Klasifikasi Harga Ponsel Menggunakan Algoritma *Logistic Regression*

Danika Najwa Ardelia[1], Hilda Desfianty Arifin[2], Sena Daniswara[3], Anggraini Puspita Sari[4*]

[1,2,3,4] *Program Studi Informatika - UPN "Veteran" Jawa Timur*

Jl. Raya Rungkut Madya Gunung Anyar – Surabaya Jawa Timur - Indonesia

22081010103@student.upnjatim.ac.id, 22081010206@student.upnjatim.ac.id, 22081010126@student.upnjatim.ac.id,

**Corresponding author email*: anggraini.puspita.if@upnjatim.ac.id

Abstrak*— Dalam era digital yang berkembang pesat, penggunaan produk teknologi dan pertumbuhan internet membuka peluang besar dalam penjualan ponsel pintar. Fitur-fitur yang semakin beragam membuat konsumen merasa bingung memilih ponsel dengan harga yang sesuai. Oleh karena itu, penggunaan model *logistic regression* menjadi pilihan yang tepat untuk mengkategorikan harga ponsel menjadi empat tingkatan: rendah, sedang, tinggi, dan sangat tinggi, yang nantinya diharapkan dapat membantu para konsumen memilih ponsel yang sesuai dengan kebutuhan mereka berdasarkan kategori harga. Penelitian ini juga mengkombinasikan *logistic regression* dengan penyetelan *hyperparameter* dimana penyetelan *hyperparameter* dilakukan untuk meningkatkan akurasi model. Penyetelan *hyperparameter* dilakukan menggunakan metode *grid search*. Dalam penelitian ini, dilakukan proses pengumpulan dataset yang kemudian akan dilakukan pengecekan terhadap nilai-nilai yang tidak valid melalui proses *preprocessing*. Data kemudian dibagi menjadi data uji dan data latih dengan menggunakan dua perbandingan, 80:20 dan 90:10. Setelah data dibagi, dilakukan pemodelan dan penyetelan *hyperparameter* untuk mengoptimalkan model *logistic regression*. Hasil tingkat akurasi yang didapatkan dalam proses ini yaitu 98% yang didapatkan dengan menggunakan perbandingan data split 90:10. Dengan demikian, penggunaan *logistic regression* dapat memprediksi kategori harga ponsel dengan tingkat akurasi yang tinggi. Hal ini dapat diharapkan membantu konsumen dalam memilih ponsel yang sesuai dengan kebutuhan dan anggaran mereka.

Kata Kunci*— Klasifikasi, harga ponsel, *logistic regression*, *hyperparameter*, *data split

Abstract*— *In the rapidly developing digital era, the use of technology products and the growth of the internet create significant opportunities in smartphone sales. The increasingly diverse features can make consumers feel confused about choosing a cellphone at the right price. Therefore, using a logistic regression model is an appropriate choice to categorize cellphone prices into four levels: low, medium, high, and very high, which is expected to help consumers select a phone that meets their needs based on price categories. This research also combines logistic regression with hyperparameter tuning, where hyperparameter tuning is conducted to improve the model's accuracy. Hyperparameter tuning is performed using the grid search method. In this research, a dataset collection process is carried out, followed by checking for invalid values through preprocessing. The data is then divided into test and training data using two ratios: 80:20 and 90:10. After the data is split, modeling and hyperparameter tuning are conducted to optimize the logistic regression model. The resulting accuracy level obtained from this process is 98%, achieved using the 90:10 data split ratio. Therefore, the use of logistic regression can predict smartphone price categories with a high level of accuracy. This can be expected to help consumers in choosing phones that match their needs and budget.

Keywords*— *Classification, mobile phone price, logistic regression, hyperparameter, split data

I. PENDAHULUAN

Era digital telah merevolusi cara kita hidup, termasuk cara kita membeli dan menggunakan produk teknologi. Pertumbuhan internet yang pesat, khususnya di Indonesia, membuka peluang besar bagi industri, termasuk penjualan ponsel pintar. Perkembangan pesat teknologi seluler mendorong inovasi dan menghadirkan berbagai pilihan ponsel dengan fitur-fitur canggih. Namun, keragaman dan kompleksitas fitur yang ditawarkan di pasar menimbulkan tantangan dalam klasifikasi harga. Konsumen seringkali kebingungan dalam memilih ponsel yang tepat dengan harga yang sesuai. Oleh karena itu, penelitian yang mengkategorikan ponsel berdasarkan harga menjadi penting [1].

