Tugas 4: Praktikum Mandiri 4 Machine Learning

**Hisyam Wildan Alfath - 0110222206**

Teknik Informatika, STT Terpadu Nurul Fikri, Depok

E-mail: [hisy22206ti@student.nurulfikri.ac.id](mailto:hisy22206ti@student.nurulfikri.ac.id)

**Abstract.** Praktikum ini membangun model prediksi keputusan pembelian mobil menggunakan algoritma Logistic Regression berdasarkan data numerik dan kategorik calon pembeli. Seluruh proses meliputi pra-pemrosesan data, eksplorasi, pemilihan fitur, pelatihan dan evaluasi model, serta interpretasi hasil prediksi. Model yang dihasilkan menunjukkan akurasi dan stabilitas tinggi dalam klasifikasi keputusan membeli mobil, dengan fitur seperti status, jumlah mobil, penghasilan, dan jenis kelamin berperan signifikan dalam hasil prediksi. Temuan ini memberikan wawasan penting mengenai faktor-faktor utama yang mempengaruhi perilaku pembelian, sehingga dapat membantu industri otomotif menyusun strategi pemasaran yang lebih efektif dan berbasis data.

1. Praktikum Mandiri 03

* Menghubungkan Colab dengan Google Drive

A screen shot of a computer

AI-generated content may be incorrect.

Kode ini memanggil modul drive dari google.colab untuk menghubungkan (mount) Google Drive ke lingkungan Google Colab, dan outputnya adalah muncul pesan “Mounted at /content/drive” yang menandakan folder Google Drive sekarang sudah bisa diakses pada path tersebut di Colab.

* Import Library Machine Learning

A screenshot of a computer program

AI-generated content may be incorrect.

Kode ini berfungsi untuk mengimpor library seperti numpy, pandas, matplotlib, serta modul-modul dari scikit-learn yang dibutuhkan dalam membangun dan mengevaluasi model machine learning di Python. Langkah ini harus dilakukan agar semua fungsi analisis data, pemrosesan, dan evaluasi model tersedia dan siap digunakan.

* Membaca file CSV menggunakan pandas

A screenshot of a computer

AI-generated content may be incorrect.

Kode ini mengimpor library pandas sebagai pd, lalu menggunakan fungsi pd.read\_csv untuk membaca file CSV dari Google Drive sesuai path yang diberikan. Outputnya berupa DataFrame yang menampilkan data awal dari file CSV tersebut, seperti kolom ID, Usia, Status, Kelamin, Memiliki\_Mobil, Penghasilan, dan Beli\_Mobil, sehingga pengguna bisa langsung melihat beberapa baris pertama data yang sudah berhasil dimuat.

* Menampilkan Informasi DataFrame dengan Pandas

A screenshot of a computer

AI-generated content may be incorrect.

Kode df.info() digunakan untuk menampilkan detail struktur DataFrame yang telah dibaca dari file CSV menggunakan pandas. Outputnya memperlihatkan jumlah baris data, nama-nama kolom, tipe data setiap kolom (int64), serta memori yang digunakan sehingga memudahkan analisis awal dan pengecekan data sebelum proses selanjutnya.

* Cek Data Kosong (Missing Value) pada DataFrame

A screenshot of a computer

AI-generated content may be incorrect.

Kode df.isna().sum() digunakan untuk menghitung jumlah data yang kosong (missing value) di setiap kolom pada DataFrame. Berdasarkan output pada gambar, seluruh kolom memiliki nilai 0, yang berarti tidak ada data yang hilang dan setiap kolom sudah lengkap di dataset ini.

* Membuat Matriks Korelasi dari Data Numerik

**A screenshot of a computer

AI-generated content may be incorrect.**

Kode df.corr(numeric\_only=True) digunakan untuk menghitung matriks korelasi antar kolom numerik pada DataFrame menggunakan pandas. Hasilnya adalah tabel yang menunjukkan hubungan linear antara variabel-variabel, dengan nilai mendekati 1 atau -1 menandakan korelasi kuat positif atau negatif, sedangkan nilai mendekati 0 menunjukkan korelasi lemah atau tidak ada korelasi. Matriks korelasi ini membantu untuk memahami hubungan antar fitur dalam dataset yang akan dianalisis lebih lanjut.

