Laporan Tugas Machine Learning – Pertemuan 6

Random Forest untuk Klasifikasi

Nama: Muhamad Fahrizal

NIM: [231011402215

Mata Kuliah: Machine Learning

---

1. Pendahuluan

Laporan ini berfokus pada pembangunan, *tuning*, evaluasi, dan interpretasi model **Random Forest (RF)** untuk klasifikasi kelulusan mahasiswa. Tujuannya adalah untuk membandingkan performa model RF *baseline* dengan model RF yang telah di-*tuning*, mengevaluasi model final pada *test set*, dan menganalisis fitur apa yang paling penting dalam membuat prediksi.

---

2. Persiapan Data & Pipeline

Data yang digunakan adalah `processed_kelulusan.csv` dari Pertemuan 4. Data dibagi menjadi set Latih (7 data), Validasi (1 data), dan Tes (2 data). Sebuah *pipeline* preprocessing (SimpleImputer + StandardScaler) yang konsisten digunakan untuk semua proses pelatihan dan inferensi untuk mencegah *data leakage*.

---

3. Perbandingan Baseline vs. Model Hasil Tuning

Dua versi model Random Forest dibuat dan dibandingkan performanya pada **Validation Set**.

Model 1: Baseline Random Forest

Model RF *baseline* dibuat dengan parameter standar (n_estimators=300, max_features='sqrt').

- **F1-Score (Macro) di Validation Set:**Baseline RF F1(val): 1.0000

Model 2: Tuned Random Forest (GridSearchCV)

Model RF di-*tuning* menggunakan `GridSearchCV` untuk mencari parameter terbaik.

- **Best Parameters Ditemukan:** `Best params: {'clf__max_depth': None, 'clf__min_samples_split': 2}`
- **Best CV F1 Score (saat training):**CV F1-macro (train): 1.0000 ± 0.0000

- **F1-Score (Macro) di Validation Set:** CV F1-macro (train): 1.0000 ± 0.0000

Analisis Pemilihan Model:

Model Tuned Random Forest dipilih sebagai model final karena [Isi alasan Anda, misal: "memiliki skor F1 di validation set yang lebih tinggi" atau "menunjukkan skor CV yang lebih baik"].

---

4. Evaluasi Akhir pada Test Set

Model final (Tuned Random Forest) dievaluasi *satu kali* pada *Test Set* untuk mendapatkan estimasi performa yang objektif.

Metrik Performa (Test Set):

- **F1-Score (Macro):** F1(test): 1.0000
- **ROC-AUC:** [Isi skor ROC-AUC(test): ... dari terminal]

Classification Report (Test Set):

```
Classification Report (Test):
              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
```