Penelitian ini akan menggunakan algoritma yang berfungsi untuk klasifikasi. Beberapa algoritma yang umum digunakan dalam klasifikasi termasuk *K-Nearest Neighbors* (KNN), *decision tree*, *random forest*, *Support Vector Machine* (SVM), dan *neural network*. *Neural network* sangat efektif dalam mengenali pola-pola kompleks dan sering digunakan dalam berbagai aplikasi canggih. Sebagai contoh, penelitian sebelumnya telah memanfaatkan *deep neural network* untuk klasifikasi kecepatan dan arah angin [2]. Namun, di antara berbagai pilihan algoritma tersebut, penelitian ini akan menggunakan *logistic regression* sebagai algoritma utama untuk klasifikasi. *Logistic regression d*ipilih karena berdasarkan penelitian sebelumnya yang dilakukan oleh Egipta Pranadjaya dkk tentang klasifikasi jenis kendaraan pada sistem tilang digital. Dalam penelitian tersebut, dilakukan perbandingan beberapa metode seperti *neural network*, *Support Vector Machine* (SVM) dan *logistic regression*, dimana hasil yang diperoleh menunjukkan bahwa

logistic regression menghasilkan tingkat akurasi yang lebih tinggi dibanding *neural network* dan *Support Vector Machine* (SVM) [3].

Penelitian ini bertujuan untuk mengembangkan model klasifikasi harga ponsel yang lebih akurat dengan menggunakan kombinasi *logistic regression* dan penyetelan *hyperparameter*. Dalam penelitian ini, data ponsel dikategorikan ke dalam empat kategori harga, yaitu rendah, sedang, tinggi, dan sangat tinggi. Model *logistic regression* digunakan untuk memprediksi kategori harga berdasarkan fitur-fitur ponsel, seperti ukuran layar, prosesor, dan memori.

Logistic regression merupakan metode yang menganalisis hubungan antara variabel independen dan variabel dependen dalam data nominal atau ordinal [4]. Metode ini memodelkan hubungan antara berbagai faktor independen (seperti memori perangkat, ukuran perangkat, dan durasi baterai) dan variabel hasil yang terdiri dari kelompok harga rendah, sedang, tinggi, atau sangat tinggi. Fungsi logistik menunjukkan peluang sebuah ponsel masuk ke dalam kategori harga tertentu. *Logistic regression* dibagi ke dalam tiga tipe, yaitu *binary logistic regression*, *multinomial logistic regression*, dan *ordinal logistic regression* [5]. Penelitian ini menggunakan tipe model *ordinal logistic regression* untuk memprediksi kategori harga dengan lebih beragam. Model ini cocok untuk respon ordinal dengan data bertingkat, banyak kategori, dan diawali dengan angka [6].

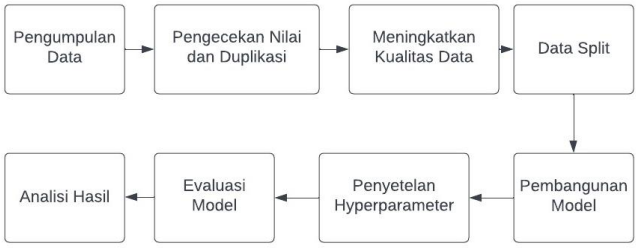
Dalam penelitian ini, penyetelan *hyperparameter* dilakukan untuk meningkatkan akurasi model. Penyetelan *hyperparameter* dilakukan menggunakan metode *grid search*. Penelitian sebelumnya menunjukkan bahwa penggunaan *grid search* pada *logistic regression* dapat meningkatkan kinerja model dalam hal akurasi prediksi. Sebagai contoh, sebuah studi yang memprediksi penyakit Diabetes Mellitus dengan metode *grid search* menunjukkan peningkatan akurasi model klasifikasi dari 72,22% menjadi 83,33% setelah menerapkan metode tersebut [7]. Hasil ini menunjukkan bahwa *grid search* efektif untuk meningkatkan akurasi prediksi model.

Penelitian ini juga akan membahas pengaruh variasi persentase pembagian data terhadap performa model klasifikasi harga ponsel. Persentase pembagian data antara data latih dan data uji memiliki peran penting dalam evaluasi dan validasi model [8]. Pentingnya pembagian data yang tepat adalah untuk memastikan bahwa model yang dibuat mampu menggeneralisasi pola dari data yang tidak terlihat. Oleh karena itu, penelitian ini akan mengeksplorasi dua variasi persentase pembagian data dan mengevaluasi dampaknya terhadap kinerja model secara menyeluruh.

Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan dalam bidang klasifikasi harga ponsel pintar dengan akurasi yang lebih tinggi. Dengan menggunakan metode *logistic regression* dan penyetelan *hyperparameter*, model yang dihasilkan dapat membantu konsumen dalam memilih ponsel yang sesuai dengan kebutuhan mereka berdasarkan kategori harga. Penelitian ini juga menggarisbawahi pentingnya pemilihan persentase pembagian data yang tepat dalam evaluasi model, sehingga model yang dikembangkan dapat memberikan prediksi yang konsisten dan dapat diandalkan. Dengan demikian, diharapkan penelitian ini dapat memberikan panduan yang berharga bagi industri untuk menyediakan pilihan ponsel yang lebih sesuai dengan kebutuhan dan preferensi konsumen.

