Pengaruh Normalisasi Data pada Klasifikasi Harga Ponsel Berdasarkan Spesifikasi Menggunakan Klasifikasi *Naive Bayes* dan *Multinomial Logistic Regression*

Ahmad Karim A.I[1], Fuad Nurhadi[2], I Ketut Okta Setiawan[3], Ichlasul Akmali Rizky[4], Rischantika Br. Manurung[5]
[1]Teknik Elektro, Fakultas Teknik
Universitas Sriwijaya
Palembang, Indonesia
Penulis Korespondensi: iakarimar03@gmail.com

***Abstract*—Mobile phone is an electronic device that is needed in today's technological advances. Mobile phones can be used for various things such as long-distance communication, entertainment facilities, as well as information service media. Along with the times, mobile phone companies are competing in producing mobile phones with various features and specifications due to the large number of customer requests. When buying a cellphone, every customer must have their own tastes and will consider the price of a cellphone based on its specifications. Cellphone price categorization can be done using machine learning algorithms such as Multinomial Logistic Regression and Naïve Bayes. However, not all machine learning algorithms can create direct classification models (using raw data) with high accuracy because the scale of each feature is not the same in predicting predictor variables. That's when a data preprocessing feature or data normalization is needed in the formation of a classification model. When the data is not normalized, the results of the evaluation of the Naive Bayes model show an accuracy value of 79.8% while the evaluation results of the Multinomial Logistic Regression model have an accuracy value of 63.5%. When the data is normalized, the results of the Naive Bayes evaluation increased to 80%, while the evaluation results of the Multinomial Logistic Regression model show a significant increase, namely to 95.8%.**

Keywords—mobile phone, Naïve Bayes, Multinomial Logistic Regression, classification, normalization

***Abstrak*—*Handphone* merupakan alat elektronik yang sangat dibutuhkan dalam kemajuan teknologi saat ini. *Handphone* dapat digunakan untuk berbagai macam hal seperti berkomunikasi jarak jauh, sarana hiburan, serta sebagai media layanan informasi. Seiring dengan perkembangan zaman, perusahaan-perusahaan *handphone* berlomba-lomba dalam memproduksi *handphone* dengan berbagai macam fitur beserta spesifikasi karena banyaknya permintaan pelanggan. Dalam pembelian *handphone*, setiap pelanggan pasti memiliki seleranya masing-masing dan pasti akan mempertimbangkan harga *handphone* berdasarkan spesifikasinya. Pengkategorian harga *handphone* dapat dilakukan dengan algoritma *machine learning* seperti *Multinomial Logistic Regression* dan Naïve Bayes. Tetapi, tidak semua algoritma *machine learning* dapat membuat model klasifikasi langsung (menggunakan data mentah) dengan akurasi yang tinggi karena skala masing-masing fitur belum sama dalam memprediksi variabel prediktor. Karena itulah dibutuhkan sebuah fitur *data preprocessing* atau *normalisasi data* dalam pembentukan suatu model klasifikasi. Ketika data tidak dinormalisasi, hasil evaluasi model *Naïve Bayes* menunjukkan nilai akurasi 79.8% sedangkan hasil evaluasi model *Multinomial Logistic Regression* memiliki nilai akurasi 63.5%. Ketika data dinormalisasi, hasil evaluasi *Naïve Bayes* naik menjadi 80%, sedangkan hasil evaluasi model *Multinomial Logistic Regression* menunjukkan kenaikan yang signifikan yaitu menjadi 95.8%.**

Kata kunci—handphone, Naïve Bayes, Multinomial Logistic Regression, klasifikasi, normalisasi

I. PENDAHULUAN

Handphone merupakan alat komunikasi yang banyak digunakan saat ini dimana hampir setiap orang saat ini menggunakannya untuk keperluan sehari-hari seperti komunikasi dan layanan informasi. Saat ini, ponsel memiliki fitur yang semakin beragam, semakin majunya teknologi, semakin banyak pula ponsel dengan fungsi dan spesifikasi yang lebih baik [1]. Fitur bawaan ponsel seperti resolusi kamera, bentuk lebih ramping, layar lebih lebar, serta kapasitas memori yang lebih besar dan lain sebagainya sangat menentukan harga sebuah ponsel.

Harga adalah nilai tukar yang paling efektif digunakan dalam pemasaran dan bisnis. Harga adalah tolak ukur pertama saat membeli atau menjual. Pelanggan yang ingin membeli telepon seluler pasti mempertimbangkan

harga telepon seluler yang akan dibelinya berdasarkan spesifikasinya. Oleh karena itu, Pengklasifikasian harga ponsel sangat diperlukan sebagai referensi pelanggan *handphone* dalam menentukan apakah *handphone* tersebut sesuai dengan keinginan pelanggan atau tidak [2]. Keuntungan lain dalam melakukan pengklasifikasian harga *handphone* yaitu pelanggan dapat memilah apakah handphone tersebut layak digunakan pada kemajuan teknologi saat ini mengingat beragam macam spesifikasi *handphone* yang ada pada saat ini.

