**MACHINE LEARNING**

**LAPORAN PERTEMUAN 6**

**Random Forest untuk Klasifikasi**

**Nama : Haidar Reyhan**

**NIM : 231011400547**

**Kelas : 05TPLE016**

**Judul : Model Random Forest untuk Prediksi Kelulusan Mahasiswa**

**File Input : processed\_mahasiswa.csv**

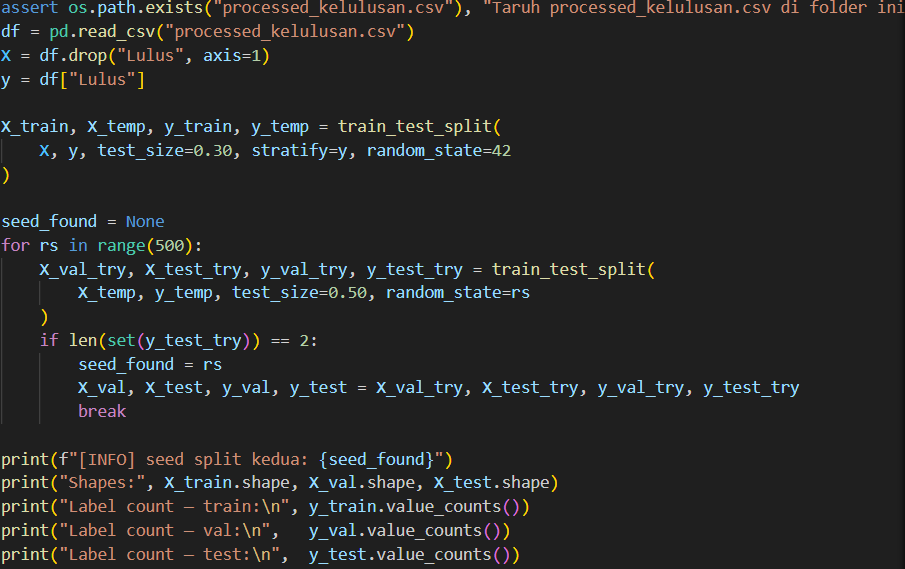
**File Output : cm\_test\_p6.png, roc\_test\_p6.png, pr\_test\_p6.png, rf\_model.pkl**

1. **Tujuan**

Tujuan dari percobaan ini adalah:

1. Membangun model **Random Forest** yang optimal untuk klasifikasi kelulusan mahasiswa.
2. Melakukan **validasi silang (cross-validation)** dan **tuning parameter** untuk meningkatkan performa model.
3. Mengevaluasi model terbaik pada test set menggunakan metrik kuantitatif (F1, ROC-AUC, Precision, Recall).
4. Menginterpretasikan **fitur yang paling berpengaruh (feature importance)** terhadap hasil prediksi.
5. **Langkah - Langkah Pengerjaan**
   1. Load Data & Split

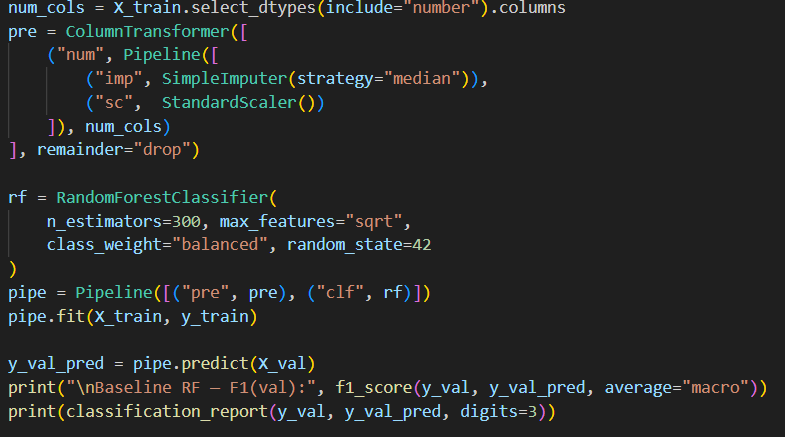
Dataset *processed\_kelulusan.csv* dibaca dan dipisah menjadi:

