LAPORAN UTS MACHINE LEARNING

Disusun oleh:

Katarina Andrea Laurentia – 212310008 – TI21PA

Michael Fernandez – 212310060 – TI21PA

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS INFORMATIKA DAN PARIWISATA

INSTITUT BISNIS DAN INFORMATIKA

KESATUAN BOGOR

2024

A. Pembahasan mengenai nama dan jenis atribut prediktor dan atribut label yang terdapat pada dataset

Atribut Prediktor dikenal juga sebagai fitur atau variabel independen. Atribut prediktor adalah variabel yang digunakan untuk memprediksi atau mempengaruhi hasil pada atribut target atau label. Contoh atribut prediktor misalnya adalah umur, jenis kelamin, pendidikan, atau pendapatan dalam dataset terkait analisis sosial.

Atribut label disebut juga variabel dependen atau target. Ini adalah variabel yang menjadi hasil atau keluaran yang ingin diprediksi atau diklasifikasikan. Contoh atribut label misalnya adalah status penyakit (sehat/sakit) dalam dataset medis, atau kategori kelulusan (lulus/tidak lulus) dalam dataset akademik.

Atribut-atribut pada **mobileprice_modified.csv** diklasifikasikan menjadi beberapa jenis berdasarkan skala pengukurannya. Atribut **kategorik** dalam dataset ini adalah **n_cores**, yang menunjukkan jumlah inti prosesor dengan nilai unik **[1, 2, 3, 4, 5, 6, 7, 8]**. Untuk atribut **biner**, terdapat atribut simetris seperti **dual_sim, touch_screen, dan wifi**, yang memiliki nilai biner (0 dan 1) dan tidak memerlukan asumsi tertentu mengenai nilai tersebut. Sementara itu, atribut **biner asimetris** meliputi **blue, four_g, dan three_g**, yang juga memiliki nilai biner (0 dan 1) tetapi dengan asumsi bahwa salah satu nilai mewakili keberadaan fitur dan yang lainnya ketiadaan. Atribut **numerik dengan skala rasio** mencakup **battery_power, clock_speed, fc , int_memory, m_dep, mobile_wt, pc, px_height, px_width, ram, sc_h, sc_w, dan talk_time**. Skala rasio ini menunjukkan bahwa data memiliki nilai nol mutlak dan perbandingan nilai-nilai numerik memiliki makna. Atribut label **price_range** akan menjadi target atau prediksi utama dari dataset ini.

B. Pembahasan mengenai statistik deskriptif dari data, baik untuk sebelum dilakukan praproses maupun data setelah dilakukan pengisian missing values dan standarisasi

Statistik deskriptif adalah koefisien informasi singkat yang meringkas

kumpulan data tertentu, yang dapat berupa representasi seluruh populasi atau sampel populasi. Statistik deskriptif dipecah menjadi ukuran kecenderungan sentral dan ukuran variabilitas (penyebaran). Ukuran kecenderungan sentral meliputi mean, median, dan modus, sedangkan ukuran variabilitas meliputi deviasi standar, varians, variabel minimum dan maksimum, kurtosis , dan kemiringan. Beberapa contoh kaitan dengan

1) Praproses

   praproses data adalah langkah awal dalam pemrosesan data yang melibatkan pembersihan dan transformasi data mentah menjadi data yang siap untuk analisis. Statistik deskriptif sangat berguna dalam tahap ini untuk mengidentifikasi anomali, distribusi, serta *outlier* (data yang sangat jauh dari nilai rata-rata), yang bisa mempengaruhi hasil analisis jika dibiarkan.

2) Missing Value (Nilai yang Hilang)

   Nilai yang hilang atau missing values adalah data yang tidak tersedia dalam *dataset*. Mengidentifikasi *missing values* adalah salah satu langkah praproses yang penting, karena data yang hilang dapat memengaruhi hasil analisis. Statistik deskriptif seperti jumlah, mean, dan median dapat membantu mendeteksi dan menangani nilai yang hilang. Metode umum untuk menangani *missing values* termasuk mengisi nilai hilang dengan mean atau median, menghapus baris yang memiliki nilai hilang, atau menggunakan model prediktif.