* Visualisasi Korelasi dengan Heatmap

A screenshot of a computer

AI-generated content may be incorrect.

Kode pada gambar menggunakan seaborn dan matplotlib untuk membuat heatmap dari matriks korelasi yang telah dihitung sebelumnya. Hasil visualisasi heatmap memperlihatkan seberapa kuat hubungan antar variabel numerik dalam bentuk warna, di mana semakin merah menunjukkan korelasi positif semakin kuat, dan semakin biru menunjukkan semakin lemah atau negatif. Heatmap ini sangat berguna untuk analisis eksplorasi karena memudahkan dalam mengidentifikasi pola dan hubungan antar fitur pada dataset.

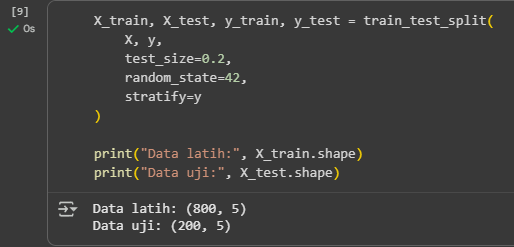
* Menyiapkan Fitur dan Target untuk Model

A screen shot of a computer program

AI-generated content may be incorrect.

Kode pada gambar bertujuan memilih fitur numerik dan biner yang akan digunakan sebagai input (X) dan menentukan target (y) dari data 'Beli\_Mobil'. Kolom-kolom yang relevan digabungkan lalu missing value dihapus dengan .dropna(), hasilnya X memiliki bentuk (1000, 5) dan y berbentuk (1000,). Langkah ini memastikan data siap dipakai untuk pelatihan model machine learning.

* Membagi Data Latih dan Uji



Kode menggunakan train\_test\_split untuk membagi dataset menjadi data latih (training) dan data uji (testing) dengan proporsi 80%:20%. Hasilnya, ukuran data latih adalah (800, 5) dan data uji (200, 5), sehingga model dapat dilatih menggunakan data latih dan dievaluasi performanya pada data uji secara terpisah.

* Melatih Model Logistic Regression dengan Pipeline

A screenshot of a computer program

AI-generated content may be incorrect.

Kode pada gambar menggabungkan transformasi fitur numerik (scaling dengan StandardScaler) dan biner ke dalam sebuah pipeline menggunakan ColumnTransformer serta Pipeline dari sklearn. Model Logistic Regression kemudian dilatih (fit) dengan data latih (X\_train, y\_train), dan proses pelatihan berjalan sukses dengan output "Model Logistic Regression berhasil dilatih". Dengan pipeline, proses pengolahan data dan pelatihan model menjadi lebih rapi dan otomatis.

* Evaluasi Model Logistic Regression

A screenshot of a computer program

AI-generated content may be incorrect.

Kode pada gambar melakukan prediksi data uji memakai model Logistic Regression yang sudah dilatih, lalu menghitung berbagai metrik evaluasi seperti akurasi, precision, recall, f1-score, dan ROC-AUC. Hasil evaluasi menunjukkan performa model sangat baik, dengan nilai akurasi 0.9300, precision 0.9748, recall 0.9143, f1-score 0.9431, dan ROC-AUC 0.9770 pada data uji. Metrik ini berguna untuk mengetahui kekuatan dan kelemahan model dalam klasifikasi.

* Visualisasi Confusion Matrix

A screenshot of a computer screen

AI-generated content may be incorrect.

Kode pada gambar menggunakan ConfusionMatrixDisplay dari sklearn untuk memvisualisasikan confusion matrix hasil prediksi model. Hasil plot confusion matrix menunjukkan performa model dalam mengklasifikasikan data, dimana dari 200 data uji: 70 data "Membeli" dan 116 data "Tidak Membeli" berhasil diprediksi dengan benar, sementara terdapat 3 prediksi yang salah pada kelas "Membeli" dan 11 yang salah pada kelas "Tidak Membeli". Visualisasi ini membantu melihat detail kesalahan dan ketepatan model dalam klasifikasi dua kelas.

