Nama : Fawzy Imam Hamdani Gunamukti

NIM : 231011401030

KLS : 05TPLE017

Penjelasan Machine Learning Pertemuan 6

1. Membaca data hasil pra-pemrosesan dari file CSV.

2. Memisahkan data menjadi fitur (X) dan target (y).

3. Membagi dataset menjadi train, validation, dan test (70%-15%-15%)

dengan stratifikasi agar seimbang.

4. Menentukan kolom numerik dan membuat pipeline pra-pemrosesan:

- SimpleImputer(strategy=‘median’) untuk mengisi nilai kosong.

- StandardScaler untuk menormalkan data numerik.

5. Membuat model awal menggunakan RandomForestClassifier:

- Parameter: 300 pohon (n\_estimators=300), max\_features=‘sqrt’,

class\_weight=‘balanced’.

6. Melatih model baseline dan mengevaluasi performa pada data validasi

menggunakan F1-score dan classification report.

7. Melakukan validasi silang (cross-validation) dengan StratifiedKFold

(5 lipatan) untuk mengukur stabilitas model.

8. Melakukan tuning hyperparameter dengan GridSearchCV:

- Parameter yang diuji: max\_depth dan min\_samples\_split.

- Evaluasi menggunakan F1-macro dengan cross-validation.

- Menampilkan parameter terbaik dan hasil evaluasi validasi.

9. Mengevaluasi model terbaik pada data test:

- Menghitung F1-score, classification report, dan confusion

matrix.

- Menghitung ROC-AUC jika model mendukung probabilitas

(predict\_proba).

- Menghasilkan grafik ROC Curve dan Precision-Recall Curve dalam

format PNG.

10. Menampilkan 10 fitur terpenting (feature importance) berdasarkan

nilai Gini importance dari Random Forest.

11. (Opsional) Menyediakan kode tambahan untuk melakukan permutation

importance bila diperlukan.

12. Menyimpan model terbaik ke file “rf\_model.pkl” menggunakan joblib.

13. Menyediakan contoh penggunaan model untuk prediksi tunggal (single

inference):

- Mengisi nilai fitur seperti IPK, Jumlah\_Absensi,

Waktu\_Belajar\_Jam, IPK\_x\_Study, dan Rasio\_Absensi.

- Menampilkan hasil prediksi dengan label:

- 1 → “LULUS TEPAT WAKTU”

- 0 → “TIDAK LULUS TEPAT WAKTU / TERLAMBAT”

Kesimpulan: Kode ini membangun, mengevaluasi, dan menyimpan model

klasifikasi Random Forest terbaik untuk memprediksi kelulusan mahasiswa.

Prosesnya mencakup pembersihan data, validasi silang, tuning parameter,

evaluasi performa, serta pembuatan prediksi individual secara otomatis.