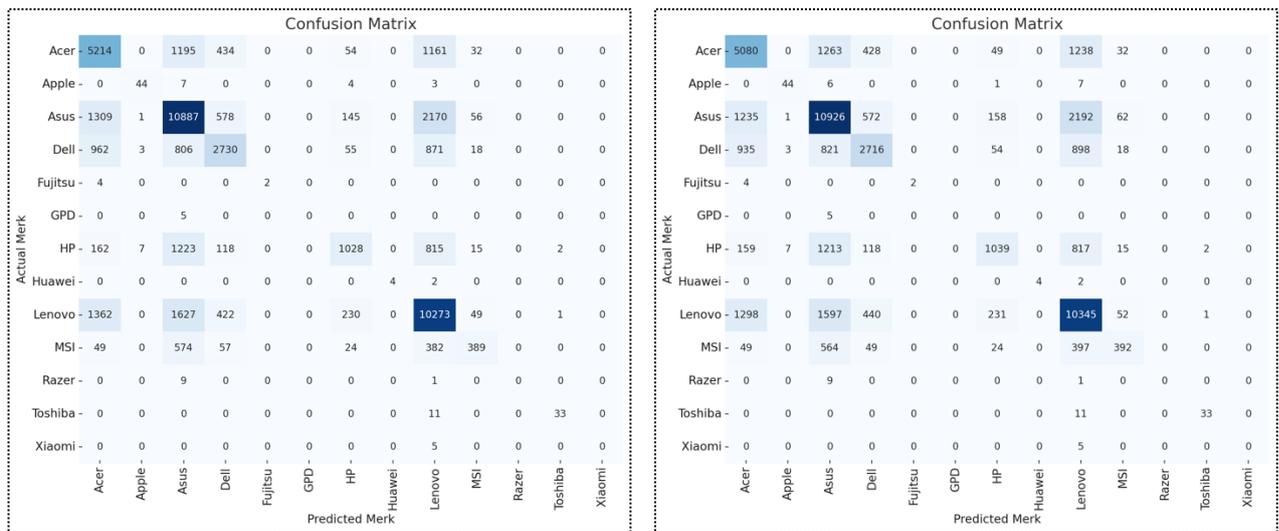
Dari Gambar 5 terlihat bahwa dengan menggunakan data histori pembelian laptop tersebut berhasil dilakukan prediksi merk laptop dengan fitur-fitur yang dipilih dengan nilai akurasi sebesar 0.6426171678145473.

```
[ ] clf.score(X_test, y_test)
0.6421342180413238

[ ] sklearn.metrics.accuracy_score(y_test, clf_pred)
0.6421342180413238
```

Gambar 6 Hasil Prediksi Klasifikasi Random Forest

Dari Gambar 6 terlihat bahwa dengan menggunakan data histori pembelian laptop tersebut berhasil dilakukan prediksi merk laptop dengan fitur-fitur yang dipilih dengan nilai akurasi sebesar 0.6421342180413238.