* Visualisasi ROC Curve Logistic Regression

A screenshot of a computer screen

AI-generated content may be incorrect.

Kode pada gambar menggunakan fungsi RocCurveDisplay.from\_estimator untuk memvisualisasikan ROC Curve hasil prediksi model Logistic Regression pada data uji. Kurva ROC memperlihatkan hubungan antara true positive rate dan false positive rate pada berbagai threshold, dengan nilai AUC yang mendekati 1 (AUC = 0.98) menandakan performa model sangat baik dalam membedakan dua kelas target. Visualisasi ini penting untuk mengevaluasi kemampuan model dalam klasifikasi secara menyeluruh.

* Classification Report Hasil Prediksi

A screenshot of a computer

AI-generated content may be incorrect.

Kode pada gambar menggunakan fungsi classification\_report dari sklearn untuk menampilkan laporan evaluasi model berdasarkan data uji dan hasil prediksi. Laporan ini memuat nilai precision, recall, f1-score, dan jumlah data (support) untuk masing-masing kelas “Tidak Membeli” dan “Membeli”, serta rata-rata akurasi keseluruhan yang mencapai 0.93. Classification report sangat membantu untuk memahami performa model pada masing-masing kelas secara detail.

* Evaluasi Model dengan Cross Validation

A screen shot of a computer

AI-generated content may be incorrect.

Kode pada gambar menggunakan fungsi cross\_val\_score untuk melakukan cross validation (validasi silang) sebanyak 5 fold pada model. Hasilnya didapatkan skor akurasi untuk masing-masing fold serta rata-rata akurasi sebesar 0.907 dan standar deviasi 0.06, menandakan model cukup stabil dan kinerjanya konsisten pada berbagai subset data. Cross validation sangat penting untuk memastikan model tidak hanya bagus pada satu bagian data tertentu, melainkan juga pada data lain secara umum.

* Interpretasi Koefisien dan Odds Ratio Logistic Regression

A computer screen shot of a program

AI-generated content may be incorrect.

Kode pada gambar mengambil nama fitur, koefisien (log-odds), serta odds ratio dari model Logistic Regression yang telah dilatih, lalu menampilkannya dalam DataFrame yang diurutkan berdasarkan nilai odds ratio tertinggi. Tabel ini memudahkan interpretasi pengaruh masing-masing fitur terhadap kemungkinan target “Beli\_Mobil”, di mana nilai koefisien positif meningkatkan odds membeli mobil, dan nilai negatif menurunkannya. Semakin besar (positif) odds ratio, semakin besar peluang seseorang diklasifikasikan akan membeli mobil jika nilai fitur tersebut meningkat.

* Contoh Prediksi Data Baru

A screen shot of a computer program

AI-generated content may be incorrect.

Kode pada gambar menunjukkan bagaimana melakukan prediksi pada data baru menggunakan model yang sudah dilatih. Dua sampel data dengan fitur seperti Usia, Status, Kelamin, jumlah mobil, dan Penghasilan diprediksi probabilitas serta label apakah akan membeli mobil (1) atau tidak (0)—hasilnya terlihat bahwa sampel pertama diprediksi membeli mobil, sementara sampel kedua tidak. Probabilitas dan prediksi diurutkan secara otomatis untuk memudahkan interpretasi hasil pada kasus baru.

* Kesimpulan

Model Logistic Regression yang diterapkan pada data calon pembeli mobil mampu menghasilkan prediksi dengan performa sangat baik, ditunjukkan oleh akurasi, precision, recall, f1-score, dan ROC-AUC yang tinggi pada data uji. Seluruh fitur yang digunakan, seperti status, jumlah mobil, penghasilan, dan jenis kelamin, memiliki pengaruh yang signifikan dalam menentukan probabilitas keputusan membeli mobil. Validasi silang menunjukkan model cukup stabil dan hasil prediksi pada data baru konsisten sesuai karakteristik input fiturnya.

Link Praktikum 04:

Tugas Mandiri 04: