

LAPORAN UJIAN TENGAH SEMESTER

ANALISIS DATA SMARTPHONE

Diajukan untuk memenuhi Ujian Tengah Semester
mata kuliah *Machine Learning*



Oleh:

Nur Zia Ratna Sari Ziani 202310009

Tika anggraeni 202310076

TI-20-KA

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS INFORMATIKA DAN PARIWISATA
INSTITUT BISNIS DAN INFORMATIKA
KESATUAN BOGOR

2023

1. Identifikasi Atribut Data

Pada bagian ini, kami akan mencari karakteristik dari dataset yang kami gunakan untuk analisis. Dataset ini disebut "Smartphone" dan terdiri dari 2000 entri data yang mewakili berbagai karakteristik dari berbagai model smartphone.

1.1 Nama dan Jenis Atribut

Berikut adalah daftar atribut beserta jenis atribut yang terdapat dalam dataset *smartphone*.

Atribut Prediktor

No	Nama Atribut	Jenis Atribut	Tipe Atribut	Deskripsi
1.	battery_power	Numerik	<i>Continuous</i>	-
2.	blue	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1
3.	clock_speed	Numerik	<i>Continuous</i>	-
4.	dual_sim	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1
5.	fc	Numerik	<i>Continuous</i>	-
6.	four_g	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1
7.	int_memory	Numerik	<i>Continuous</i>	-
8.	m_dep	Numerik	<i>Continuous</i>	-
9.	mobile_wt	Numerik	<i>Continuous</i>	-
10.	n_cores	Numerik	<i>Continuous</i>	-
11.	pc	Numerik	<i>Continuous</i>	-
12.	px_height	Numerik	<i>Continuous</i>	-
13.	px_width	Numerik	<i>Continuous</i>	-
14.	ram	Numerik	<i>Continuous</i>	-

15.	sc_h	Numerik	<i>Continuous</i>	-
16.	sc_w	Numerik	<i>Continuous</i>	-
17.	talk_time	Numerik	<i>Continuous</i>	-
18.	three_g	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1
19.	touch_screen	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1
20.	wifi	Kategorik	<i>Binary</i>	Memiliki nilai 0 dan 1

Atribut Label

No	Nama Atribut	Jenis Atribut	Tipe Atribut	Deskripsi
1.	price_range	Kategorik	-	Memiliki nilai 0,1,2,3

1.1. Atribut Kategorik

Dalam dataset ini, terdapat beberapa atribut kategorik *binary* yang merupakan bagian penting dari data analisis, yaitu:

- blue
- dual_sim
- four_g
- three_g
- touch_screen - wifi.

Atribut-atribut ini memiliki dua nilai unik, yaitu 0 dan 1, yang mengindikasikan keberadaan atau ketiadaan fitur tertentu pada *smartphone*.

2. Statistik Deskriptif Data

2.1. Data Sebelum Praproses

Sebelum kami melakukan tahap persiapan awal terhadap data, berikut adalah gambaran statistik deskriptif dari beberapa atribut dalam dataset "Smartphone." Informasi statistik ini membantu dalam memahami distribusi data sebelum kami melakukan perbaikan, seperti yang ditunjukkan di bawah ini:

Nama Atribut	battery_power	blue	clock_speed	dual_sim
--------------	---------------	------	-------------	----------

Jumlah Data	1990	2000	2000	2000
<i>Mean</i>	1237.87	0.495	1.522	0.5095
Standar Deviasi	439.68	0.5001	0.816	0.500035
Nilai Minimum	501	0	0.5	0
25% (Q1)	850.25	0	0.7	0
50% (Q2)	1225	0	1.5	1
75% (Q3)	1615	1	2.2	1
Nilai Maksimum	1998	1	3	1

2.2. Data Setelah Praproses

Berikut adalah statistik deskriptif dari data setelah menjalani proses praproses, termasuk pengisian *missing values* dan standarisasi.

