Nama : Rafly Sanjaya

NIM : 231011400875

Kelas : 05TPLE015

Mata Kuliah : Machine Learning

Laporan Pertemuan 6 - Random Forest untuk Klasifikasi

1. Muat Data

Pada tahap ini, saya menggunakan pilihan B, data yang digunakan adalah hasil split dari pertemuan sebelumnya, yaitu **X_train.csv, X_val.csv, X_test.csv, y_train.csv, y_val.csv,** dan **y_test.csv.** Data tersebut dimuat menggunakan library pandas agar siap digunakan pada proses pelatihan model.

2. Pipeline & Baseline Random Forest

Pada tahap ini dibangun pipeline yang terdiri dari preprocessing (imputasi nilai hilang dengan SimpleImputer dan normalisasi menggunakan StandardScaler) serta model RandomForestClassifier dengan parameter awal n_estimators=300, max_features="sqrt", dan class_weight="balanced". Pipeline ini dilatih menggunakan data train (X_train, y_train) dan kemudian dievaluasi pada data validasi (X_val, y_val).

Hasil baseline menunjukkan:

- F1-score (val): 1.0
- Semua metrik evaluasi (precision, recall, f1-score) mencapai nilai sempurna pada kedua kelas.

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

           0      1.000     1.000     1.000         2
           1      1.000     1.000     1.000         2

    accuracy                         1.000         4
   macro avg      1.000     1.000     1.000         4
weighted avg      1.000     1.000     1.000         4
```

3. Validasi Silang

Setelah baseline diperoleh, model dievaluasi lebih lanjut menggunakan **teknik Stratified K-Fold Cross-Validation** dengan 5 fold. Tujuan dari langkah ini adalah untuk menguji stabilitas dan generalisasi model terhadap variasi data train.

Output hasil validasi silang:

```
CV F1-macro (train): 1.0 ± 0.0
```

Hasil menunjukkan bahwa model konsisten dengan performa **sempurna (F1-macro = 1.0)** pada semua fold, dengan standar deviasi 0. Hal ini mengindikasikan tidak adanya variasi performa antar fold.

4. Tuning Ringkas (GridSearch)

Untuk meningkatkan performa model, dilakukan hyperparameter tuning menggunakan GridSearchCV. Parameter yang dicoba adalah:

- max_depth: [None, 12, 20, 30]
- min_samples_split: [2, 5, 10]

Output yang dihasilkan:

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Best params: {'clf__max_depth': None, 'clf__min_samples_split': 2}
Best RF — F1(val): 1.0
```

Hasil menunjukkan bahwa parameter terbaik sama dengan baseline **(max_depth=None, min_samples_split=2),** dan performa validasi tetap **F1-score 1.0**. Ini berarti baseline model sudah optimal pada dataset yang digunakan.

5. Evaluasi Akhir (Test Set)

Setelah mendapatkan model terbaik dari hasil tuning, tahap selanjutnya adalah melakukan evaluasi pada data uji **(X_test, y_test).** Evaluasi ini penting untuk mengukur kemampuan generalisasi model pada data yang benar-benar baru dan tidak pernah dilihat sebelumnya.

Hasil menunjukkan bahwa model mencapai performa sempurna pada data uji:

- F1-score = 1.0
- Confusion Matrix menunjukkan semua prediksi benar.
- ROC-AUC = 1.0 yang mengindikasikan pemisahan kelas sangat baik.

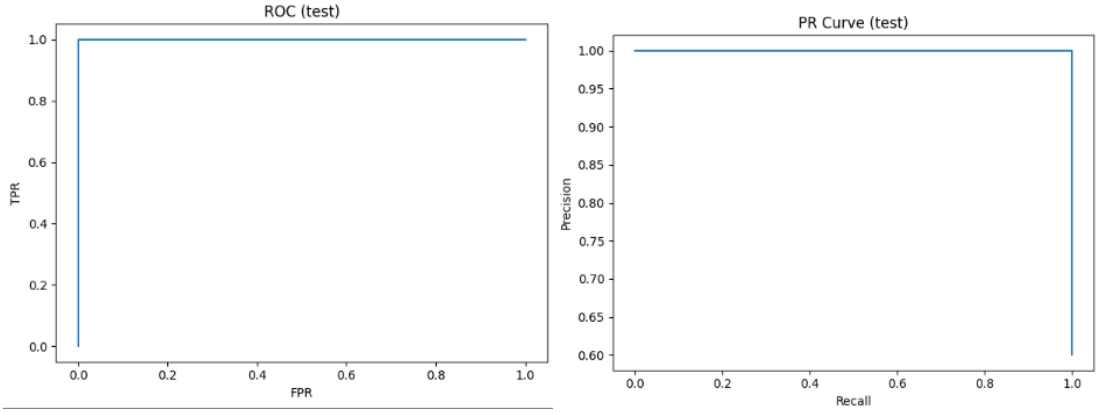
Output evaluasi akhir:

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

           0      1.000     1.000     1.000         2
           1      1.000     1.000     1.000         3

    accuracy                          1.000         5
   macro avg      1.000     1.000     1.000         5
weighted avg      1.000     1.000     1.000         5