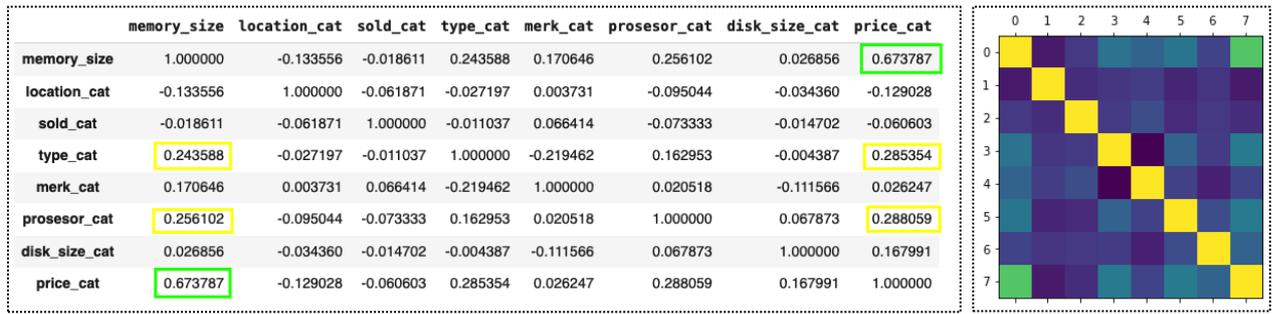


Gambar 7 Hasil Uji Coba Klasifikasi Decision Tree ID3 (plot kiri) dan Random Forest (plot kanan)

Begitu pula hasil ujicoba pengklasifikasian ID3 dan *random forest* dengan menginput secara manual data fitur yang digunakan menghasilkan hasil klasifikasi yang sama untuk ID3 maupun *Random Forest* seperti pada ditunjukkan Gambar 7.

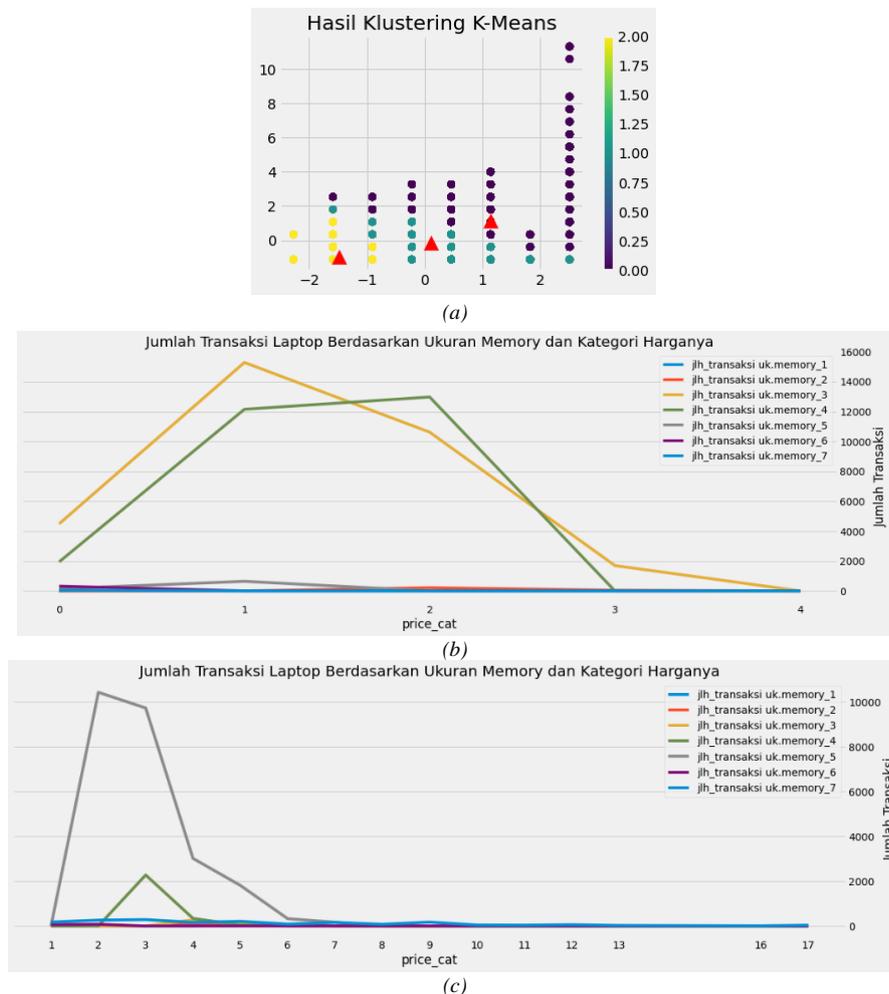
B. Seleksi Fitur dan Klusterisasi dengan K-Means

Pada proses klusterisasi yang pertama kali dilakukan adalah membuang fitur yang tidak perlu seperti *name* dan *date*. Selanjutnya dari fitur yang tersisa akan dilakukan proses *label encoding* untuk membuatnya ke dalam bentuk kategori. Hasil dari *label encoding* dicari korelasinya dengan Korelasi Pearson.