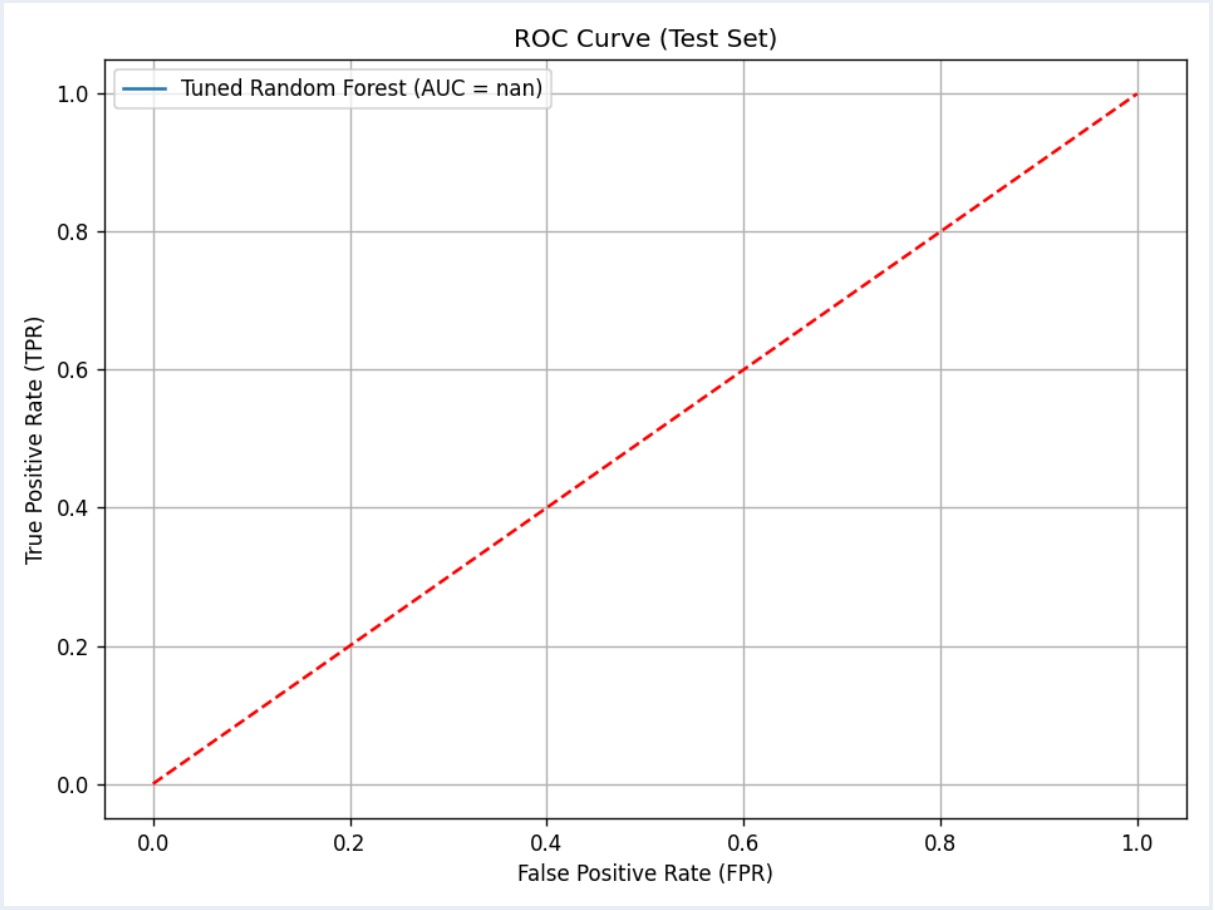
Confusion Matrix (Test Set):