Pembelajaran mesin atau *machine learning* menawarkan kepada kita teknik *artificial intelligent* terbaik seperti klasifikasi, klasterisasi, regresi, dan lain sebagainya. Pembuatan model klasifikasi harga ponsel pada penelitian ini menggunakan metode *Naïve Bayes* dan *Multinomial Logistic Regression*. *Naïve Bayes* yaitu metode klasifikasi yang menghitung probabilitas dengan mengkalkulasikan kombinasi frekuensi dan nilai dalam kumpulan data yang diberikan untuk menentukan probabilitas suatu hasil [3]. Regresi logistik multinomial berguna untuk menjelaskan hubungan antara suatu fitur (biasanya dalam bentuk variabel independen) dengan variabel dependen (variabel prediktor) yang lebih dari 2 dengan menggunakan hubungan peluang dengan fungsi eksponensial dan logaritma.

Pembagian data dalam menentukan suatu variabel prediktor dapat berupa berbagai macam bentuk seperti bentuk gambar atau citra, angka, biner (0 atau 1), bahkan dalam bentuk suara (amplitudo, frekuensi, periode). Fitur-fitur tersebut digunakan untuk menentukan variabel prediktor. Tetapi, seringkali hasil yang didapat tidak sesuai dengan hasil prediksi sebenarnya karena pada masing-masing variabel independennya tidak memiliki suatu skala yang sama. Perbedaan skala tersebut sangat berpengaruh dalam penentuan suatu hasil prediksi dalam pembentukan suatu model *machine learning.* Oleh karena itu, dibutuhkan suatu normalisasi data agar akurasi yang dihasilkan lebih baik.

Penelitian serupa pernah dilakukan oleh Vanissa [4] yaitu klasifikasi harga *handphone* menggunakan metode *Random Forest* dengan jumlah data yang sama*.* Vanissa hanya menggunakan 7 fitur yang terdapat pada spesifikasi *handphone,* sedangkan pada penelitian ini menggunakan sebanyak 20 fitur. Selain itu, pada penelitian tersebut tidak menjelaskan apakah pembuatan model klasifikasinya menggunakan normalisasi atau tidak, sedangkan pada penelitian ini akan membandingkan evaluasi hasil model klasifikasi berdasarkan data dinormalisasi atau tidak. Penelitian kali ini menggunakan metode klasifikasi terawasi *Naïve Bayes* dan *Multinomial Logistic Regression* karena kedua metode ini memiliki kesamaan dalam perhitungan yaitu menggunakan konsep peluang. *Naïve Bayes* menggunakan metode probabilitas yang sederhana sedangkan pada *Multinomial Logistic Regression* menggunakan perhitungan peluang menggunakan logaritma dan eksponensial dalam memprediksi variabel prediktor. Keuntungan dari menggunakan kedua metode ini dibandingkan metode lainnya yaitu kedua metode ini cenderung memiliki *size model* yang lebih kecil serta waktu pelatihan yang cukup cepat dalam pembuatan model klasifikasi dari data numerik dibandingkan metode lain seperti *Artificial Neural Network* yang memiliki *size model* yang besar serta waktu pelatihan yang cenderung lama.

Tujuan penelitian ini adalah membandingkan hasil pengaruh data yang belum dinormalisasi dengan data yag telah dinormalisasi menggunakan metode *Multinomial Logistic Regression* dan juga metode *Naive Bayes*. Dalam memperkirakan spesifikasi harga ponsel dengan menggunakan kedua metode tersebut, diharapkan peneliti dapat mengetahui bagaimana pengaruh normalisasi data dalam pembuatan model klasifikasi.

II. TINJAUAN PUSTAKA

A. Normalisasi Data

Normalisasi data biasa disebut dengan proses *preprocessing.* Proses ini mengubah suatu bentuk data mentah menjadi bentuk data yang lebih pantas untuk dibuat suatu model klasifikasi [5]. Proses *preprocessing* ini dapat menjadikan sebuah data memiliki skala *impact* pada masing-masing variabel independen yang hampir sama pada saat memprediksi variabel dependen berdasarkan hasil normalisasi masing-masing variabel independen.

Tahap pertama untuk melakukan normalisasi data adalah menghitung rata-rata masing masing kolom variabel fitur dengan menggunakan persamaan 1. Nilai $\bar{x}_b$ merupakan rata-rata fitur pada kolom ke-b, $\sum x_{nb}$ adalah jumlah semua data pada fitur ke b dengan banyak data n.

$$\bar{x}_b = \frac{\sum x_{nb}}{n} \tag{1}$$

Tahap selanjutnya yaitu penghitungan standar deviasi pada masing-masing fitur yang dapat dinyatakan dengan menggunakan persamaan 2. s_b yang merupakan standar deviasi pada fitur ke-b. Nilai x_{nb} merupakan nilai pada baris ke-n dan kolom atau fitur ke-b.

$$s_b = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{nb} - \bar{x}_b)^2} \tag{2}$$

Langkah akhir untuk melakukan normalisasi data adalah menggantikan semua data dengan nilai hasil normalisasi yang dapat dinyatakan dengan persamaan (3). Nilai Z_{nb}merupakan nilai hasil normalisasi data yang terletak pada baris ke-n kolom dan fitur ke-b.

$$Z_{nb} = \frac{x_{nb} - \bar{x}_b}{s_b} \tag{3}$$

Apabila diasumsikan sebuah *dataset* yang memiliki banyak data n dengan banyak fitur b (dengan syarat semua fitur merupakan variabel independen), serta nilai pada data ke-n dan fitur ke-b adalah x_{nb} maka data tersebut dapat kita representasikan seperti Tabel 1.