II. METODE PENELITIAN

Penelitian ini mengajukan konsep model pembelajaran mesin yang menggunakan algoritma *logistic regression* untuk mengkategorikan harga ponsel ke dalam empat tingkatan: 0 (rendah), 1 (sedang), 2 (tinggi), dan 3 (sangat tinggi). Metodologi penelitian melibatkan serangkaian proses yang dimulai dari pengumpulan data, *preprocessing*, pemisahan data, pembangunan model, penyetelan *hyperparameter*, evaluasi model, hingga analisis hasil yang diperoleh, seperti yang terlihat pada Gbr. 1.



Gbr. 1 Alur Metode Penelitian

A. *Pengumpulan* Data

Dataset yang digunakan dalam penelitian ini diperoleh dari sumber terbuka Kaggle, dengan judul "Mobile Price Classification" [9]. Kumpulan data yang diteliti terdiri dari 2000 sampel, dengan masing-masing sampel dilengkapi variabel-variabel yang merinci spesifikasi teknis dan karakteristik ponsel seperti yang terlihat pada tabel 1.

TABEL I
VARIABEL PENELITIAN

Variabel	**Deskripsi**
battery_power	Daya baterai dalam mAh
blue	Dukungan Bluetooth (0 = tidak, 1 = ya)
clock_speed	Kecepatan clock prosesor dalam GHz
dual_sim	Dukungan dual SIM (0 = tidak, 1 = ya)
fc	Resolusi kamera depan dalam megapiksel
four_g	Dukungan 4G (0 = tidak, 1 = ya)
int_memory	Memori internal dalam GB
m_dep	Ketebalan perangkat dalam cm
mobile_wt	Berat perangkat dalam gram
n_cores	Jumlah core prosesor
pc	Resolusi kamera utama dalam megapiksel
px_height	Resolusi piksel layar (tinggi)
px_width	Resolusi piksel layar (lebar)

ram	Kapasitas RAM dalam MB
sc_h	Tinggi layar dalam cm
sc_w	Lebar layar dalam cm
talk_time	Waktu bicara dalam jam
three_g	Dukungan 3G (0 = tidak, 1 = ya)
touch_screen	Layar sentuh (0 = tidak, 1 = ya)
wifi	Dukungan WiFi (0 = tidak, 1 = ya)

B. Preprocessing

Tahap pertama dalam *preprocessing* melibatkan pengecekan terhadap nilai-nilai yang tidak ada atau duplikasi. Setelah itu, untuk meningkatkan kualitas data, *outlier* diidentifikasi dan dieliminasi dengan menggunakan metode Rentang Antar Kuartil (IQR). *Outlier* adalah data yang mencolok karena nilainya yang sangat berbeda dari mayoritas data lainnya. Kehadiran *outlier* dapat berdampak pada analisis data, seperti membuat proses klasifikasi menjadi kurang akurat [10].

Untuk setiap fitur, dihitung korelasinya dengan variabel target '*price_range*' dengan menggunakan koefisien korelasi Pearson. Koefisien korelasi Pearson merupakan ukuran statistik yang digunakan untuk menentukan kekuatan dan arah hubungan linier antara dua variabel kontinu. Biasanya dilambangkan dengan simbol r.

$$r = \frac{\Sigma(xi - \underline{x})(yi - \underline{y})}{\sqrt{\Sigma(xi - \underline{x})^2 \Sigma(yi - \underline{y})^2}} \quad (1)$$

Nilai xi dan yi mewakili nilai individu dari variabel x dan y, sementara x dan y adalah rata-rata dari masing-masing variabel tersebut. Koefisien korelasi Pearson dapat berkisar dari -1 (korelasi negatif sempurna) hingga 1 (korelasi positif sempurna), dengan nilai 0 menunjukkan tidak adanya korelasi antara variabel tersebut. Semakin mendekati nilai r ke 1 atau -1, semakin kuat korelasi antara kedua variabel tersebut. Pada penelitian ini, hanya fitur-fitur yang memiliki nilai korelasi lebih dari 0.1 yang akan dianalisis lebih mendalam, seperti yang terlihat pada tabel 2 [11][12].

TABEL II
NILAI KORELASI FITUR DENGAN VARIABEL TARGET

Fitur	Nilai Korelasi
ram	0.9170457362649905
battery_power	0.20072261211373094
px_width	0.16581750172625515
px_height	0.14885755500042175

C. Data Split

Data latih dan data uji adalah komponen penting dalam pembelajaran mesin, di mana data latih digunakan untuk membangun model dan data uji untuk mengevaluasi akurasinya. Pembagian persentase antara keduanya sangat mempengaruhi performa model. Penelitian menunjukkan bahwa komposisi yang tepat dapat mengoptimalkan akurasi, sementara komposisi yang tidak tepat dapat menyebabkan fluktuasi akurasi, terutama pada dataset dengan distribusi data yang tidak merata [13]. Penelitian ini akan menggunakan dua persentase *data split* yang berbeda yaitu, 80:20 dan 90:10 dengan jumlah data seperti pada Tabel 3.