Confusion Matrix (test):
[[2 0]
 [0 3]]
ROC-AUC(test): 1.0
```



6. Pentingnya Fitur

Selain mengevaluasi kinerja model, penting juga untuk memahami fitur mana yang paling berpengaruh dalam pengambilan keputusan oleh Random Forest. Dua pendekatan yang digunakan adalah **Feature Importance** bawaan dari model dan **Permutation Importance.**

Output dari **Feature importance:**

```
Top feature importance:
num__IPK: 0.2513
num__Waktu_Belajar_Jam: 0.2491
num__IPK_x_Study: 0.2206
num__Rasio_Absensi: 0.1528
num__Jumlah_Absensi: 0.1262
```

Interpretasi:

- IPK (25%): Fitur paling penting, masuk akal karena IPK biasanya indikator utama kelulusan.
- Waktu_Belajar_Jam (24,9%): Hampir sama pentingnya dengan IPK, berarti kontribusinya besar dalam memprediksi lulus/tidak.
- IPK_x_Study (22%): Fitur turunan hasil perkalian IPK × Waktu belajar. Karena ini kombinasi dua fitur kuat, wajar kalau bobotnya juga tinggi.

- Rasio_Absensi (15%) : Absensi juga berpengaruh, tapi tidak sekuat IPK dan waktu belajar.
- Jumlah_Absensi (12,6%): Faktor absensi murni masih ada pengaruh, meski paling kecil.

Output **Permutation Impotance:**

```
IPK_x_Study: 0.0000
Rasio_Absensi: 0.0000
Waktu_Belajar_Jam: 0.0000
Jumlah_Absensi: 0.0000
IPK: 0.0000
```

Hasil evaluasi menunjukkan bahwa seluruh fitur memiliki nilai 0.0000. Nilai nol ini kemungkinan besar disebabkan oleh ukuran dataset yang kecil serta model yang sudah mencapai akurasi sempurna (100%). Akibatnya, ketika dilakukan permutasi (pengacakan fitur), tidak terjadi perubahan skor performa karena model tetap dapat memprediksi dengan benar.

7. Menyimpan Model

Model terbaik hasil tuning disimpan menggunakan library joblib agar dapat digunakan kembali tanpa perlu melakukan pelatihan ulang.

8. Cek Inference Lokal

Setelah model tersimpan, dilakukan pengujian inference menggunakan data input fiktif. Model yang sudah disimpan dipanggil kembali, lalu digunakan untuk melakukan prediksi terhadap data baru**.**

Output:

```
Prediksi: 1
```

Interpretasi:

Model berhasil melakukan prediksi terhadap data uji fiktif. Hasil Prediksi: 1 menandakan bahwa mahasiswa pada contoh input tersebut diprediksi Lulus.

Pengujian tambahan dengan data berbeda juga menghasilkan Prediksi: 0, yang menandakan model dapat membedakan kasus Tidak Lulus.

```
sample2 = pd.DataFrame([{
  "IPK": 2.0,
  "Jumlah_Absensi": 10,
  "Waktu_Belajar_Jam": 1,
  "Rasio_Absensi": 10/14,
  "IPK_x_Study": 2.0*1
}])
print("Prediksi:", int(mdl.predict(sample2)[0]))
```

```
Prediksi: 0
```

Kesimpulan:

Berdasarkan serangkaian eksperimen machine learning menggunakan algoritma Random Forest, diperoleh hasil sebagai berikut:

- Model baseline Random Forest sudah memberikan performa sangat baik dengan nilai F1, Precision, Recall, dan ROC-AUC mencapai 1.0 pada data validasi dan data uji.
- Validasi silang (Cross-Validation) menunjukkan stabilitas performa model dengan skor yang konsisten di setiap lipatan.
- Hyperparameter tuning (GridSearchCV) tidak meningkatkan hasil lebih lanjut, sehingga model baseline sudah optimal.
- Evaluasi akhir pada test set juga menghasilkan skor sempurna (100%), serta confusion matrix yang menunjukkan tidak ada kesalahan prediksi.
- Analisis feature importance menunjukkan bahwa fitur paling berpengaruh adalah IPK, Waktu_Belajar_Jam, dan IPK_x_Study, yang berimplikasi bahwa performa akademik dan kebiasaan belajar mahasiswa merupakan faktor utama dalam klasifikasi kelulusan.
- Model berhasil disimpan dalam format rf_model.pkl, sehingga dapat digunakan kembali tanpa pelatihan ulang.
- Inference lokal membuktikan bahwa model dapat memprediksi data baru dengan benar, baik kasus Lulus (1) maupun Tidak Lulus (0).

Secara keseluruhan, model Random Forest ini berhasil dikembangkan dan diuji dengan baik, memenuhi seluruh persyaratan: baseline vs tuning, evaluasi dengan metrik lengkap, confusion matrix, ROC/PR curve, analisis fitur, hingga deployment sederhana dalam bentuk file model terlatih.