Gambar 8 Hasil Korelasi Pearson dari dataset

Gambar 8 menunjukkan hasil korelasi dari setiap fitur. Plot di sebelah hasil korelasi Pearson merupakan bentuk heatmap-nya. Label 0 = *memory_size* (ukuran memori/RAM), 1 = *location_cat* (lokasi transaksi), 2 = *sold_cat* (jumlah laptop terjual), 3 = *type_cat* (jenis laptop (*gaming/office*)), 4 = *merk_cat* (merk laptop), 5 = *prosesor_cat* (jenis prosesor), 6 = *disk_size_cat* (ukuran *storage*), dan 7 = *price_cat* (kategori harga). Pada hasil korelasi di Gambar 8, korelasi 0.5 ke atas hanya *price_cat* dengan *memory_size*. Pasangan fitur ini akan diuji coba dalam klusterisasi. Penentuan jumlah *cluster* menggunakan penentuan *knee* atau *elbow*. Hasilnya adalah tiga kluster.



Gambar 9 Hasil Klusterisasi pada fitur *price_cat* dan *memory_size* (a), Jumlah Transaksi Laptop untuk Kluster 0 (b), dan Jumlah Transaksi Laptop untuk Kluster 1 (c)

Gambar 9a menunjukkan grafik klusterisasi dari *price_cat* dengan *memory_size*. Kluster 0 (warna ungu) memiliki data sebesar 60591, Kluster 1 (warna biru langit) memiliki data sebesar 30661 baris, dan Kluster 2 (warna kuning) memiliki data sebesar 27804 baris. Kluster 0 dan 1 dipilih untuk identifikasi karakteristiknya lebih lanjut karena jumlah datanya yang banyak. Gambar 9b menunjukkan Kluster 0 terdiri dari data pembelian laptop dengan kategori harga dari 12 *jutaan* sampai 22 *jutaan* dengan ukuran memori 16 GB. Sedangkan Gambar 9c menunjukkan Kluster 1 terdiri dari data pembelian laptop dengan kategori harga dari 7 *jutaan* sampai 17 *jutaan* dengan ukuran memori 8 GB dan 12 GB. Kedua fitur ini juga akan digunakan dalam Regresi Logistik untuk dikomparasi dengan fitur dalam hipotesis yaitu *merk_cat*.

C. Regresi Logistik:

Regresi logistik dalam penelitian ini menggunakan tiga fitur, yaitu *merk*, sesuai dengan rumusan masalah dalam penelitian ini, *memory_size* dan *price_cat* yang merupakan fitur terbaik dari hasil klusterisasi yang dilakukan sebelumnya. Ketiga fitur tersebut dicari regresi logistiknya dengan mengacu pada banyaknya laptop yang terjual dari setiap kategori pada masing-masing fitur tersebut.

