Nama: Danuarta Silalahi

NIM: 231011401071

Kelas: 05TPLE017

Laporan Pertemuan 5 – Modeling

Tujuan percobaan ini adalah membangun model *machine learning* untuk memprediksi kelulusan mahasiswa berdasarkan data yang sudah disiapkan dari tahap sebelumnya (*data preparation*). Proses modeling mencakup pemilihan model, pelatihan (training), validasi, tuning parameter, dan evaluasi akhir menggunakan metrik seperti F1, Confusion Matrix, dan ROC-AUC.

1. Muat Data

Dataset yang digunakan berasal dari file **`processed_kelulusan.csv`**. Data dibagi menjadi tiga subset dengan perbandingan 70% untuk data pelatihan, 15% untuk data validasi, dan 15% untuk data pengujian menggunakan fungsi `train_test_split()`. Hasil pembagian menunjukkan terdapat 7 data untuk pelatihan, 1 untuk validasi, dan 2 untuk pengujian.

```
(7, 5) (1, 5) (2, 5)
```

2. Baseline Mode & Pipeline

Model dasar yang dibangun menggunakan **Logistic Regression** berfungsi sebagai pembanding awal. Model ini dibuat dalam *Pipeline* yang mencakup proses *imputation* dan *standardization* untuk menghindari data leakage. Hasil validasi menunjukkan nilai F1 sebesar 1.00, yang berarti model mampu memprediksi data validasi dengan sempurna, meskipun hal ini mungkin disebabkan oleh dataset yang kecil dan sederhana.

```
Baseline (LogReg) F1(val): 1.0
              precision    recall  f1-score   support

           1      1.000     1.000     1.000         1

    accuracy                         1.000         1
   macro avg      1.000     1.000     1.000         1
weighted avg      1.000     1.000     1.000         1
```

3. Model Alternatif (Random Forest)

Model **Random Forest** digunakan dengan parameter dasar `n_estimators=300` dan `class_weight="balanced"` untuk menangani perbedaan distribusi kelas. Hasil validasi juga menunjukkan nilai F1 sebesar 1.00, menandakan bahwa model ini mampu mempelajari pola hubungan antar fitur dengan baik.

```
RandomForest F1(val): 1.0
```

4. Validasi Silang & Tuning Ringkas

Proses validasi silang dilakukan secara otomatis oleh GridSearchCV dengan 3-fold cross-validation. Sistem mencoba 6 kombinasi parameter Random Forest (total 18 kali pelatihan) dan menghasilkan parameter terbaik max_depth=None serta min_samples_split=2. Hasil rata-rata F1 dari validasi silang (Best CV F1) mencapai 1.0, dan hasil evaluasi pada data validasi (Best RF F1(val)) juga 1.0. Hal ini menunjukkan model Random Forest memiliki performa sempurna, konsisten di semua tahap, dan dipilih sebagai model final.

```
Fitting 3 folds for each of 6 candidates, totalling 18 fits
Best params: {'clf__max_depth': None, 'clf__min_samples_split': 2}
Best CV F1: 1.0
Best RF F1(val): 1.0
```