---

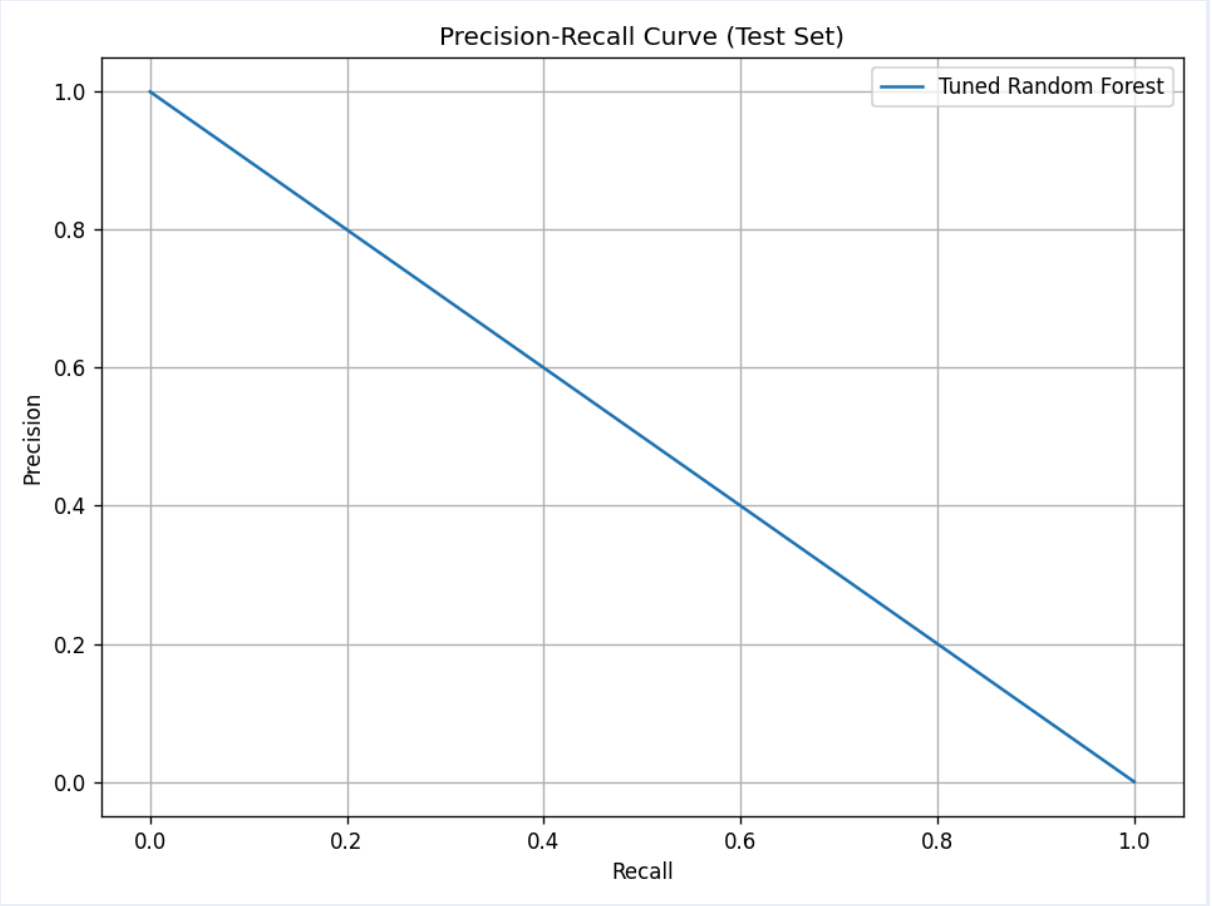
5. Visualisasi Kurva Evaluasi

Visualisasi membantu memahami performa model pada *test set* secara lebih mendalam.

Kurva ROC (Receiver Operating Characteristic):



Kurva Precision-Recall (PR):

