

LAPORAN UJIAN TENGAH SEMESTER

ANALISIS DATA SMARTPHONE

Diajukan untuk memenuhi Ujian Tengah Semester mata kuliah
Machine Learning



Oleh:

Diva Lutfiando (202310028)

Eki Virgiana (202310061)

Verrel Ardhan R.S.P (202310012)

TI-20-PA

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

2023

1. Identifikasi Atribut Data

Dalam bagian ini, kami akan mengidentifikasi atribut-atribut yang ada dalam dataset yang digunakan dalam analisis kami. Dataset yang digunakan adalah dataset "Smartphone" yang mengandung informasi tentang berbagai atribut dari berbagai model *smartphone*. Dataset ini terdiri dari 2000 data yang mewakili 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

1.2. 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 dilakukan praproses terhadap data, berikut adalah statistik deskriptif dari beberapa atribut dalam dataset "Smartphone." Statistik ini memberikan gambaran tentang bagaimana data terdistribusi sebelum diterapkan perbaikan apa pun:

Nama Atribut	battery_power	blue	clock_speed	dual_sim
Jumlah Data	1990	2000	2000	2000
Mean	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
Mean	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

Data tersebut mengalami perubahan setelah melalui tahap pengisian *missing value* dengan strategi *mean* dan proses standarisasi menggunakan *StandardScaler*. Data setelah praproses memiliki *mean* mendekati nol dan standar deviasi mendekati satu untuk setiap atribut, yang mengindikasikan bahwa data telah diubah ke dalam skala yang seragam. Perubahan ini bertujuan untuk memastikan data siap digunakan dalam analisis lebih lanjut.

3. Model Klasifikasi: Decision Tree

Dalam analisis ini, digunakan algoritma *Decision Tree* sebagai model klasifikasi. Algoritma *Decision Tree* adalah algoritma pembelajaran mesin yang digunakan untuk mengklasifikasikan data berdasarkan serangkaian keputusan berhierarki yang dibentuk dalam bentuk pohon. Keputusan-keputusan ini didasarkan pada atribut-atribut dalam dataset dan berfungsi untuk memprediksi label atau kategori tertentu.

3.1. Alasan Pemilihan Algoritma

Pemilihan algoritma *Decision Tree* didasari oleh keunggulan algoritma ini dalam hal interpretabilitas dan kemampuan untuk mengekstraksi pengetahuan yang bermakna dari data. Hasil keputusan dalam bentuk pohon dapat dijelaskan dengan mudah, sehingga mempermudah pemahaman faktor-faktor apa yang mempengaruhi prediksi harga smartphone, yang merupakan tujuan analisis.

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 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 adalah 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:

- 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 *confusion matrix*:

```
[90  15  0  0]
[5   74 12  0]
[    0  0  0  0]
[0   16 64 12]
```

[0 0 13 99]

- Classification Report

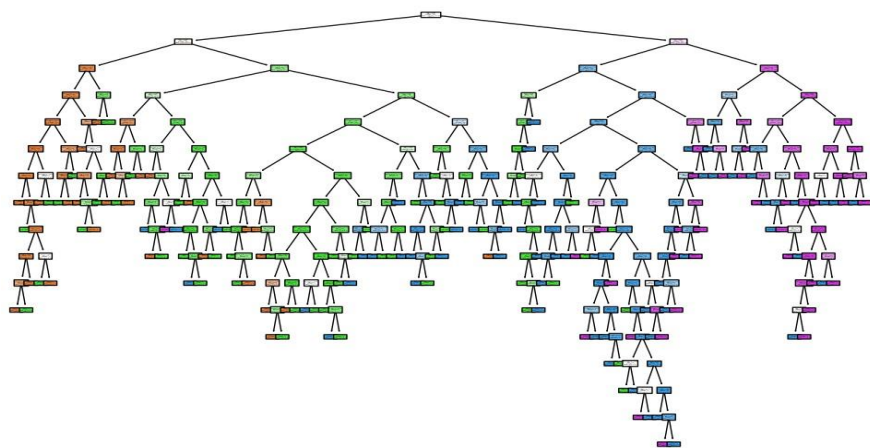
Classification Report menyediakan informasi yang lebih detail tentang *precision*, *recall*, *F1-score*, dan *support* untuk setiap kategori. *Report* ini membantu dalam pemahaman performa model dengan lebih mendalam.

Hasil *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
Accuracy			0.82	400
Macro Avg.	0.82	0.81	0.81	400
Weighted Avg.	0.82	0.82	0.82	400

- Visualisasi model Decision Tree

Berikut adalah visualisasi dari model Decision Tree dalam bentuk gambar.



4. Model Clustering: K-Means

Dalam analisis ini, algoritma K-Means digunakan untuk melakukan pengelompokan data. Algoritma K-Means adalah salah satu algoritma *clustering*

yang bertujuan untuk mengelompokkan data ke dalam beberapa kluster berdasarkan kesamaan atribut-atribut tertentu. Selain itu dilakukan juga pencarian jumlah kluster optimal dengan metode *Elbow Point* dan mengevaluasi hasil *clustering* dengan menggunakan metrik *Silhouette Coefficient*.