TABEL III
JUMLAH DATA LATIH DAN DATA UJI

Persentase (%)	Jumlah Data Latih	Jumlah Data Uji
80:20	1600	400
90:10	1800	200

D. Pemodelan

Dalam pengembangan model pembelajaran mesin, *pipeline* yang diberikan menggunakan *StandardScaler* untuk menormalkan data dan *SelectKBest* untuk seleksi fitur, diikuti oleh penerapan algoritma *logistic regression* untuk klasifikasi. Pendekatan ini memastikan proses yang sistematis dan dapat diulang, dengan *preprocessing* yang efektif dan pelatihan model yang koheren, meningkatkan kemungkinan model yang akurat dan dapat diandalkan. Dalam penelitian ini, teknik *logistic regression* yang digunakan adalah *ordinal logistic regression*.

Ordinal logistic regression merupakan metode yang digunakan untuk menganalisis variabel respons yang berskala ordinal dengan tiga kategori atau lebih, yang memiliki urutan. Model ordinal logistic regression ini diwujudkan dalam bentuk model logit kumulatif. Variabel prediktor menggunakan data yang bersifat kualitatif. Variabel respons *Y* dalam model logit ini memiliki karakteristik ordinal yang diekspresikan melalui peluang kumulatif. Oleh karena itu, model logistik kumulatif ini diperoleh dengan membandingkan peluang keseluruhan, yaitu probabilitas bahwa variabel respons berada di atau di bawah kategori respons ke-j.

Jika diasumsikan terdapat variabel respons Y yang berskala ordinal dengan J kategori, dan $X^{\mathrm{T}} = (x_1, x_2, \ldots, x_p)$ merupakan vektor dari variabel-variabel penjelas, maka probabilitas bahwa Y berada dalam kategori ke-j, pada nilai tertentu dari X, dapat dinyatakan sebagai $P[Y = j|\mathrm{x}] = \pi j(x)$, di mana $\pi j(x)$ merujuk pada probabilitas kategori ke-j dari Y berdasarkan nilai X yang diberikan. Peluang kumulatifnya adalah sebagai berikut:

$$P[Y \leq j|\mathrm{X}] = \frac{exp\ (\alpha J + XT\beta)}{1 + exp\ (\alpha J + XT\beta)} \quad (2)$$

Dimana $xi = (xi1\ ,\ xi2\ ,\ \ldots\ ,\ xip)$ adalah nilai yang diamati untuk observasi ke-i (dengan i=1,2,…,n) dari setiap dari p variabel prediktor. Estimasi parameter regresi dilakukan dengan cara mendekomposisi menggunakan transformasi logit dari $[Y \leq j|x]$ [14].

E. Penyetelan Hyperparameter

GridSearchCV digunakan untuk penyetelan *hyperparameter*, dan lima kali validasi silang dilakukan untuk mendapatkan hasil terbaik. *GridSearchCV* digunakan dalam penelitian ini untuk mengoptimalkan model *logistic regression* dalam mengestimasi kategori harga ponsel. Nilai regularization C dan tipe *solver logistic regression* merupakan contoh parameter yang ditingkatkan. Dalam *GridSearchCV*,

validasi silang membantu dalam mengevaluasi seberapa baik kinerja kombinasi *hyperparameter* model yang berbeda. Menggabungkan *hyperparameter* dengan performa terbaik memungkinkan dalam menemukan kombinasi *hyperparameter* yang optimal untuk model. [15]

F. Evaluasi Model

Model yang telah dilatih dievaluasi menggunakan beberapa metrik, antara lain akurasi, *precision*, *recall*, *f1-score*, *support*, dan *confusion matrix*. Hasil evaluasi menunjukkan tingkat akurasi model dan memberikan gambaran tentang kinerja model pada data uji.

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN} \text{ x } 100\% \quad (3)$$

$$Precision = \frac{TP}{TP + FN} \text{ x } 100\% \quad (4)$$

$$Recall = \frac{TP}{FP + TP} \text{ x } 100\% \quad (5)$$

$$F1\text{-}score = 2 \text{ x } \frac{Precision \; x \; Recall}{Precision + Recall} \text{ x } 100\% \quad (6)$$

Dimana TP merupakan *True Positive*, TN merupakan *True Negative*, FP merupakan *False Positive* dan FN merupakan *False Negative.*

G. Analisis Hasil

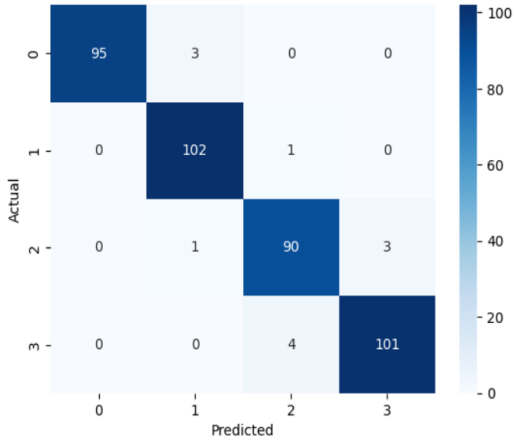
Hasil dari evaluasi model divisualisasikan menggunakan *heatmap confusion matrix.* Selain itu, dilakukan analisis terhadap prediksi yang benar dan salah untuk memberikan wawasan lebih lanjut tentang kinerja model dalam kondisi nyata.