TABEL 1
HASIL PERHITUNGAN REGRESI LOGISTIK LAPTOP YANG TERJUAL BERDASARKAN MERK

Merk	Terjual	Jumlah Transaksi	Probabilitas	Odds (P/(1-P))	Odds (Terjual/Tidak Terjual)	Logit/ln(Odds) (Y)
0	113793	754734	0.150772325	0.177540522	0.177540522	-1.728556401
1	2283	754734	0.003042907	0.003048085	0.003048085	-5.797845981
2	216429	754734	0.286761959	0.40205455	0.40205455	-0.911627655
3	38783	754734	0.05186316	0.054169908	0.054169908	-2.916527985
4	120	754734	0.000158996	0.000159022	0.000159022	-8.744569869
5	54	754734	7.15484E-05	7.1555E-05	7.1555E-05	-9.54506051
6	49485	754734	0.06556146	0.070166707	0.070166707	-2.658881339
7	268	754734	0.000355092	0.000355218	0.000355218	-7.942897365
8	298640	754734	0.395889077	0.654777735	0.654777735	-0.42244566
9	31559	754734	0.041861472	0.043639056	0.043639056	-3.129580237
10	48	754734	6.35986E-05	6.36026E-05	6.36026E-05	-9.662856037
11	3260	754734	0.004319403	0.004338141	0.004338141	-5.440347339
12	12	754734	1.58996E-05	1.58999E-05	1.58999E-05	-11.0491981

Berdasarkan Tabel 1, dengan variabel x adalah merk laptop, diperoleh persamaan regresi $y = -0,393x - 3,022$ dengan nilai r^2 sebesar 0,173. Nilai r^2 ini menunjukkan bahwa hanya sekitar 17,3% variasi dalam keputusan pembelian laptop yang dapat dijelaskan oleh merk laptop. Ini merupakan indikasi bahwa merk laptop tidak memiliki pengaruh yang kuat terhadap keputusan pembelian. Hasil ini dapat disebabkan oleh beberapa faktor, termasuk preferensi konsumen yang bervariasi dan tidak selalu berfokus pada merk sebagai faktor utama dalam keputusan pembelian. Banyak konsumen mungkin lebih memperhatikan faktor-faktor lain seperti harga, spesifikasi teknis, atau ulasan produk daripada merk itu sendiri. Oleh karena itu, meskipun merk merupakan elemen penting dalam pemasaran, hasil analisis ini menunjukkan bahwa dalam konteks pembelian laptop di Tokopedia, merk bukanlah prediktor utama yang kuat.

TABEL 2
HASIL PERHITUNGAN REGRESI LOGISTIK LAPTOP YANG TERJUAL BERDASARKAN MEMORY

Merk	Terjual	Jumlah Transaksi	Probabilitas	Odds (P/(1-P))	Odds (Terjual/Tidak Terjual)	Logit/ln(Odds) (Y)
0	2001	754734	0.002651265	0.002658313	0.002658313	-5.930063528
1	179810	754734	0.238242878	0.31275481	0.31275481	-1.16233712
2	30920	754734	0.040960876	0.042718157	0.042718157	-3.15313123
3	184088	754734	0.24391111	0.322595795	0.322595795	-1.131551833
4	210996	754734	0.279533396	0.388047184	0.388047184	-0.946628357
5	135831	754734	0.179972017	0.219470579	0.219470579	-1.51653079
6	6502	754734	0.008614956	0.008698818	0.008698818	-4.735602363
7	4586	754734	0.006073613	0.00611346	0.00611346	-5.079762336

Berdasarkan Tabel 2, dengan variabel x adalah besar memori laptop, diperoleh persamaan regresi $y = -0,083x - 2,669$ dengan nilai r^2 sebesar 0,0099. Nilai r^2 yang sangat rendah ini menunjukkan bahwa hanya sekitar 0,99% variasi dalam keputusan pembelian laptop yang dapat dijelaskan oleh besar memori. Ini menandakan bahwa besar memori laptop hampir tidak memiliki korelasi dengan keputusan pembelian. Alasan di balik temuan ini bisa beragam, salah satunya adalah bahwa

konsumen mungkin tidak menganggap besar memori sebagai faktor krusial dalam pembelian mereka, atau mereka mungkin melihat spesifikasi lainnya seperti prosesor atau harga sebagai lebih penting. Selain itu, dengan kemajuan teknologi, perbedaan dalam ukuran memori mungkin tidak lagi signifikan bagi banyak pengguna karena standar memori yang lebih tinggi telah menjadi umum di banyak laptop. Temuan ini menggarisbawahi pentingnya mempertimbangkan berbagai faktor lain yang mungkin lebih berpengaruh dalam keputusan pembelian laptop.