TABLE I. *DATASET*

	Fitur 1	**Fitur 2**	**Fitur 3**	**…**	**Fitur b**
1.	x_{11}	x_{12}	x_{13}	…	x_{1b}
2.	x_{21}	x_{22}	x_{23}	…	x_{2b}
3.	x_{31}	x_{32}	x_{33}	…	x_{3b}
…	…	…	…	…	…
n	x_{n1}	x_{n2}	x_{n3}	…	x_{nb}

Nilai x_{nb} dapat dinormalisasikan menjadi nilai Z_{nb} menggunakan persamaan sehingga data mengalami transformasi yang dapat dinyatakan oleh Tabel 2. Data hasil normalisasi inilah yang sering digunakan untuk pelatihan model klasifikasi data karena persebaran nilai pada masing-masing fitur memiliki skala yang sama dalam penentuan suatu variabel prediktor.

TABLE II. TRANSFORMASI *DATASET*

	Fitur 1	**Fitur 2**	**Fitur 3**	**…**	**Fitur b**
1.	Z_{11}	Z_{12}	Z_{13}	…	Z_{1b}
2.	Z_{21}	Z_{22}	Z_{23}	…	Z_{2b}
3.	Z_{31}	Z_{32}	Z_{33}	…	Z_{3b}
…	…	…	…	…	…
n	Z_{n1}	Z_{n2}	Z_{n3}	…	Z_{nb}

B. Klasifikasi Naïve Bayes

Naïve Bayes adalah salah satu metode pengklasifikasian secara terawasi yang cocok digunakan untuk data numerik. *Naïve Bayes* sendiri digunakan untuk proses klasifikasi data dengan metode prediksi probabilitas dan statistik sederhana [6]. Metode probabilitas dan statistik yang digunakan dikenal sebagai Teorema Bayes, metode ini melakukan suatu prediksi berdasarkan probabilitas kedepannya berdasarkan pengalaman sebelumnya

Naïve Bayes sendiri memiliki kelebihan. Kelebihan tersebut antara lain kecepatan dalam perhitungan dan algoritma yang sederhana sekaligus berakurasi tinggi. Hal ini dikarenakan metode ini dapat digunakan untuk mengklasifikasikan suatu data numerik yang sangat banyak dan sangat tinggi serta memiiki model dengan *size* yang cukup kecil untuk proses pengklasifikasian dalam memprediksi variabel prediktor yang ditentukan. Ditambah dengan alur perhitungan yang tidak terlalu panjang menjadikan metode ini lebih mudah untuk digunakan [7]. Persamaan 4 menunjukkan perhitungan probabilitas menggunakan metode *Naïve Bayes*

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \tag{4}$$

$P(Y|X)$ = Probabilitas suatu kelas variabel independen terhadap suatu variabel tujuan (dependen)

$P(Y)$ = Probabilitas suatu kelas variabel dependen

$P(X)$ = Probabilitas suatu kelas variabel independen

$\prod_{i=1}^{q} P(X_i|Y)$ = Perkalian seluruh probabilitas independent kelas Y dari semua fitur variabel X

C. Klasifikasi Multinomial Logistic Regression

Regresi Logistik *(Logistic Regression)* adalah metode analisis statistik yang menggambarkan hubungan antara variabel respon yang pada dasarnya kualitatif (variabel dependen) dengan dua atau lebih kategori dengan satu atau lebih variabel penjelas (variabel independen) dengan menggunakan fungsi logaritma pada skala kategoris atau interval [8]. Kemudian menggunakan hubungan antara dua faktor data tersebut untuk memprediksi nilai dari salah satu faktor tersebut berdasarkan faktor yang lain. Prediksi tersebut biasanya memiliki jumlah hasil yang

terbatas, seperti ya atau tidak. Regresi Logistik ini sendiri merupakan regresi nonlinier yang digunakan untuk menjelaskan hubungan non-linier antara suatu fitur dan variabel prediktor (dinyatakan dalam Y), anomali dalam distribusi Y, dan juga variabilitas respons non-konstan yang tidak dapat dipertanggungjawabkan oleh model regresi linier biasa. Tujuan metode Regresi Logistik ini sendiri untuk mendapatkan model yang bagus sekaligus sederhana yang menjelaskan secara singkat variabel respons dengan variabel prediktor [9].

