5. Rangkum apa saja yang dilakukan pada keempat tugas di nomor sebelumnya

a. Nama dan jenis atribut

NAMA	JENIS
battery_power	Numerik
blue	Kategorik
clock_speed	Numerik
dual_sim	Kategorik
fc	Numerik
four_g	Kategorik
int_memory	Numerik
m_dep	Numerik
mobile_wt	Numerik
n_cores	Kategorik
pc	Numerik
px_height	Numerik
px_width	Numerik
ram	Numerik
sc_h	Numerik
sc_w	Numerik
talk_time	Numerik
three_g	Kategorik
touch_screen	Kategorik
wifi	Kategorik
price_range	Kategorik

b. Data diolah untuk menangani nilai yang hilang dan untuk standarisasi data. Langkah-langkah yang dilakukan meliputi:
   - Statistik Deskriptif Awal: Sebelum praproses, informasi statistik deskriptif tentang atribut prediktor ditampilkan untuk mengetahui rentang, rata-rata, median, dan distribusi data. Hal ini juga menentukan atribut yang mungkin tidak memiliki nilai.
   - Mengisi Missing Values: Missing values diatasi dengan menggunakan SimpleImputer dengan strategi rata-rata (mean). Data yang dimasukkan kemudian diperiksa statistiknya untuk memastikan tidak ada lagi nilai yang kosong.
   - Standarisasi Data: Atribut numerik distandarisasi dengan menggunakan StandardScaler agar nilai dari setiap atribut menjadi seragam (rata-rata = 0 dan deviasi standar = 1). Statistik deskriptif setelah standarisasi juga ditampilkan untuk melihat perbedaan dalam rentang nilai yang lebih konsisten.
   - Statistik deskriptif pada setiap tahap praproses ini memberikan informasi tentang perubahan data dan persiapan untuk analisis lebih lanjut.

c. Model klasifikasi dibuat dengan menggunakan algoritma Decision Tree untuk memperkirakan rentang harga berdasarkan atribut-atribut prediktor lainnya. Langkah-langkah yang dilakukan:

- Dataset dibagi menjadi data training (85%) dan data testing (15%) menggunakan metode holdout.
- Model Decision Tree dilatih pada data training, dan kemudian dilakukan prediksi terhadap data testing.
- Model dievaluasi menggunakan confusion matrix dan akurasi. Confusion matrix adalah tabel yang menunjukkan detail prediksi benar dan salah pada setiap kelas. Sementara akurasi adalah persentase prediksi yang benar secara keseluruhan.

d. Pengelompokan model dilakukan dengan menggunakan algoritma K-Means. Karena atribut label (price_range) punya 4 kategori, model K-Means dibuat dengan 4 cluster untuk mencoba mengelompokkan data yang mirip. Langkah-langkah yang dilakukan:
   - Membangun Model: Model K-Means dengan 4 cluster dibuat dan dilatih dengan data yang sudah diimputasi dan distandarisasi.
   - Evaluasi Clustering: Kualitas clustering dinilai dengan menggunakan silhouette score. Skor siluet menunjukkan seberapa mirip objek dalam satu cluster dengan objek dalam cluster tersebut dibandingkan dengan objek dari cluster lain.
   - Skor siluet yang tinggi menunjukkan clustering yang lebih baik, sedangkan skor yang rendah menandakan bahwa data kurang terpisah dengan baik di antara cluster.