III. HASIL DAN PEMBAHASAN

Pada pembahasan ini, akan dijelaskan secara rinci mengenai hasil klasifikasi harga ponsel menggunakan model *logistic regression*. Hasil ini akan dilakukan evaluasi *data split* menggunakan dua perbandingan yaitu 80:20 dan 90:10. Untuk penyajian hasil dilakukan menggunakan Colab dengan bahasa pemrograman Python.

Pada perbandingan *data split* 80:20, sebanyak 80% dari total data digunakan untuk data latih dan 20% sisanya digunakan untuk data uji. Hasil akurasi pada perbandingan ini adalah 0.97 atau tingkat keberhasilan dalam memprediksi harga ponsel sebesar 97%.

Confusion matrix digunakan untuk mengevaluasi kinerja *logistic regression* dalam memperhitungkan berapa banyak prediksi yang benar dan salah pada klasifikasi setiap kelas [16].



Gbr. 2 *Confusion Matrix* 80:20

Seperti pada Gbr. 2, ditunjukkan *confusion matriks* dari perbandingan *data split* 80:20. *Confusion matrix* menyajikan hasil bahwa pada kelas 0, terdapat 95 sampel dimana semuanya benar diprediksi sebagai kelas 0. Pada kelas 1, terdapat 102 sampel dimana sebanyak 101 sampel benar diprediksi sebagai kelas 1 dan terdapat 1 sampel diprediksi sebagai kelas 2. Pada kelas 2, dari 94 sampel, 90 sampel diantaranya benar diprediksi sebagai kelas 2, 1 sampel diprediksi sebagai kelas 1 dan 3 sampel diprediksi sebagai kelas 3. Pada kelas 3, dari 105 sampel semuanya benar diprediksi sebagai kelas 3 kecuali 4 sampel yang diprediksi sebagai kelas 2.

	precision	recall	f1-score	support
0	1.00	0.97	0.98	98
1	0.96	0.99	0.98	103
2	0.95	0.96	0.95	94
3	0.97	0.96	0.97	105
accuracy			0.97	400
macro avg	0.97	0.97	0.97	400
weighted avg	0.97	0.97	0.97	400

Gbr. 3 *Classification Report* 80:20

Berdasarkan nilai dari *classification report* pada Gbr. 3, dijelaskan bahwa model *logistic regression* pada perbandingan *data split* 80:20 memiliki kinerja yang sangat baik dan akurat dalam memprediksi klasifikasi harga ponsel berdasarkan data latih dengan tingkat akurasi 97%. Hal tersebut juga ditunjukkan dengan nilai presisi, *recall*, dan *f1-score* yang tinggi untuk setiap kelas. Selain itu, terlihat bahwa presisi dan *recall* untuk setiap kelas cenderung mendekati nilai 1. Hal ini menandakan bahwa *logistic regression* pada perbandingan *data split* 80:20 memiliki keseimbangan optimal antara kemampuan untuk memberikan prediksi yang akurat dan kemampuan untuk mengidentifikasi sebagian besar contoh yang sebenarnya positif.

Untuk memberikan gambaran yang lebih jelas tentang kinerja model, berikut merupakan data-data prediksi yang benar dan salah untuk pembagian persentase *data split* 80:20.

	battery_power	px_height	px_width	ram	Actual	Predicted	Correct
256	601	356	765	532	0	0	True
352	1604	134	939	916	0	0	True
298	928	221	1243	666	0	0	True
581	1512	1079	1897	3607	3	3	True
1288	1541	796	1052	1108	1	1	True
...	...	...	...	...	...	...	...
1616	1986	251	599	3476	3	3	True
650	1315	59	575	3278	2	2	True
261	728	526	1529	2039	1	1	True
1305	1023	5	1744	2086	1	1	True
966	1910	985	1284	309	0	0	True