TABEL 3
HASIL PERHITUNGAN REGRESI LOGISTIK LAPTOP YANG TERJUAL BERDASARKAN HARGA LAPTOP

Merk	Terjual	Jumlah Transaksi	Probabilitas	Odds (P/(1-P))	Odds (Terjual/Tidak Terjual)	Logit/ln(Odds) (Y)
1	201892	754734	0.267500868	0.365189331	0.365189331	-1.007393944
2	211708	754734	0.280560775	0.389867152	0.389867152	-0.941948997
3	70362	754734	0.093227548	0.102815205	0.102815205	-2.274488885
4	13276	754734	0.017590303	0.017905629	0.017905629	-4.022660766
5	7922	754734	0.010496413	0.010602779	0.010602779	-4.546196137
6	932	754734	0.001234872	0.001236399	0.001236399	-6.695779146
7	518	754734	0.000686305	0.000686806	0.000686806	-7.28453583
8	99	754734	9.6722E-05	9.6725E-05	9.6725E-05	-9.240828885
9	281	754734	0.000372317	0.000372455	0.000372455	-7.895293533
10	56	754734	7.4245E-05	7.68425E-05	7.68425E-05	-9.477654964
11	58	754734	7.68435E-05	7.6845E-05	7.6845E-05	-9.477480548
12	76	754734	0.000100696	0.000100699	0.000100699	-9.203891685
13	23	754734	3.04734E-05	3.04752E-05	3.04752E-05	-10.39773829
14	24	754734	3.17993E-05	3.18002E-05	3.18002E-05	-10.35851155
17	99	754734	0.000131172	0.000131189	0.000131189	-8.938869617

Berdasarkan Tabel 3, dengan variabel x adalah harga laptop, diperoleh persamaan regresi $y = -0,628x - 1,546$ dengan nilai r^2 sebesar 0,819. Nilai r^2 yang tinggi ini menunjukkan bahwa sekitar 81,9% variasi dalam keputusan pembelian laptop dapat dijelaskan oleh harga laptop. Ini menunjukkan bahwa harga adalah faktor yang sangat signifikan dalam mempengaruhi keputusan pembelian. Konsumen cenderung sangat sensitif terhadap harga ketika membeli laptop, yang dapat dijelaskan oleh keterbatasan anggaran atau keinginan untuk mendapatkan nilai terbaik untuk uang yang mereka keluarkan. Hasil ini konsisten dengan banyak penelitian sebelumnya yang menemukan bahwa harga sering kali menjadi faktor penentu utama dalam keputusan pembelian produk, terutama barang elektronik.

IV. SIMPULAN

Dalam proses pengolahan data, ada banyak data yang dibuang karena dalam penulisan nama produknya tidak mencantumkan spesifikasi yang digunakan dalam penguraian untuk membuat fitur yang baru. Sehingga disarankan Tokopedia dapat membuat aturan penulisan yang seragam untuk penjualan laptop. Didalamnya terdapat merk, tipe, ukuran memori, ukuran disk, dan prosesor. Hal ini akan sangat membantu dalam mengeksplorasi dan menganalisis histori pembelian laptop dari setiap toko/merchant.

Secara keseluruhan, hasil dari analisis regresi logistik ini menunjukkan bahwa merk laptop dan besar memori bukanlah faktor dominan dalam mempengaruhi keputusan pembelian, sementara harga laptop memiliki pengaruh yang sangat signifikan. Ini menekankan perlunya mempertimbangkan variabel lain yang mungkin lebih relevan selain merk dan besar memori, seperti harga, fitur tambahan, atau ulasan pengguna dalam model prediksi pembelian. Temuan ini memberikan wawasan berharga bagi pemasar dan pengelola inventaris di Tokopedia untuk mengarahkan upaya mereka pada faktor-faktor yang lebih signifikan dalam mempengaruhi perilaku konsumen. Dengan demikian, strategi pemasaran yang lebih efektif dapat difokuskan pada penetapan harga yang kompetitif dan promosi yang menarik untuk meningkatkan keputusan pembelian dan kepuasan pelanggan.

