Nama : Dyas Tri Apriliansyah  
Kelas : TI -21- PA

NPM : 212310015

UTS Machine Learning

NO 1.

Mengidentifikasi nama dan jenis atribut dari seluruh data atribut predictor dan atribut label yang terdapat pada dataset tersebut. Sebelum menentukan jenis atribut , anda dapat memastikan dahulu berapa nilai unik yang terdapat dari suatu atribut, untuk atribut yang berjenis kategorik, sebutkan nilai unik yang terdapat pada atribut tersebut

Dari hasil identifikasi, berikut atribut-atribut yang bersifat kategorik beserta nilai uniknya:

1. **blue**: [0, 1]
2. **dual\_sim**: [0, 1]
3. **four\_g**: [0, 1]
4. **n\_cores**: [1, 2, 3, 4, 5, 6, 7, 8]
5. **three\_g**: [0, 1]
6. **touch\_screen**: [0, 1]
7. **wifi**: [0, 1]
8. **price\_range** (label): [0, 1, 2, 3]

Selain atribut ini, atribut lainnya berisi nilai kontinu atau diskrit non-kategorik (seperti ukuran layar dan kapasitas baterai) yang digunakan sebagai atribut predictor

Dalam dataset ini, atribut-atribut dibagi menjadi dua jenis:

1. **Atribut Prediktor**: Atribut-atribut ini digunakan untuk memprediksi label.

* **battery\_power**
* **blue**
* **clock\_speed**
* **dual\_sim**
* **fc** (front camera)
* **four\_g**
* **int\_memory**
* **m\_dep** (mobile depth)
* **mobile\_wt**
* **n\_cores**
* **pc** (primary camera)
* **px\_height**
* **px\_width**
* **ram**
* **sc\_h** (screen height)
* **sc\_w** (screen width)
* **talk\_time**
* **three\_g**
* **touch\_screen**
* **wifi**

1. **Atribut Label**: Atribut ini adalah target yang ingin diprediksi.

* **price\_range**

NO 5.

**Nama dan Jenis Atribut**

Dataset ini terdiri dari atribut-atribut prediktor dan satu atribut target atau label yang menjadi variabel respon. Secara umum, berikut penjelasan mengenai masing-masing tipe atribut:

* **Atribut Prediktor**: Merupakan fitur atau variabel independen yang digunakan untuk memprediksi atau menentukan nilai target. Nama dan jenis atribut prediktor dalam dataset perlu diperhatikan untuk memastikan bahwa tipe data sesuai dengan metode analisis yang digunakan.
* **Atribut Target (Label)**: Variabel target adalah variabel yang nilainya ingin kita prediksi. Ini merupakan variabel yang berfungsi sebagai respons atau keluaran dari model.

Pastikan bahwa jenis data pada masing-masing atribut (baik numerik maupun kategori) sesuai dengan teknik pemrosesan dan analisis yang diterapkan

**Statistik Deskriptif**

**Sebelum Pra-pemrosesan**

Sebelum melakukan pra-pemrosesan, statistik deskriptif digunakan untuk memahami distribusi data awal, termasuk nilai rata-rata, median, standar deviasi, serta rentang nilai minimum dan maksimum. Statistik deskriptif ini memberikan gambaran mengenai distribusi data, seperti adanya nilai yang hilang atau outlier. Misalnya, atribut dengan nilai rata-rata yang sangat berbeda dari median dapat menunjukkan adanya outlier atau distribusi yang miring.

**Setelah Pengisian Missing Value dan Standarisasi**

Setelah menangani missing value (mengisi dengan nilai rata-rata) dan melakukan standarisasi, distribusi data menjadi lebih seragam, terutama karena nilai setiap atribut prediktor telah dinormalisasi agar memiliki skala yang sama. Standarisasi ini mengubah nilai setiap atribut menjadi memiliki rata-rata 0 dan standar deviasi 1, yang penting untuk model seperti SVM dan K-Means yang sensitif terhadap skala data.

Statistik deskriptif setelah standarisasi menunjukkan bagaimana nilai atribut berubah dari berbagai skala menjadi skala yang seragam, membuatnya lebih sesuai untuk algoritma yang sensitif terhadap perbedaan skala antar variabel.

**Model Klasifikasi**

Model klasifikasi yang digunakan adalah **Support Vector Machine (SVM)**, yang merupakan algoritma yang bertujuan memisahkan data ke dalam kelas yang berbeda berdasarkan margin maksimum. Berikut adalah tahapan proses yang dilakukan:

* **Pembagian Data**: Data dibagi menggunakan metode holdout dengan 85% data sebagai data latih dan 15% sebagai data uji.
* **Pelatihan Model**: Model dilatih pada data latih untuk mencari garis pemisah optimal antara kelas-kelas yang ada.
* **Evaluasi Model**: Setelah model SVM dilatih, data uji digunakan untuk mengevaluasi kinerjanya dengan menghitung confusion matrix dan akurasi. Confusion matrix memberikan informasi mengenai jumlah prediksi benar dan salah untuk setiap kelas, sementara akurasi menunjukkan persentase prediksi yang benar.

SVM efektif dalam klasifikasi data yang terpisah dengan margin yang jelas, dan evaluasi dengan confusion matrix serta akurasi memberikan informasi mengenai kinerja prediksi model ini.

**Model Clustering**

Model clustering yang digunakan adalah **K-Means**, sebuah algoritma clustering yang mengelompokkan data ke dalam k klaster berdasarkan jarak antar titik data. Berikut adalah langkah-langkah yang dilakukan:

* **Penentuan Jumlah Klaster Optimal**: Elbow Method digunakan untuk menentukan jumlah klaster optimal dengan melihat nilai inertia yang mengecil secara signifikan pada titik tertentu.
* **Pelatihan Model K-Means**: Setelah jumlah klaster optimal dipilih, model K-Means dilatih pada data yang telah distandarisasi.
* **Evaluasi dengan Silhouette Score**: Silhouette Score digunakan untuk mengevaluasi kualitas clustering, dengan skor yang lebih tinggi menunjukkan klaster yang lebih baik. Silhouette Score mengukur konsistensi antar titik data dalam satu klaster dibandingkan dengan titik data di klaster lain.

K-Means sangat berguna untuk mengelompokkan data berdasarkan kesamaan, dan Silhouette Score memberikan indikasi mengenai seberapa baik model membentuk klaster yang terpisah satu sama lain.