Gbr. 4 *Prediksi Benar Data Split* 80:20

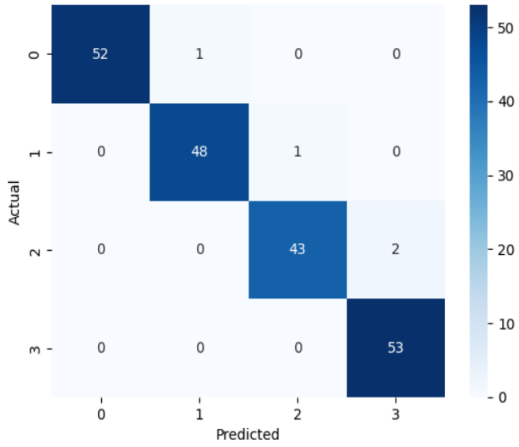
Berdasarkan Gbr. 4, terdapat 97% yang diprediksi benar oleh model dengan total 388 data. Gambar ini membandingkan nilai aktual dengan nilai prediksi, di mana terdapat kecocokan yang konsisten antara keduanya. Sebagai contoh, pada baris pertama, perangkat dengan daya baterai sebesar 601, tinggi piksel 356, lebar piksel 765, dan RAM 532 diprediksi dengan benar sebagai kategori 0, sesuai dengan nilai aktualnya. Contoh-contoh serupa dalam gambar menunjukkan bahwa model klasifikasi yang digunakan berfungsi dengan baik pada dataset ini.

	battery_power	px_height	px_width	ram	Actual	Predicted	Correct
1890	1991	298	1037	1861	1	2	False
1147	1851	293	1967	735	0	1	False
538	704	251	1013	3653	2	3	False
824	1722	1179	1638	2376	3	2	False
733	720	1347	1733	2799	3	2	False
819	1236	517	809	1406	0	1	False
1962	635	599	1299	2452	2	1	False
1164	860	1573	1581	2832	2	3	False
266	1876	546	1564	2513	3	2	False
1547	1611	163	1011	3078	2	3	False
1418	908	154	941	3518	3	2	False
1437	1933	229	1473	838	0	1	False

Gbr. 5 *Prediksi Salah Data Split* 80:20

Gbr. 5 menunjukkan dua belas data yang diprediksi salah oleh model ketika metode pembagian data 80:20 digunakan. Data ini memperlihatkan perbedaan antara nilai aktual dan prediksi, yang menunjukkan adanya kesalahan dalam prediksi model. Sebagai contoh, pada baris pertama, perangkat dengan daya baterai 1991, tinggi piksel 298, lebar piksel 1037, dan RAM 1861 memiliki nilai aktual 1 tetapi diprediksi sebagai 2 oleh model. Kesalahan seperti ini penting untuk diidentifikasi dan dianalisis lebih lanjut untuk meningkatkan akurasi model dalam memprediksi kategori yang benar.

Untuk perbandingan *data split* 90:10, sebanyak 90% dari total data digunakan untuk data latih dan 10% sisanya digunakan untuk data uji. Jika pada perbandingan sebelumnya mendapatkan hasil tingkat akurasi sebesar 97%, pada perbandingan ini, hasil akurasi yang didapatkan adalah 0.98 atau tingkat keberhasilan dalam memprediksi harga ponsel sebesar 98%.



Gbr. 6 *Confusion Matrix* 90:10

Pada Gbr. 6, *confusion matrix* menyajikan hasil bahwa pada kelas 0, dari 53 sampel, 52 sampel diprediksi benar sebagai kelas 0 dan terdapat 1 sampel yang diprediksi sebagai kelas 1. Pada kelas 1, dari banyaknya 49 sampel, sebanyak 48 sampel yang diprediksi benar dan sisanya yaitu 1 salah yang diprediksi sebagai kelas 2. Pada kelas 2, terdapat 45 sampel dimana 43 sampel diantaranya diprediksi benar dan 2 sampel salah diprediksi sebagai kelas 3. Kemudian, dari banyaknya 53 sampel yang terdapat pada kelas 3, semuanya diprediksi benar.

	precision	recall	f1-score	support
0	1.00	0.98	0.99	53
1	0.98	0.98	0.98	49
2	0.98	0.96	0.97	45
3	0.96	1.00	0.98	53
accuracy			0.98	200
macro avg	0.98	0.98	0.98	200
weighted avg	0.98	0.98	0.98	200

Gbr. 7 *Classification Report* 90:10

Pada Gbr. 7, ditunjukkan hasil bahwa model *logistic regression* pada perbandingan ini menunjukkan kinerja yang sangat baik dengan akurasi 98%, secara konsisten memprediksi kategori harga ponsel dari data pengujian. Meskipun terdapat variasi dalam presisi, *recall*, dan *f1-score* di berbagai kelas, tingkat konsistensi menunjukkan bahwa model ini memberikan hasil yang dapat diandalkan dan stabil. Nilai presisi, *recall*, dan *f1-score* yang tinggi membuktikan bahwa kemampuan *logistic regression* pada perbandingan 90:10 untuk mengklasifikasikan kategori harga sangat akurat dengan tingkat kesalahan yang rendah.