DAFTAR PUSTAKA

- [1] D. Prawira, R. Kurniawan, R. E. Indrajit and E. Dazki, "Effect of E-Service Quality, Brand Image and E-Security Seals on TOKOPEDIA E-Customer Satisfaction (Case Study on Tokopedia Customers)," *Jurnal Informasi & Teknologi (JiIT)*, vol. 5, no. 1, pp. 153-164, 2023.
- [2] S. Mahardini, I. I. Mudafia and R. Apricuansyah, "Kualitas Produk, Harga, dan Kualitas Layanan Terhadap Keputusan Pembelian Tokopedia di Wilayah DKI Jakarta," *Jurnal Manajemen dan Perbankan (JUMPA)*, vol. 10, no. 2, 2023.
- [3] F. Himmah, T. Wuryandari and A. Hoyyi, "Analisis Pengaruh Strategi Bauran Pemasaran terhadap Pemilihan Merek Laptop Menggunakan Regresi Logistik Multinomial (Studi Kasus Mahasiswa Universitas Diponegoro)," *Media Statistika*, vol. 5, no. 1, pp. 17-26, 2012.
- [4] C. Sunita, "Artificial Intelligence in E-Commerce: Exploring the Purchase Decisions through Logistic Regression Analysis," *Quing: International Journal of Commerce and Management*, 2023.

- [5] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama and P. Yanque-Churo, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020.
- [6] K. Sunday, P. Ocheja, S. Hussain, S. S. Oyelere, B. O. Samson and F. J. Agbo, "Analyzing Student Performance in Programming Education Using Classification Techniques," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 15, no. 2, p. 127, 2020.
- [7] S. Ratnasingam and J. Muñoz-Lopez, "Distance Correlation-Based Feature Selection in Random Forest," *Entropy*, vol. 25, 2023.
- [8] B. Cui, "Prediction and Analysis of Shoes Popularity Based on Random Forest Regression Model," *Advances in Economics, Management and Political Sciences*, vol. 42, pp. 157-169, 2023.
- [9] L. Zhu, D. Qiu, D. Ergu, C. Ying and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503-513, 2019.
- [10] S. G, S. M and D. N, "Regression Analysis-Based Predictive Model for E-Commerce Application," in *2023 International Conference on Networking and Communications (ICNWC)*, Chennai, 2023.
- [11] U. Lilhore, S. S., D. Prasad and D. Verma, "Hybrid Weighted Random Forests Method for Prediction & Classification of Online Buying Customers," *Journal of Information Technology Management*, vol. 13, pp. 245-259, 2021.
- [12] V. R. Prasetyo, F. A. Miranti and S. Limanto, "Implementation of Feature Selection to Reduce the Number of Features in Determining the Initial Centroid of K-Means Algorithm," in *2022 International Conference on Informatics Electrical and Electronics (ICIEE)*, Yogyakarta, 2022.
- [13] X. Zeng and H. Zheng, "CS Sparse K-means: An Algorithm for Cluster-Specific Feature Selection in High-Dimensional Clustering," *ArXiv*, 2019.
- [14] S. Kumar, B. Bhushan, L. Bhambhu, M. Thakur, U. M. Mohapatra and D. K. Choubey, "Medical Datasets Classification using a Hybrid Genetic Algorithm for Feature Selection based on Pearson Correlation Coefficient," in *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, Bhubaneswar, 2022.
- [15] A. Kusuma, "E-commerce Laptop Price Scraping," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/artakusuma/laptopecomercee>. [Accessed 15 June 2024].