4.1. Inisialisasi Model K-Means

Pertama-tama, model K-Means diinisialisasi dengan beberapa parameter seperti jumlah kluster yang diinginkan, inisialisasi acak, dan pengaturan lainnya sebagai berikut.

```
# Inisialisasi model K-Means
kmeans_model = KMeans(init="random", n_clusters=4, random_state=42, n_init=10)
# Latih model dengan data yang telah diproses
kmeans_model.fit(x_scaled_df)
```

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 kluster dari pusat kluster. Semakin rendah nilai SSE, semakin baik model K-Means dalam membentuk kluster yang kompak.
- Koordinat Pusat Kluster: Koordinat pusat dari setiap kluster dalam bentuk vektor.
- Jumlah Iterasi: Jumlah iterasi yang diperlukan oleh algoritma KMeans hingga mencapai konvergensi.
- Label Kluster: Setiap data dalam dataset diberi label kluster yang menunjukkan kluster mana yang mereka masuk.

4.3. Penentuan Jumlah Kluster Optimal

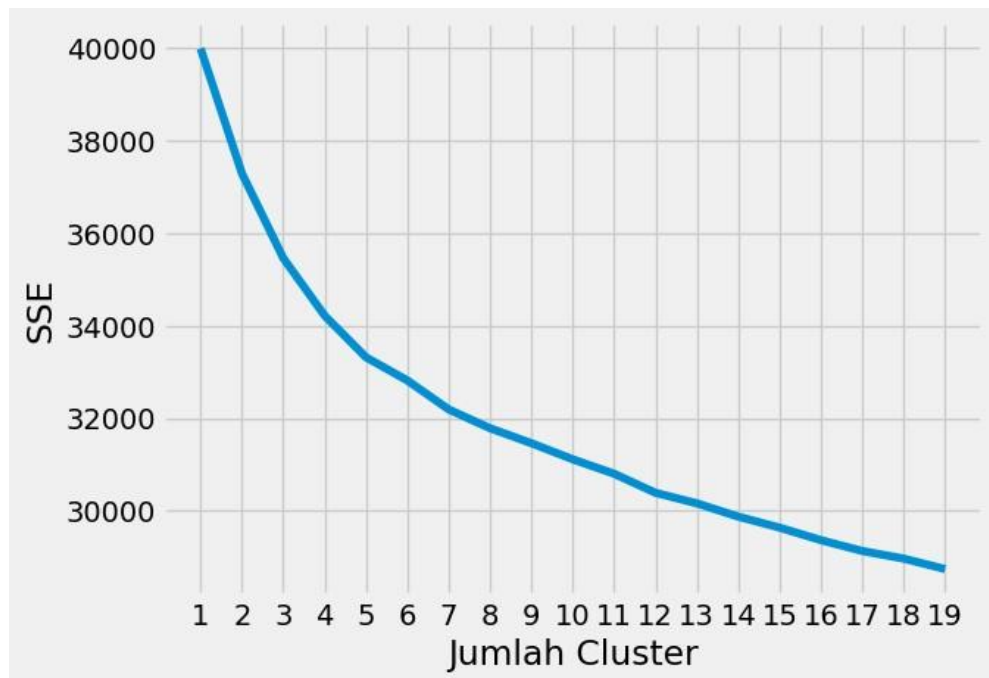
Untuk menentukan jumlah kluster yang optimal, digunakan metode *Elbow Point*. Pada metode ini, berbagai jumlah kluster dicoba, dan nilai SSE

untuk setiap jumlah kluster dicatat. Titik "elbow" dalam grafik SSE menunjukkan jumlah kluster yang optimal.

Hasil menunjukkan bahwa jumlah kluster optimal adalah 5.

4.4. Grafik Metode Elbow

Grafik metode Elbow digunakan untuk memvisualisasikan proses penentuan jumlah kluster optimal. Dengan menemukan titik siku dalam grafik, jumlah kluster yang optimal dapat ditentukan. Dan berikut grafik dari metode Elbow yang sudah dilakukan.