Berikut merupakan gambar yang menunjukkan data prediksi yang benar dan salah untuk pembagian persentase *data split* 90:10.

	battery_power	px_height	px_width	ram	Actual	Predicted	Correct
256	601	356	765	532	0	0	True
352	1604	134	939	916	0	0	True
298	928	221	1243	666	0	0	True
581	1512	1079	1897	3607	3	3	True
1288	1541	796	1052	1108	1	1	True
...	...	...	...	...	...	...	...
1865	1748	718	1109	2633	2	2	True
1034	1949	951	1178	356	0	0	True
1937	1396	560	1177	2694	2	2	True
1108	808	526	1324	3431	3	3	True
746	1884	451	819	3619	3	3	True

Gbr. 8 *Prediksi Benar Data Split* 90:10

Gbr. 8 menampilkan 98% data yang diprediksi benar oleh model, yaitu sebanyak 196 data. Nilai aktual dan nilai prediksi dibandingkan, yang menunjukkan kecocokan yang konsisten antara keduanya. Sebagai contoh, berdasarkan nilai aktualnya, perangkat dengan daya baterai 1604, lebar piksel 939, tinggi piksel 134, dan RAM 916 diprediksi dengan benar sebagai kategori 0 pada baris pertama. Tabel menunjukkan bahwa model klasifikasi yang digunakan berfungsi dengan baik pada dataset ini.

	battery_power	px_height	px_width	ram	Actual	Predicted	Correct
1890	1991	298	1037	1861	1	2	False
745	894	286	1300	3377	2	3	False
1147	1851	293	1967	735	0	1	False
538	704	251	1013	3653	2	3	False

Gbr. 9 *Prediksi Salah Data Split* 90:10

Gbr. 9 menunjukkan bahwa model yang menggunakan metode pembagian 90:10 membuat prediksi salah untuk empat data. Perbedaan antara nilai aktual dan prediksi menunjukkan kesalahan prediksi model. Sebagai contoh, baris pertama menunjukkan bahwa perangkat dengan daya baterai 1991 memiliki nilai aktual 1 tetapi diprediksi 2 oleh model, serta lebar piksel 1037, tinggi piksel 298, dan RAM 1861. Kesalahan seperti ini harus ditemukan dan dianalisis lebih lanjut untuk meningkatkan akurasi model dalam memilih kategori yang tepat.

Dengan akurasi masing-masing 97% dan 98%, teknik *data split* (80:20 dan 90:10) menghasilkan model *logistic regression* yang sangat akurat untuk memprediksi harga ponsel. Model dengan *data split* 90:10 sedikit lebih akurat, meskipun tidak ada signifikansi statistik dalam perbedaan akurasinya. Nilai presisi, *recall*, dan *f1-score* yang tinggi merupakan indikasi kinerja yang baik dalam *confusion matrix* dan laporan klasifikasi dari kedua perbandingan. Meskipun area yang perlu ditingkatkan diidentifikasi melalui analisis prediksi benar dan salah, model *logistic regression* berkinerja baik secara keseluruhan dalam kedua skenario pemisahan data [17].

IV. KESIMPULAN DAN SARAN

Dalam penelitian ini, model *logistic regression* berhasil digunakan untuk mengkategorikan harga ponsel menjadi empat kategori atau tingkatan: 0 (rendah), 1 (sedang), 2 (tinggi), dan 3 (sangat tinggi). Evaluasi model dilakukan dengan dua pendekatan berbeda dalam *data split* untuk data , yaitu dengan perbandingan 80:20 dan 90:10. Pada perbandingan data 80:20, model menunjukkan tingkat akurasi sebesar 97%, dengan nilai *precision*, *recall*, dan *f1-score* yang tinggi untuk setiap kelas. Hal ini menunjukkan bahwa model mampu melakukan prediksi dengan cukup akurat dan memiliki keseimbangan yang baik antara kemampuan mendeteksi kelas yang benar dan menghindari kesalahan prediksi. Untuk perbandingan *data split* 90:10 menghasilkan tingkat akurasi sebesar 98%. Tingkat akurasi ini sangat tinggi, menunjukkan bahwa model mampu memprediksi dengan benar sebagian besar data uji dengan kesalahan yang sangat sedikit. Nilai *precision* dan *recall* yang mendekati 1 untuk setiap kategori harga juga menunjukkan bahwa model mampu mengidentifikasi dengan tepat dan lengkap kategori harga ponsel yang sebenarnya dalam data pengujian. Hasil ini menunjukkan bahwa model *logistic regression* adalah model yang efektif dan akurat untuk mengklasifikasikan harga ponsel ke dalam kategori yang telah ditentukan. Terutama penggunaan *data split* 90:10 memberikan hasil yang sangat baik, karena menghasilkan tingkat akurasi yang hampir sempurna. Dengan demikian, penelitian ini memberikan kontribusi yang signifikan dalam pemahaman dan penerapan model *logistic regression* dalam klasifikasi harga ponsel.