Regresi Logistik biasa hanya digunakan memprediksi variabel prediktor dengan 2 kategori (seperti iya atau tidak, 1 atau 0). Untuk memprediksikan variabel prediktor yang memiliki kategori yang lebih dari 2 dapat menggunakan metode *Multinomial Logistic Regression*. Adapun persamaan dasar yang digunakan untuk perhitungan metode *Multinomial Logistic Regression* adalah pada persamaan 5. Pr adalah probabilitas ketika nilai variabel prediktor adalah ke-k dengan b_k adalah himpunan koefisien regressi yang berhubungan dengan prediksi ke-k, serta x_i adalah himpunan variabel eksplanatori yang berhubungan dengan observasi i.

$$\Pr(Y_z = k) = \frac{e^{b_k x_i}}{1 + \sum_j^{k-1} e^{b_j x_i}} \quad (5)$$

III. METODOLOGI PENELITIAN

A. Pengumpulan dan Pembagian Variabel Data

Dataset yang didapatkan bersumber dari website kaggle [10] yang berisikan 2000 data spesifikasi ponsel beserta rentang harga berdasarkan spesifikasinya. Terdapat 20 fitur pada dataset ini antara lain battery_power (kapasitas baterai dalam satuan mAh), blue (mempunyai *bluetooth* atau tidak), clock_speed (kecepatan mikroprosesor), dual_sim (mempunyai fitur *dual sim* atau tidak), fc (resolusi kamera depan), four_g (mempunyai fitur 4G atau tidak), int_memory (kapasitas memori internal), m_dep (ketebalan *handphone dalam satuan* cm), mobile_wt (lebar *handphone*), n_cores (jumlah *core* pada prosesor), pc (resolusi kamera belakang), px_height (resolusi ketinggian layar), px_width (resolusi lebar layar), ram (kapasitas RAM), sc_h (ketinggian *handphone* dalam satuan cm), sc_w (lebar *handphone*), talk_time (waktu terlama untuk menghabiskan seluruh kapasitas baterai), three_g (mempunyai fitur 3G atau tidak), touch_screen (memiliki fitur layar sentuh atau tidak), dan wifi (mempunyai fitur koneksi *wireless* atau tidak). Semua fitur tersebut merupakan variabel independen yang akan menentukan variabel price_range (rentang harga *handphone*) dalam bentuk numerik yaitu dinyatakan dalam angka 0 bila dikategorikan sebagai murah, angka 1 bila sedang, angka 2 bila mahal, dan angka 3 bila sangat mahal.

B. Data Split

Pada pembentukan model klasifikasi pada kedua metode, digunakan data *test* dan data *train* yang sama. Jumlah data *test* yang digunakan yaitu sebanyak 600 data dengan data *train* sebanyak 1400 data. Data tersebut diambil secara acak. Pada pembuatan model klasifikasi tanpa normalisasi menggunakan data seperti yang terlihat pada Tabel 3 dan 4.

TABLE III. DATA *TEST* MENTAH

	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
1860.	1646	0	2.5	0	…	11	1	1	0
353.	1182	0	0.5	1	…	19	1	0	0
1333.	1972	0	2.9	1	…	8	1	1	0
905.	989	1	2.0	0	…	19	1	1	0
…	…	…	…	…	…	…	…	…	…

TABLE IV. DATA *TRAIN* MENTAH

	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
836.	902	1	0.6	1	…	15	0	1	1
575.	1197	1	0.5	1	…	14	1	1	1
557.	1519	0	2.1	0	…	15	1	1	0
1235.	1971	1	0.5	1	…	17	1	1	0
…	…	…	…	…	…	…	,,,	…	…

Pada pembuatan model klasifikasi menggunakan normalisasi, data yang sama ditransformasikan sehingga data *test* dan data *train* akan berubah seperti pada Tabel 5 dan 6.

TABLE V. DATA *TEST* TELAH DINORMALISASI

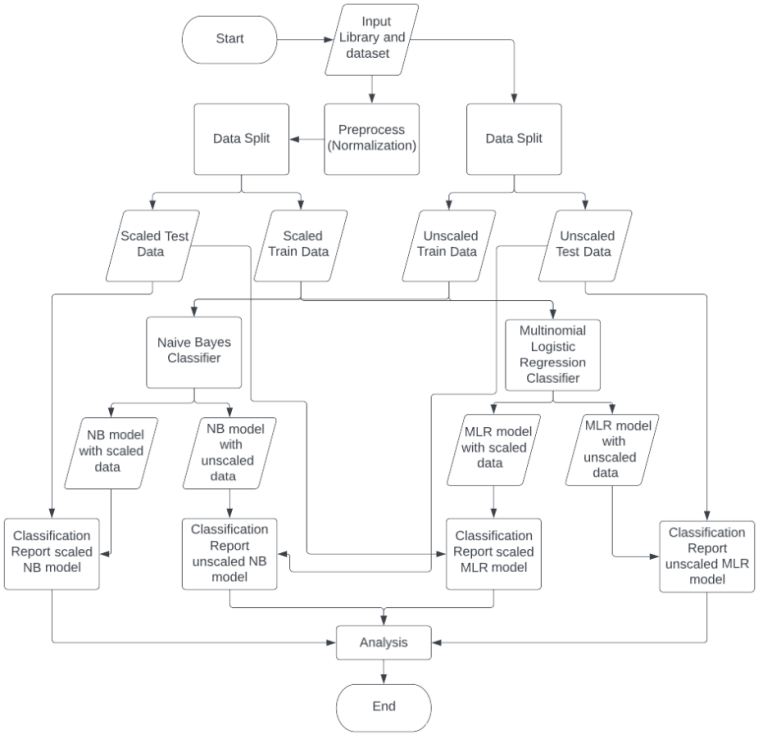
	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
1860.	0.9275	-0.9900	1.1985	-1.0191	…	-0.0020	0.5596	0.9940	-1.0140
353.	-0.1286	-0.9900	-1.2530	-1.0191	…	2.5791	0.5596	-1.0060	-1.0140
1333.	1.6696	-0.9900	1.6888	-1.0191	…	-1.0945	0.5596	0.9940	-1.0140
905.	-0.5679	1.0100	0.5856	-1.0191	…	0.2831	0.5596	0.9940	-1.0140
…	…	…	…	…	…	…	…	…	…