3) Standarisasi

   Standarisasi adalah proses mengubah skala data sehingga setiap fitur memiliki skala yang sama, biasanya dengan mengonversi data menjadi nilai yang berdistribusi normal (nilai mean 0 dan standar deviasi 1). Standarisasi penting ketika data memiliki skala yang berbeda, karena model analisis seperti regresi atau *machine learning* akan lebih optimal jika fitur-fitur memiliki skala yang seimbang. Statistik deskriptif membantu dalam standarisasi dengan memberikan informasi mengenai mean dan standar deviasi, yang digunakan dalam formula standarisasi.

Pra-pemrosesan data dimulai dengan memisahkan atribut prediktor dan label **(price_range)** untuk fokus analisis, kemudian menampilkan statistik

deskriptif awal seperti rata-rata, standar deviasi, minimum, dan maksimum sebagai gambaran distribusi awal data. Langkah selanjutnya adalah imputasi nilai hilang dengan mean, sehingga kolom yang sebelumnya kurang dari **2000** data menjadi lengkap. Setelah imputasi, statistik utama seperti rata-rata dan standar deviasi tetap stabil, tetapi variabilitas data berkurang.

Langkah terakhir adalah standarisasi menggunakan **StandardScaler**, yang menyesuaikan semua atribut agar memiliki rata-rata 0 dan standar deviasi 1. Proses ini menyamakan skala data untuk memaksimalkan akurasi pada algoritma berbasis jarak, seperti K-Means

Output statistik deskriptif menunjukkan perbedaan signifikan pada tiga tahap pemrosesan. Sebelum imputasi, beberapa kolom memiliki kurang dari 2000 data, yang mengindikasikan adanya nilai hilang. Statistik deskriptif menunjukkan variasi besar di beberapa fitur, seperti battery_power dan ram. Setelah imputasi, jumlah data di setiap kolom menjadi seragam dengan 2000 entri, dan rata-rata serta standar deviasi berubah sedikit pada kolom seperti int_memory, menunjukkan bahwa proses imputasi telah berhasil tanpa memengaruhi distribusi data secara signifikan. Setelah standarisasi, seluruh kolom memiliki rata-rata mendekati nol dan standar deviasi mendekati satu, menandakan bahwa data telah distandarisasi sepenuhnya. Ini penting untuk memastikan skala yang seragam, sehingga semua fitur memberikan kontribusi yang seimbang pada model pembelajaran mesin yang akan digunakan.

C. Pembahasan mengenai model klasifikasi

Klasifikasi adalah *supervised machine learning* yang dimana model mencoba memprediksi label yang benar dari data masukan yang diberikan. Dalam klasifikasi, model dilatih sepenuhnya menggunakan data pelatihan, lalu dievaluasi pada data uji sebelum digunakan untuk melakukan prediksi pada data baru yang belum terlihat. Ada beberapa algoritma untuk model klasifikasi pada machine learning, contohnya:

1. Decision Tree

   Decision Tree adalah model klasifikasi di mana proses pembelajaran adalah metode untuk mendekati fungsi target diskrit yang direpresentasikan oleh

decision tree. Kata tree merujuk pada mathematical graph theory, yang didefinisikan sebagai grafik tidak berarah di mana dua simpul (*node*) terhubung oleh satu jalur (*path*). Decision tree akan membuat simpul yang terus membagi berdasarkan pembelajaran data dan akan berhenti sampai parameter yang telah ditentukan. Model ini cukup populer digunakan oleh banyak ahli karena cepat dan mudah dijelaskan. namun model ini sering dapat mengalami masalah *overfitting*.

2. SVM (Support Vector Machine)

   SVM adalah pengklasifikasi yang membuat batasan untuk memisahkan kelas-kelas yang berbeda. Data disebut support vektor untuk membantu membuat batasan. Batasan itu disebut hyperplane atau pembagi. Ini dihitung berdasarkan dataset dan dengan mengukur margin terbaik dengan memindahkan hyperplane. Ketika data berada dalam dimensi yang lebih tinggi atau ketika ada data yang tidak dapat dipisahkan secara linear, kita akan menggunakan Kernel trick untuk menemukan hyperplane

Pada kode program nomor 3, algoritma **Decision Tree** digunakan sebagai model klasifikasi untuk memprediksi kategori **price_range** dalam dataset. **Decision Tree** adalah algoritma pembelajaran mesin yang membentuk pohon keputusan berdasarkan fitur-fitur dari data, memisahkan dataset secara berulang hingga mendapatkan aturan yang paling tepat untuk mengklasifikasikan data. Dalam kasus ini, dataset dibagi menjadi data ***training*** **(85%)** dan **data *testing* (15%)** dengan metode ***holdout***, di mana data training digunakan untuk melatih model dan data testing digunakan untuk menguji kinerja model yang telah dilatih.