Sebagai saran untuk penelitian selanjutnya, disarankan untuk mencoba beberapa metode lain, seperti *decision tree*, *random forest*, atau SVM (*Support Vector Machine*). Selain itu, pengaturan *hyperparameter* juga dapat dioptimalkan untuk mendapatkan hasil yang lebih baik. Penentuan perbandingan *data split* yang berbeda guna menguji konsistensi dan stabilitas model dalam berbagai skenario data juga dapat dipertimbangkan untuk penelitian selanjutnya.

V. DAFTAR PUSTAKA

[1] A. Arisusanto, N. Suarna, And G. Dwilesatari, “Analisa Klasifikasi Data Harga Handphone Menggunakan Algortima Random Forest Dengan Optimize Parameter Grid,” Jurnal Teknologi Ilmu Komputer., Vol. 1, No. 2, Pp. 43–47, 2023.

[2] A. P. Sari, H. Suzuki, T. Kitajima, T. Yasuno, and D. A. Prasetya, “PREDICTION MODEL OF WIND SPEED AND DIRECTION USING DEEP NEURAL NETWORK,” JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science), vol. 3, no. 1, Feb. 2020.

[3] P. Egipta, P. E. Sudira, S. C. Olivia, O. Sandra, and D. Marten, "Perbandingan Algoritma Machine Learning menggunakan Orange Data Mining untuk Klasifikasi Jenis Kendaraan pada Sistem Tilang Digital" Jurnal Elektro, Vol. 17, No 1. pp. 41-47, April, 2024.

[4] A. Avini, K. W. Patunduk, S. Sumarni, H. Harbianti, A. Pratiwi, and R. Hidayat, “Analisis Model Cox Proportional Hazard dan Regresi Logistik sebagai Upaya Pencegahan Covid-19 di Kota Palopo,” Inferensi, vol. 5, no. 2, pp. 105–114, Sep. 2022.

[5] Y. Tampil, H. Komaliq, and Y. Langi, “Analisis Regresi Logistik untuk menentukan Faktor-Faktor yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado,” D’Cartesian: Jurnal Matematika Dan Aplikasi/D ’ Cartesian, vol. 6, no. 2, p. 56, Aug. 2017.

[6] D. U. Setyawati, B. D. Korida, and B. R. A. Febrilia, “Analisis Regresi Logistik Ordinal Faktor-Faktor yang Mempengaruhi IPK Mahasiswa,” Jurnal Varian, vol. 3, no. 2, pp. 65–72, May 2020.

[7] M. I. Gunawan, D. Sugiarto, and I. Mardianto, “Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression,” Jurnal Edukasi dan Penelitian Informatika (JEPIN), vol. 6, no. 3, p. 280, Dec. 2020.

[8] N. B. N. Azmi, N. A. Hermawan, and N. D. Avianto, “Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan

PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," JTIM : Jurnal Teknologi Informasi Dan Multimedia/Jurnal Teknologi Informasi Dan Multimedia, vol. 4, no. 4, pp. 281–290, Feb. 2023.

[9] "Mobile Price Classification," www.kaggle.com. https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/data

[10] M. R. Irianto, A. Maududie, and F. N. Arifin, "Implementation of K-Means Clustering Method for Trend Analysis of Thesis Topics (Case Study: Faculty of Computer Science, University of Jember)," BERKALA SAINSTEK, vol. 10, no. 4, Pp. 210–226, Dec. 2022.

[11] S. Perveen, M. A. Khalid, and O. Ahsan, "CORRELATION OF SERUM URIC ACID LEVELS WITH MODIFIED RANKIN SCORE IN PATIENTS WITH ACUTE ISCHEMIC STROKE," Pakistan Armed Forces Medical Journal, vol. 69, no. 6, pp. 1199-1203, 2019.

[12] K. Stewart, "Pearson's correlation coefficient | Definition, Formula, & Facts," Encyclopedia Britannica, May 08, 2024. https://www.britannica.com/topic/Pearsons-correlation-coefficient

[13] W. Musu, A. Ibrahim, and H. Heriadi, "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5," SISITI : Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi, vol. 10, no. 1, pp. 186–195, Mar. 2021.

[14] D. B. Lanini, S. U. Rahmi, and M. F. Siddiq, "Klasifikasi Harga Ponsel dengan Feature Selection Menggunakan Metode Machine Learning", Proceeding KONIK (Konferensi Nasional Ilmu Komputer), vol. 6, pp. 049–053, 2023.

[15] I. M. M. Matin, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," MULTINETICS: Jurnal Multimedia Networking Informatics, vol. 9, no. 1, pp. 43–50, May 2023.

[16] Suhliyyah, H. H. Hikmayanti, and B. K. Ahmad, "Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke", Jurnal Informatika, vol. 12, no. 01, pp. 15-23, 2023.

[17] E. S. I. Aksoy, and S. Murat, "Mobile Phone Price Classification Using Machine Learning", International Journal of Advanced Natural Sciences and Engineering Researches (IJANSER), vol. 7, no. 4, pp. 458-462, 2023.