TABLE VI. DATA *TRAIN* TELAH DINORMALISASI

	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
836.	-0.7660	1.0100	-1.1304	0.9811	…	0.7302	-1.7868	0.9940	0.9860
575.	-0.0945	1.0100	1.2530	0.9811	…	0.5471	0.5596	0.9940	0.9860
557.	0.6384	-0.9900	0.7082	-1.0191	…	0.7302	0.5596	0.9940	-1.0140
1235.	1.6673	1.0100	-1.2530	0.9811	…	1.0963	0.5596	0.9940	-1.0140
…	…	…	…	…	…	…	…	…	…

C. Langkah Analisis

Langkah yang dilakukan pertama kali agar dapat melakukan analisis yaitu memasukkan *library python* beserta *dataset* ke dalam aplikasi Google Colab. *Dataset* tersebut akan mengalami proses normalisasi atau *preprocessing.* Terdapat juga *dataset* yang tidak mengalami normalisasi agar kita dapat mengetahui analisis dari data yang mengalami normalisasi dan data yang tidak mengalami normalisasi. Masing-masing *dataset* akan dibagi atau *split* menjadi 2 bagian yaitu data *train* dan data *test.* Seluruh data *train* akan mengalami proses pengklasifikasian berdasarkan modelnya. Peneliti menggunakan klasifikasi Naïve-Bayes dengan *Multinomial Logistic Regression* dalam metode pengklasifikasian. Proses *training data* akan menghasilkan suatu model klasifikasi. Model klasifikasi akan dievaluasi dengan cara membandingkan prediksi variabel price_range berdasarkan model klasifikasi dengan hasil price_range yang telah tertera pada data test. Gambar 1 menunjukkan diagram alir yang berisikan tahapan agar dapat menganalisis model klasifikasi.



Gambar. 1. Diagram alir penelitian

IV. HASIL DAN PEMBAHASAN

A. Identifikasi Data

1) Data Un-Preprocessed

Tabel 7 menunjukkan tampilan 5 data *train*ing awal yang akan dilakukan proses pelatihan.

TABLE VII. DATA *TRAIN*

	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
1.	842	0	2.2	0	…	19	0	0	1
2.	1021	1	0.5	1	…	7	1	1	0
3.	563	1	0.5	1	…	9	1	1	0
4.	615	1	2.5	0	…	11	1	0	0
5.	1821	1	1.2	0	…	15	1	1	0

2) Deskripsi Data

Deskripsi data berguna untuk melihat bagaimana representasi masing-masing fitur untuk menentukan nilai normalisasi pada saat men-*train* data untuk membuat suatu model klasifikasi. Fitur yang paling penting untuk digunakan pada saat proses normalisasi adala std (standar deviasi) dan mean (rata-rata) seperti yang tertera pada Tabel 8.

TABLE VIII. DESKRIPSI DATA

	battery_power	blue	clock_speed	dual_sim	…	three_g	touch_screen	wifi	price_range
mean	1238.518	0.495	1.522	0.509	…	0.762	0.503	0.507	1.500
std	439.418	0.500	0.816	0.500	…	0.426	0.500	0.500	1.118
min	501	0	0.5	0	…	0	0	0	0
max	1998	1	3	1	…	1	1	1	3