Nama Atribut	battery_power	blue	clock_speed	dual_sim
Jumlah Data	2000	2000	2000	2000
<i>Mean</i>	2.84e-17	-1.24e-17	-1.54e-16	8.08e-17
Standar Deviasi	1.00025	1.00025	1.00025	1.00025
Nilai Minimum	-1.68041	-0.99005	-1.25306	-1.01918
25% (Q1)	-0.88047	-0.99005	-1.00791	-1.01918
50% (Q2)	-0.0292	-0.99005	-0.02727	0.98118
75% (Q3)	0.85798	1.01005	0.83078	0.98118
Nilai Maksimum	1.73376	1.01005	1.81141	0.98118

Setelah melewati tahap pengisian nilai yang hilang dengan strategi mean, data tersebut mengalami perubahan. Setelah praproses, data memiliki mean mendekati nol dan standar deviasi mendekati satu untuk setiap atribut, yang menunjukkan bahwa data telah diubah ke dalam skala yang seragam. Perubahan ini dilakukan untuk memastikan data siap untuk analisis.

3. Model Klasifikasi: Decision Tree

Analisis ini menggunakan algoritma Decision Tree sebagai model klasifikasi untuk analisis ini. Algoritma Decision Tree merupakan metode pembelajaran mesin yang digunakan untuk mengelompokkan data dengan mengandalkan serangkaian keputusan hierarkis yang terbentuk dalam struktur pohon. Keputusan-keputusan ini dibuat berdasarkan atribut dataset, yang membantu memprediksi label atau kategori tertentu.

3.1. Alasan Pemilihan Algoritma

Keputusan untuk menggunakan algoritma Decision Tree didasarkan pada pertimbangan interpretasi yang jelas dan kemampuan untuk mengekstraksi pengetahuan penting dari data. Hasil proses ini berupa pohon keputusan yang dapat dengan mudah diuraikan, yang memudahkan pemahaman tentang faktor-faktor yang mempengaruhi prediksi harga smartphone, yang merupakan tujuan utama dari analisis data ini.

3.2. Pelatihan Model dan Evaluasi

Model *Decision Tree* dilatih menggunakan data pelatihan sebesar 80% dari dataset, dengan pengaturan `random_state = 42` untuk memastikan hasil yang dapat direproduksi. Selanjutnya, model tersebut digunakan untuk melakukan prediksi pada data pengujian (20%) menggunakan perintah `dtree_model.predict(x_test)`.

Akurasi model dihitung untuk mengukur tingkat keberhasilan dalam memprediksi kategori harga *smartphone*. Hasil akurasi yang diperoleh adalah sebesar **81.75%**. Akurasi merupakan metrik yang mengukur sejauh mana model mampu memprediksi kategori yang benar.

3.3. Evaluasi Tambahan

Evaluasi model tidak hanya didasarkan pada akurasi, namun juga dilakukan dengan menggunakan metode berikut ini:

- *Confusion Matrix*

Confusion matrix digunakan untuk menggambarkan sejauh mana model berhasil atau gagal dalam memprediksi setiap kategori, seperti *price_range* 0, 1, 2 dan 3. Hal ini membantu dalam memahami area di mana model memiliki kesulitan dalam melakukan prediksi.

Hasil dari *confusion matrix* sebagai berikut :

[[-90Classification 15 0 0] [5 74Report 12 0] [0 16 64 12] [0 0 13 99]]

Classification Report menyediakan informasi yang lebih detail tentang *precision*, *recall*, *F1-score* dan *support* untuk setiap kategori. *Report* ini membantu dalam pemahaman performa model dengan lebih mendalam.

Berikut ini hasil dari *classification report* :

	Precision	Recall	F1-Score	Support
0	0.95	0.86	0.90	105
1	0.70	0.81	0.76	91
2	0.72	0.70	0.71	92
3	0.89	0.88	0.89	112

<i>Accur</i>		0.8	400
<i>acy</i>		2	
<i>Macro</i>	0.82	0.8	0.8 400
<i>Avg</i>		1	1
<i>Weighted</i>	0.82	0.8	0.8 400
<i>Avg</i>		2	2

4. Model Clustering: K-Means

Dalam analisis ini, menggunakan algoritma K-Means untuk mengelompokkan data dalam analisis ini. Algoritma ini adalah salah satu metode clustering yang dimaksudkan untuk menggabungkan data ke dalam berbagai kelompok berdasarkan kesamaan karakteristik tertentu. Selain itu, kami menemukan jumlah cluster yang ideal dengan metode Elbow Point dan menilai hasil clustering dengan metrik Silhouette Coefficient.