4.5. Menampilkan Silhouette Coefficients

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

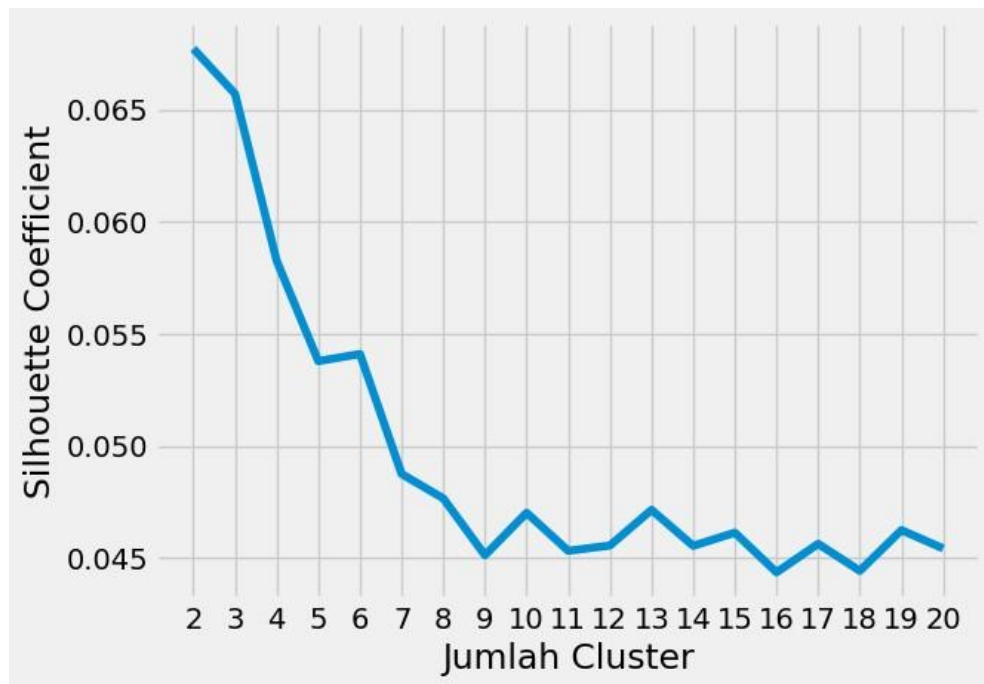
- *Silhouette Score for 2 clusters: 0.067724062611456*
- *Silhouette Score for 3 clusters: 0.06571220910132038*
- *Silhouette Score for 4 clusters: 0.05826485726578026*
- *Silhouette Score for 5 clusters: 0.05378860103239825*
- *Silhouette Score for 6 clusters: 0.05410647031902022*

- *Silhouette Score for 7 clusters: 0.048755651831570705*
- *Silhouette Score for 8 clusters: 0.04766816256107165*
- *Silhouette Score for 9 clusters: 0.04513294827395241*
- *Silhouette Score for 10 clusters: 0.04702478666445728*
- *Silhouette Score for 11 clusters: 0.045328424388324365*
- *Silhouette Score for 12 clusters: 0.0455647744862188*
- *Silhouette Score for 13 clusters: 0.04715344249564453*
- *Silhouette Score for 14 clusters: 0.04555050225716178*
- *Silhouette Score for 15 clusters: 0.0461384020607766*
- *Silhouette Score for 16 clusters: 0.044368538660926785*
- *Silhouette Score for 17 clusters: 0.0456396567650925*
- *Silhouette Score for 18 clusters: 0.04443805684585118*
- *Silhouette Score for 19 clusters: 0.046250819921978374*
- *Silhouette Score for 20 clusters: 0.04543362328682459*

Hasil evaluasi menunjukkan bahwa nilai Silhouette Score tertinggi diperoleh saat menggunakan 2 klaster.

4.6. Grafik Silhouette Coefficients

Grafik Silhouette Coefficients digunakan untuk memvisualisasikan hasil evaluasi dengan lebih jelas. Grafik ini menunjukkan tren penurunan nilai Silhouette Coefficients seiring peningkatan jumlah klaster.



5. Kesimpulan

Setelah dilakukan analisis data terhadap dataset “Smartphone” terdapat beberapa hasil temuan penting yang dapat diambil:

1. Teridentifikasi berbagai atribut dalam dataset, termasuk atribut prediktor dan atribut label. Atribut prediktor meliputi berbagai atribut numerik dan atribut kategorik *binary* yang berkaitan dengan spesifikasi *smartphone*. Atribut label adalah atribut kategorik yang merupakan target prediksi, yaitu "price_range."
2. Sebelum praproses, dilakukan analisis statistik deskriptif pada beberapa atribut utama. Statistik ini memberikan gambaran tentang sebaran data awal sebelum dilakukan perbaikan. Data kemudian mengalami transformasi melalui pengisian *missing values* dan standarisasi.
3. Algoritma *Decision Tree* digunakan sebagai model klasifikasi untuk memprediksi kategori harga *smartphone*. Algoritma ini dipilih karena keunggulan dalam interpretabilitas. Model ini dilatih dengan akurasi sebesar **81.75%** dan dievaluasi menggunakan metrik *Confusion Matrix* dan *Classification Report*.

4. Melalui algoritma K-Means, dilakukan pengelompokan data ke dalam kluster. Jumlah kluster optimal ditentukan menggunakan metode *Elbow Point* dan *Silhouette Coefficients*. Hasilnya menunjukkan bahwa jumlah kluster optimal adalah **5**. Selain itu, hasil *clustering* juga divisualisasikan.

Dengan demikian, analisis data ini memberikan wawasan yang berguna dalam pemahaman dan prediksi harga *smartphone*, serta mengidentifikasi kelompok-kelompok *smartphone* dengan karakteristik serupa. Analisis ini dapat menjadi dasar untuk pengambilan keputusan lebih lanjut terkait dengan penawaran dan perbandingan *smartphone*.

Berikut link untuk google colabs

<https://colab.research.google.com/drive/14EumMra8GShx5iD3sSCjn4sy8EHA-zAC?usp=sharing>