Setelah model dilatih, dilakukan prediksi pada data testing, dan hasil prediksi dievaluasi menggunakan Confusion Matrix dan akurasi. Confusion Matrix adalah tabel yang menampilkan perbandingan antara label asli (*True Labels*) dan label yang diprediksi (*Predicted Labels*) pada masing-masing kelas. Matriks ini memberikan gambaran mengenai jumlah prediksi yang benar (diagonal utama) serta kesalahan klasifikasi yang terjadi untuk setiap kelas.

Dalam contoh ini, terdapat empat kategori dari price_range, yang masing-masing ditampilkan dalam *Confusion Matrix*.

Akurasi model sebesar **0.84** atau 84%, menunjukkan bahwa model mampu mengklasifikasikan data dengan benar sebanyak 84% pada data testing. Angka ini menunjukkan kinerja yang cukup baik, tetapi perlu ditinjau lebih lanjut apakah angka ini sudah memenuhi kriteria yang diinginkan atau masih dapat ditingkatkan, misalnya dengan pengoptimalan parameter atau menggunakan algoritma klasifikasi lainnya.

D. Pembahasan mengenai model clustering

Clustering adalah salah satu teknik dari algoritma machine learning yaitu unsupervised learning. Algoritma clustering membagi populasi atau data point dengan sifat yang sama ke beberapa kelompok kecil untuk dikelompokkan. Contoh salah satu dari algorima clustering ada metode K-Means.

K-Means Clustering adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan *unssupervised* learning dan menggunakan metode yang mengelompokan data berbagai partisi. K Means Clustering memiliki objective yaitu meminimalisasi object function yang telah di atur pada proses clasterisasi. Dengan cara minimalisasi variasi antar 1 cluster dengan maksimalisasi variasi dengan data di *cluster* lainnya.

Langkah pertama dalam menggunakan algoritma K-Means untuk clustering adalah menentukan jumlah *cluster*, yang dalam hal ini ditetapkan menjadi 3. K-Means mengelompokkan data berdasarkan jarak Euclidean untuk meminimalkan variasi dalam setiap cluster. Setelah model K-Means dibentuk, kualitas *clustering* dievaluasi menggunakan *Silhouette Score*, yang berkisar antara -1 hingga 1. Skor mendekati 1 menunjukkan bahwa data dikelompokkan dengan baik, sedangkan skor mendekati 0 menunjukkan data berada di antara dua cluster tanpa pemisahan jelas.

Silhouette Score sebesar 0.06 mengindikasikan bahwa data kurang terkelompok dengan baik, kemungkinan karena jumlah cluster yang mungkin tidak optimal atau struktur alami data yang tidak mudah dipisahkan dalam tiga

kelompok berbeda. Dalam konteks ini, eksperimen lebih lanjut dengan jumlah *cluster* atau metode *clustering* lain mungkin diperlukan untuk menemukan representasi terbaik dari data

Daftar pustaka

Source code

https://github.com/Katkatarr/uas_machine_learning/blob/main/UAS.ipynb

7 Algoritma Klasifikasi untuk Machine Learning

Cornellius Yudha Wijaya, 2 Februari **2024**

https://www.berdata.com/post/7-algoritma-klasifikasi-untuk-machine-learning

Classification in Machine Learning: An Introduction

Zoumana Keita, 8 Agustus 2024

https://www.datacamp.com/blog/classification-machine-learning

Clustering Algoritma (K-Means)

Bryan Orleans dan Edi Purnomo Putra, 31 Januari 2022

https://sis.binus.ac.id/2022/01/31/clustering-algoritma-k-means/

Understanding Clustering: Definition, Types, and How It Works

Coding Studio Team, 4 November 2021

https://codingstudio.id/blog/mengenal-clustering/

PPT Machine Learning