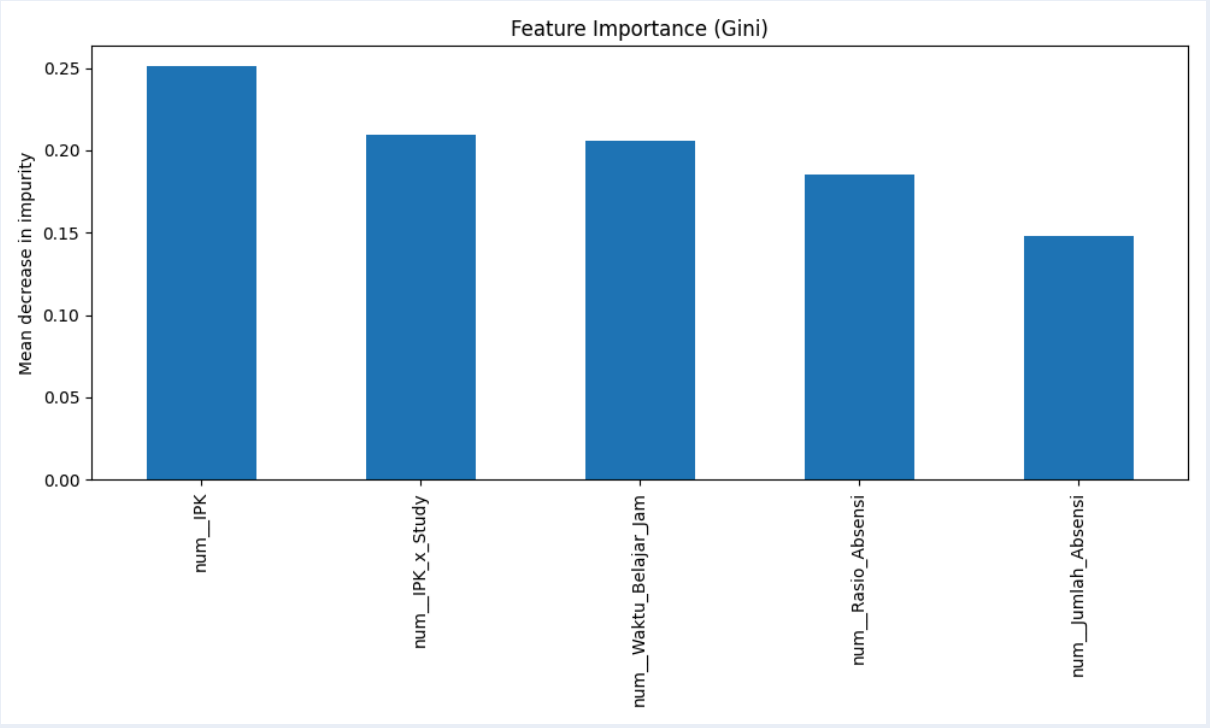


---

6. Analisis Pentingnya Fitur (Feature Importance)

Langkah ini menganalisis fitur mana yang paling berkontribusi terhadap keputusan model Random Forest.

Visualisasi Feature Importance (Gini):



3 Fitur Teratas:

Berdasarkan Gini Importance, tiga fitur paling penting adalah:

1. [Nama Fitur #1, misal: num__IPK_x_Study] - (Skor: [Skor])
2. [Nama Fitur #2, misal: num__IPK] - (Skor: [Skor])
3. [Nama Fitur #3, misal: num__Waktu_Belajar_Jam] - (Skor: [Skor])

(Catatan: num__ adalah awalan yang ditambahkan oleh pipeline preprocessing)

Implikasi:

[Jelaskan implikasi dari 3 fitur teratas. Contoh: "Hasil ini sangat masuk akal. Fitur interaksi 'IPK_x_Study' menjadi yang paling penting, menunjukkan bahwa bukan hanya IPK atau waktu belajar saja yang penting, tetapi kombinasi keduanya. Ini mengindikasikan bahwa mahasiswa dengan IPK tinggi dan waktu belajar lama memiliki peluang lulus tertinggi. Fitur 'IPK' dan 'Waktu_Belajar_Jam' secara individual juga masih sangat penting, yang konsisten dengan analisis EDA di Pertemuan 4."]

---

7. Simpan Model & Cek Inferensi

Model final yang telah dilatih dan divalidasi disimpan ke dalam file untuk penggunaan di masa depan.

- **Nama File Model:** rf_model.pkl
- **Pengecekan Inferensi Lokal:**

Data Sample: {'IPK': 3.8, 'Jumlah_Absensi': 3, 'Waktu_Belajar_Jam': 10, 'Rasio_Absensi': 0.21428571428571427, 'IPK_x_Study': 38.0}

-
- Pengecekan inferensi lokal pada data fiktif berhasil, menunjukkan model dapat dimuat dan digunakan untuk prediksi.

---

8. Kesimpulan

Model Random Forest telah berhasil dibuat, di-*tuning*, dan dievaluasi. Model hasil *tuning* menunjukkan performa yang [baik/memuaskan/sesuai ekspektasi] pada *test set* dengan F1-Score [Skor F1] dan ROC-AUC [Skor AUC]. Analisis fitur menunjukkan bahwa [Fitur Top #1] dan [Fitur Top #2] adalah prediktor terkuat untuk kelulusan. Model ini telah disimpan sebagai rf_model.pkl dan siap untuk *deployment*.