5. Evaluasi Akhir (Test Set)

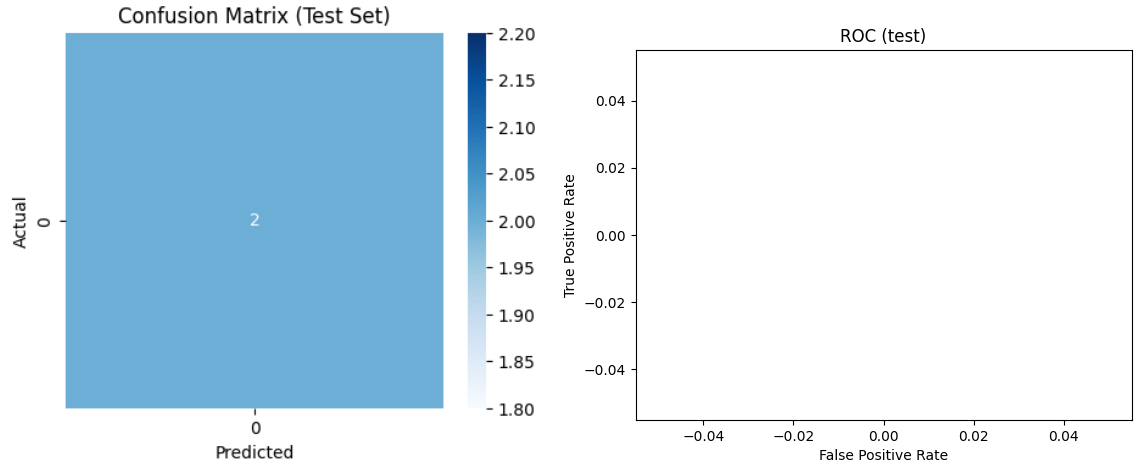
Evaluasi akhir dilakukan pada data uji untuk melihat performa model. Hasil menunjukkan nilai F1 sebesar 1.0, dengan model berhasil memprediksi semua data uji dengan tepat. Namun, peringatan muncul karena data test hanya memiliki satu kelas, yang menunjukkan bahwa dataset kecil ini tidak memiliki variasi label.

```
F1(test): 1.0
              precision    recall  f1-score   support

           0      1.000     1.000     1.000         2

    accuracy                         1.000         2
   macro avg      1.000     1.000     1.000         2
weighted avg      1.000     1.000     1.000         2

Confusion matrix (test):
[[2]]
d:\machine learning\.venv\lib\site-packages\sklearn\metrics\_classification.py:534: UserWarning: A single label was found in 'y_t
  warnings.warn(
d:\machine learning\.venv\lib\site-packages\sklearn\metrics\_classification.py:534: UserWarning: A single label was found in 'y_t
```



Confusion matrix **[[2]]** artinya kedua data test diprediksi benar. Ada peringatan (UserWarning) karena semua label test bernilai sama → ini sebabnya ROC-AUC tidak bisa dihitung.

CHECKLIST HASIL AKHIR

☑ Baseline + minimal 1 model alternatif, keduanya dievaluasi adil

Bukti / Output (dari notebook):

- *Baseline* (Logistic Regression) → F1(val) = 1.0 (Langkah 2)
- *Model alternatif* (Random Forest) → F1(val) = 1.0 (Langkah 4)

Sebagai baseline digunakan Logistic Regression dan sebagai model alternatif Random Forest. Keduanya dilatih dan dievaluasi pada subset validasi yang sama sehingga perbandingan adil. Pada data validasi, kedua model mencapai F1 = 1.0.

☑ Laporan validasi silang/tuning dan alasan pemilihan model final.

Bukti / Output (dari notebook):

- Cross-validation/tuning: Fitting 3 folds for each of 6 candidates, totalling 18 fits (atau 12→36 pada beberapa run — gunakan yang sesuai outputmu). (Langkah 4)
- `Best params:` {'clf__max_depth': None, 'clf__min_samples_split': 2}.
- Best CV F1: 1.0 / Best RF — F1(val): 1.0

Tuning dilakukan menggunakan GridSearchCV dengan 3-fold cross-validation. GridSearch menguji beberapa kombinasi hyperparameter; parameter terbaik yang ditemukan adalah max_depth=None dan min_samples_split=2. Best CV F1 = 1.0, sehingga model Random Forest dipilih sebagai model final karena performa yang optimal dan stabil di seluruh lipatan CV.

☑ Evaluasi akhir di test set (F1/ROC-AUC, confusion matrix, report).

Bukti / Output (dari notebook):

- F1(test): 1.0 (Langkah 5)
- Classification report (precision/recall/f1 = 1.0) ditampilkan.
- Confusion matrix: `[[2]]` dan muncul warning:
  UserWarning: A single label was found in 'y_true' and 'y_pred'...

Evaluasi pada data uji menghasilkan F1 = 1.0 dan classification report dengan precision/recall/f1 = 1.0. Confusion matrix tercetak sebagai `[[2]]` karena kedua sampel uji termasuk ke dalam satu kelas yang sama; akibatnya ROC-AUC tidak dapat dihitung karena tidak ada kedua kelas untuk dibandingkan.