3) Data Preprocessed

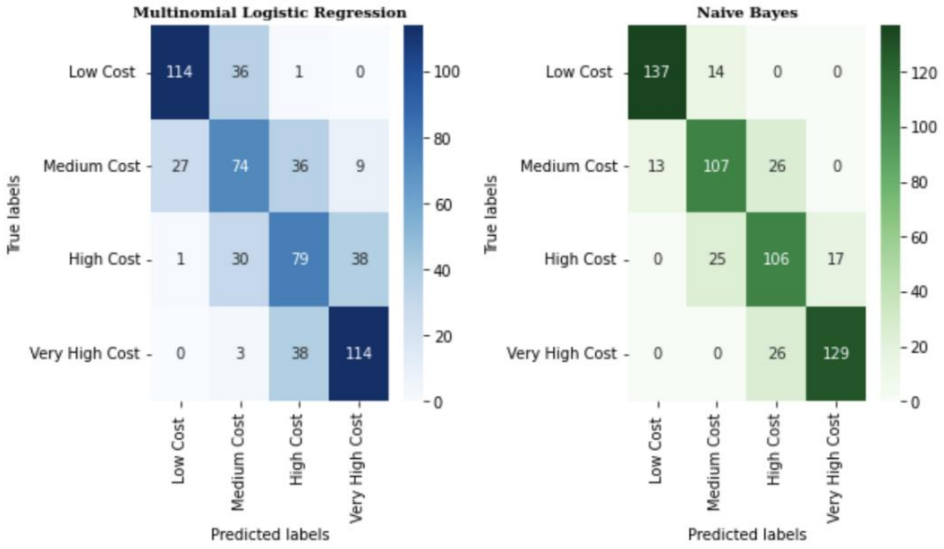
Data yang telah dinormalisasi dari 5 data *testing* awal dapat dilihat pada Tabel 9. Nilai-nilai pada masing-masing fitur terlihat lebih kecil dan lebih terstruktur dibandingkan data mentah sehingga skala masing-masing fitur untuk memprediksi rentang harga sama.

TABLE IX. NORMALISASI DATA

	battery_power	blue	clock_speed	dual_sim	…	talk_time	three_g	touch_screen	wifi
1.	-0.9025	-0.9900	0.8307	-1.0191	…	1.4624	-1.7868	-1.0060	0.9860
2.	-0.4951	1.0100	-1.2530	0.9811	…	-0.7342	0.5596	0.9940	-1.014
3.	-1.5376	1.0100	-1.2530	0.9811	…	-0.3681	0.5596	0.9940	-1.014
4.	-1.4193	1.0100	1.1985	-1.0191	…	-0.0020	0.5596	-1.0060	-1.014
5.	1.3259	1.0100	-0.3950	-1.0191	…	0.7302	0.5596	0.9940	-1.014

B. Confusion Matrix

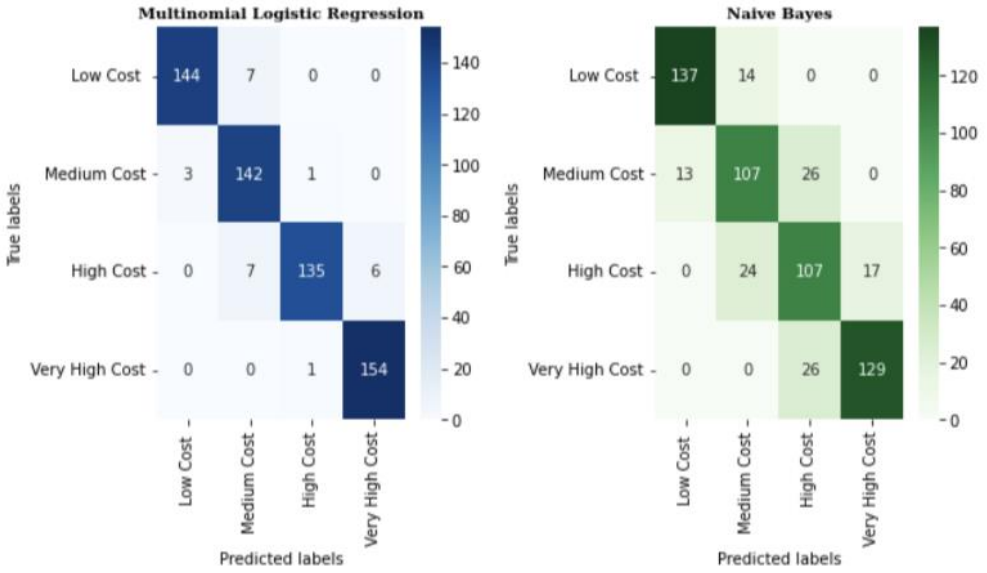
Ketidakakuratan hasil prediksi suatu model klasifikasi dapat divisualisasikan seperti pada gambar 2 dan gambar 3. Visualisasi dibentuk dari suatu bentuk matriks 4x4. Bagian baris merupakan hasil data test yang seharusnya, sedangkan bagian kolom merupakan hasil prediksi dari model klasifikasi yang telah dilatih. Berdasarkan Gambar 2, pada 151 data yang seharusnya menunjukkan harga rendah (*low cost*) terdapat 114 prediksi data yang benar (menunjukkan harga rendah), terdapat 36 prediksi menunjukkan harga sedang (*medium cost*), dan 1 prediksi menunjukkan harga mahal (*high cost*) pada metode MLR. Sedangkan pada metode *Naïve Bayes*, ketika hasil data menunjukkan ada 151 data yang memiliki harga rendah, terdapat 137 prediksi yang benar dan hanya 14 prediksi yang menunjukkan harga prediksi sedang. Untuk hasil *confusion matriks* mengenai rentang harga lainnya dapat dilihat pada gambar 2. Hasil *confusion matrix* yang didapatkan pada saat melatih model klasifikasi menggunakan data yang belum dinormalisasi menunjukkan bahwa model klasifikasi *Naïve Bayes* lebih akurat daripada model klasifikasi *Multinomial Logistic Regression*.



Gambar. 2. *Confusion Matrix* model klasifikasi menggunakan data mentah

Berdasarkan Gambar 3, hasil *confusion matrix* yang didapatkan pada model yang dilatih menggunakan data yang telah dinormalisasi menunjukkan bahwa pada saat 151 data yang seharusnya menunjukkan harga rendah, ternyata terdapat 144 prediksi benar (harga rendah) dan 7 prediksi harga sedang pada model MLR, sedangkan

pada model *Naïve Bayes* terdapat 137 prediksi benar dan 14 prediksi harga sedang. Untuk hasil *confusion matrix* pada rentang harga lainnya dapat dilihat pada gambar 3. Hasil *confusion matrix* yang didapatkan pada saat model klasifikasi dilatih menggunakan data yang telah dinormalisasi menunjukkan bahwa model *Multinomial Logistic Regression* lebih akurat dibandingkan dengan model *Naïve Bayes*.



Gambar. 3. *Confusion Matrix* model klasifikasi menggunakan data yang telah dinormalisasikan

C. Analisis Klasifikasi Data

Setiap model klasifikasi pasti memiliki perbedaan dalam hal akurasi (perbandingan antara prediksi yang benar terhadap prediksi yang sebenarnya). Akurasi dapat diukur dengan menggunakan persamaan (6). *Total True* merupakan total data yang memprediksi hasil yang benar (total data diagonal pada *confusion matrix*) dan *Total Predicted* adalah total data yang diprediksi.

$$Acc = \frac{Total\ True}{Total\ Predicted} \tag{6}$$

Untuk mengevaluasi proses klasifikasi, dapat kita perhitungkan dengan menggunakan persamaan (7) yang merupakan perhitungan *recall* dan persamaan (8) untuk menghitung *presicion. Recall* adalah ukuran perbandinganhasil prediksi yang benar terhadap nilai sebenarnya yang positif. *Presicion* merupakan ukuran ketepatan dengan cara membandingkan hasil prediksi yang benar terhadap hasil prediksi positif.

$$Rec = \frac{T_+}{T_+ + F_-} \tag{7}$$

$$Pres = \frac{T_+}{T_+ + F_+} \tag{8}$$

Kombinasi dari *presicion* dan *recall* dalam evaluasi data dapat menggunakan persamaan (9). F1-*score* digunakan untuk melihat nilai yang seimbang diantara nilai *recall* dan *akurasi.*

$$F1 = \frac{2(Rec * Pres)}{Rec + Pres} \tag{9}$$

1) Evaluasi Model Klasifikasi Data Mentah

Data testing yang digunakan untuk mengevaluasi model yang telah dilatih berjumlah 600 data dengan data yang digunakan untuk *training* sebanyak 1400 data. Dari 600 data *testing* didapatkan bahwa untuk rentang harga ponsel rendah sebanyak 151 data, rentang harga ponsel sedang sebanyak 146 data, rentang harga ponsel tinggi sebanyak 148 data, dan rentang harga ponsel tinggi sebanyak 155 data.

Berdasarkan hasil evaluasi pada Tabel 10 dan 11, hasil akurasi parameter *presicion, recall, f1-score,* dan *accuracy* pada model klasifikasi *Naïve Bayes* yang terlatih menggunakan data mentah menunjukkan nilai 0.798 atau 79.8% sedangkan hasil akurasi parameter pada model klasifikasi MLR menunjukkan nilai 0.635 atau 64.5%. Hasil *f1-score* pada perhitungan model *Naïve Bayes* menunjukkan bahwa hasil akurasi prediksi kategori yang benar dari tertinggi dan terendah adalah 0-3-1-2 (murah- sangat mahal – sedang-mahal), sedangkan pada model MLR adalah 0-3-2-1 (murah- sangat mahal- sedang -mahal).

TABLE X. PENGHITUNGAN AKURASI MODEL *NAÏVE BAYES* MENGGUNAKAN DATA MENTAH

	pres	rec	f1	supp
murah (0)	0.9133	0.9072	0.9102	151
sedang (1)	0.7328	0.7328	0.7328	146
mahal (2)	0.6708	0.7162	0.6938	148
sangat mahal (3)	0.8835	0.8322	0.8571	155
acc (4)	0.7983	0.7983	0.7983	0.798

TABLE XI. PENGHITUNGAN AKURASI MODEL *MULTINOMIAL LOGISTIC REGRESSION* MENGGUNAKAN DATA MENTAH

	pres	rec	f1	supp
murah (0)	0.8028	0.7549	0.7781	151
sedang (1)	0.5174	0.5068	0.5121	146
mahal (2)	0.5129	0.5337	0.5231	148
sangat mahal (3)	0.7080	0.7354	0.7215	155
acc (4)	0.6350	0.6350	0.6350	0.635