4.1. Inisialisasi Model K-Means

Pertama-tama, model K-Means diinisialisasi dengan beberapa parameter seperti jumlah *cluster* yang diinginkan, inisialisasi acak.

4.2. Detail Model K-Means

Setelah pelatihan model K-Means, beberapa detail model dieksplorasi, seperti:

- *Sum of Squared Error (SSE)*: SSE digunakan untuk mengukur sejauh mana data dalam setiap *cluster* dari pusat *cluster*. Semakin rendah nilai SSE, semakin baik model K-Means dalam membentuk *cluster* yang kompak.
- *Koordinat Pusat Cluster*: Koordinat pusat dari setiap *cluster* dalam bentuk vektor.
- *Jumlah Iterasi*: Jumlah iterasi yang diperlukan oleh algoritma K-Means hingga mencapai konvergensi.
- *Label Cluster*: Setiap data dalam dataset diberi label *cluster* yang menunjukkan *cluster* masuk.

4.3. Penentuan Jumlah *Cluster* Optimal

Untuk menentukan jumlah *cluster* yang optimal, digunakan metode *Elbow Point*. Pada metode ini, berbagai jumlah *cluster* dicoba, dan nilai SSE untuk setiap jumlah *cluster* dicatat. Titik "*elbow*" dalam grafik SSE menunjukkan jumlah *cluster* yang optimal.

Hasil menunjukkan bahwa jumlah *cluster* optimal adalah 5.

4.4. Menampilkan *Silhouette Coefficients*

Silhouette Coefficients digunakan untuk mengevaluasi kualitas pengelompokan data dalam *cluster* dengan berbagai jumlah *cluster* yang berbeda. Hasil evaluasi dengan *Silhouette Coefficients* adalah sebagai berikut :-

Silhouette Score : 0.05830501834188593

5. Kesimpulan

Setelah melakukan analisis data pada kumpulan data "Smartphone", beberapa hasil penting ditemukan:

1. Berbagai atribut diidentifikasi dalam dataset, termasuk atribut prediktor dan label. Atribut prediktor terdiri dari berbagai atribut kategorik binary dan numerik yang terkait dengan spesifikasi smartphone, sedangkan atribut label adalah atribut kategorik yang merupakan target prediksi, yaitu "price_range".
2. Sebelum praproses, analisis statistik deskriptif dilakukan pada beberapa fitur penting. Sebelum perbaikan dilakukan, statistik ini memberikan gambaran tentang sebaran data awal. Data kemudian diubah dengan standarisasi dan penambahan nilai yang tidak ada.
3. Untuk memprediksi kategori harga smartphone, Algoritma Decision Tree digunakan sebagai model klasifikasi. Algoritma ini dipilih karena keunggulannya dalam interpretabilitas. Dengan menggunakan matrik Konflik dan Laporan Klasifikasi, model ini dilatih dengan akurasi 81,75%.
4. Data digabungkan ke dalam klaster menggunakan algoritma K-Means.

Dengan menggunakan metode Elbow Point dan Silhouette Coefficients, jumlah cluster ideal adalah 5. Hasilnya menunjukkan bahwa jumlah cluster ideal adalah 5. Hasil clustering juga ditunjukkan.

Akibatnya, analisis data ini memberikan informasi penting yang membantu kita memahami dan memprediksi harga smartphone, dan membantu kita mengenali kelompok-kelompok smartphone dengan karakteristik yang sama. Hasil analisis ini dapat membantu kita membuat keputusan tentang penawaran dan perbandingan smartphone.