2) Evaluasi Model Klasifikasi Data Normalisasi

Berdasarkan Tabel 12 dan 13, didapatkan bahwa akurasi total yang didapatkan pada model klasifikasi *Naïve Bayes* yaitu sebesar 0.8 atau sekitar 80% sedangkan akurasi total yang didapatkan pada model klasifikasi MLR adalah sebesar 0.95 atau sekitar 95%. Hasil *f1-score* dalam memprediksi kategori rentang harga pada metode *Naïve Bayes* dari urutan yang tertinggi ke terendah adalah sama dengan hasil evaluasi model klasifikasi data mentah dengan kenaikan akurasi yang tidak terlalu signifikan, sedangkan hasil *f1-score* dalam memprediksi rentang harga pada metode MLR menunjukkan urutan dari yang tertinggi ke terendah sama dengan model klasifikasi data mentah, tetapi mengalami kenaikan yang signifikan yaitu pada rentang harga murah dari 0.77 menjadi 0.96, pada rentang harga sedang dari 0.51 menjadi 0.94, pada rentang harga mahal dari 0.52 menjadi 0.94, dan pada rentang harga sangat mahal dari 0.72 menjadi 0.97.

TABLE XII. PENGHITUNGAN AKURASI MODEL *NAÏVE BAYES* MENGGUNAKAN DATA NORMALISASI

	pres	rec	f1	supp
murah (0)	0.9133	0.9072	0.9102	151
sedang (1)	0.7379	0.7328	0.7353	146
mahal (2)	0.6729	0.7229	0.6970	148
sangat mahal (3)	0.8835	0.8322	0.8571	155
acc	0.8000	0.8000	0.8000	0.8

TABLE XIII. PENGHITUNGAN AKURASI MODEL *MULTINOMIAL LOGISTIC REGRESSION* MENGGUNAKAN DATA NORMALISASI

	pres	rec	f1	supp
murah (0)	0.9795	0.9536	0.9664	151
sedang (1)	0.9102	0.9726	0.9403	146
mahal (2)	0.9854	0.9121	0.9473	148
sangat mahal (3)	0.9625	0.9935	0.9777	155
acc	0.9583	0.9583	0.9583	0.9583

V. KESIMPULAN

Hasil model klasifikasi data yang mengalami normalisasi data memiliki tingkat keakurasian yang lebih tinggi (95.8% untuk metode MLR dan 80% untuk metode *Naïve Bayes*) dibandingkan dengan hasil model klasifikasi data yang tidak mengalami normalisasi (63.5% untuk metode MLR dan 79.8% untuk metode *Naïve Bayes*). Pengaruh normalisasi data pada model klasifikasi *Naïve Bayes* dalam klasifikasi harga handphone berdasarkan spesifikasi menunjukkan hubungan yang tidak terlalu signifikan karena pertambahan akurasi dengan normalisasi data hanya sebesar 0.2%. Pengaruh normalisasi data pada model klasifikasi MLR menunjukkan bahwa hasil keakuratan metode MLR sangat bergantung dengan normalisasi data karena terjadi pertambahan akurasi yang sangat signifikan ketika dilakukan normalisasi data, yaitu sebesar 32.3%. Hasil visualisasi pada *confusion matrix* menunjukkan bahwa model klasifikasi yang memiliki tingkat keakuratan yang paling tinggi sekitar 95.8% adalah model klasifikasi *Multinomial Logistic Regression* dengan pelatihan menggunakan data yang dinormalisasi.

DAFTAR PUSTAKA

[1] S. H. Putra and B. T. Putra, "Klasifikasi Harga Cell Phone menggunakan Metode K-Nearest Neighbor (KNN)," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 242–245, 2018.

[2] O. D. Megawati and A. E. Minarno, "Klasifikasi Harga Ponsel Menggunakan Support Vector Machine (SVM)," vol. 3, no. 4, pp. 393–400, 2021.

[3] R. Y. Hayuningtyas, "Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita," *J. Inform.*, vol. 6, no. 1, pp. 18–22, 2019, doi: 10.31311/ji.v6i1.4685.

[4] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 144–147, 2018.

[5] D. B. Setyohadi, F. A. Kristiawan, and E. Ernawati, "Perbaikan Performansi Klasifikasi Dengan Preprocessing Iterative Partitioning Filter Algorithm," *Telematika*, vol. 14, no. 01, pp. 12–20, 2017, doi: 10.31315/telematika.v14i01.1960.

[6] F. Ristianto, N. Nurmalasari, and A. Yoraeni, "Impementasi Metode Naive Bayes Untuk Prediksi Harga Emas," *Comput. Sci.*, vol. 1, no. 1, pp. 62–71, 2021, doi: 10.31294/coscience.v1i1.201.

[7] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[8] R. P. Kurniadi, V. P. Widartha, and U. Telkom, "Perbandingan Akurasi Algoritma K-Nearest Neighbor Dan Logistic," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9757–9764, 2021.

[9] D. A. Novitasari and M. Yaskun, "Analisis Regresi Logistik Ordinal Pada Kepuasaan Pelanggan Mebel Lamongan," *J. Manaj.*, vol. 4, no. 1, p. 841, 2019, doi: 10.30736/jpim.v4i1.226.

[10] A. Sharma, "Mobile Price Classification," 2017. https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification (accessed Oct. 24, 2022).