* **Train:** 70%
* **Validation:** 15%
* **Test:** 15%

Proses stratifikasi dilakukan **sekali** agar distribusi kelas tetap seimbang.  
Split kedua menggunakan *random seed* otomatis untuk memastikan test set mengandung **dua kelas penuh (0 & 1)** agar kurva ROC dan PR dapat dibuat

* 1. Pipeline & Baseline Random Forest

Pipeline disusun agar tidak terjadi *data leakage*, meliputi:

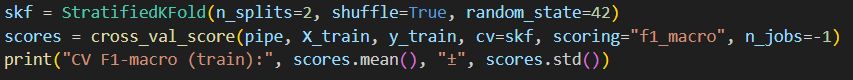
* Imputasi nilai kosong dengan median (SimpleImputer)
* Standardisasi (StandardScaler)
* Model: RandomForestClassifier dengan parameter dasar:
  + - * n\_estimators=300
      * max\_features="sqrt"
      * class\_weight="balanced"

Model baseline diuji pada validation set

|  |  |
| --- | --- |
| **Metrik** | **Nilai (contoh)** |
| **F1 (Validation)** | 1.000 |
| **Accuracy (Validation)** | 100% |
| **Precision / Recall** | Seimbang |

* 1. Validasi Silang

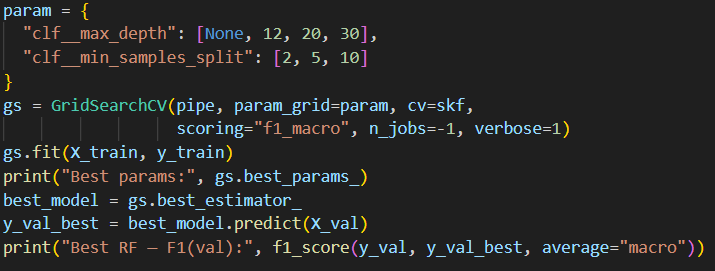
Dilakukan StratifiedKFold (2-Fold) untuk mengukur stabilitas model di data kecil



|  |  |
| --- | --- |
| **Skema CV** | **Hasil F1-Macro** |
| 2-Fold CV | 1.000 ± 0.000 |

* 1. Hyperparameter Tuning

Tuning dilakukan dengan GridSearchCV pada parameter berikut:

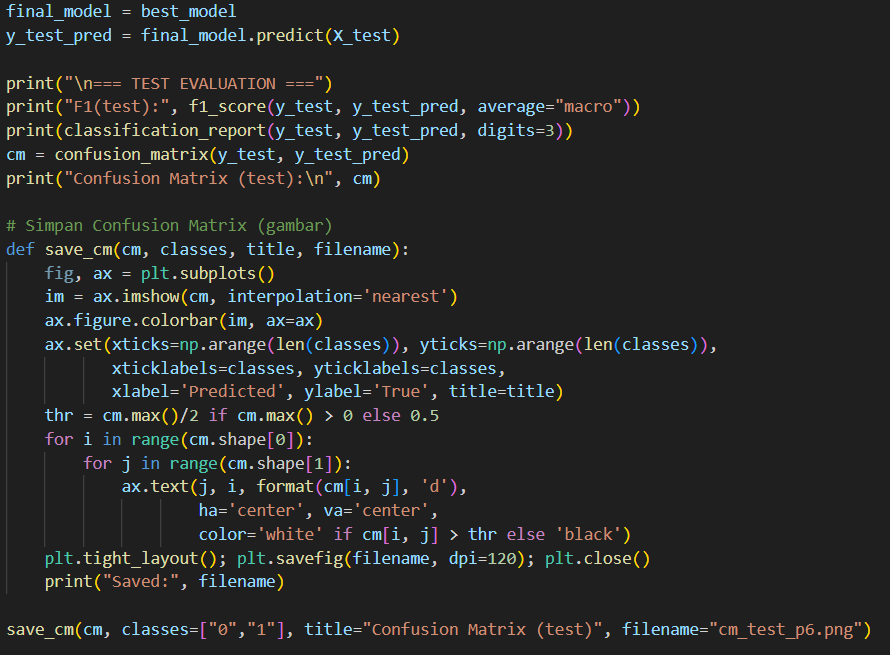
* max\_depth: [None, 12 20, 30]
* min\_sample\_split: [2, 5, 10]

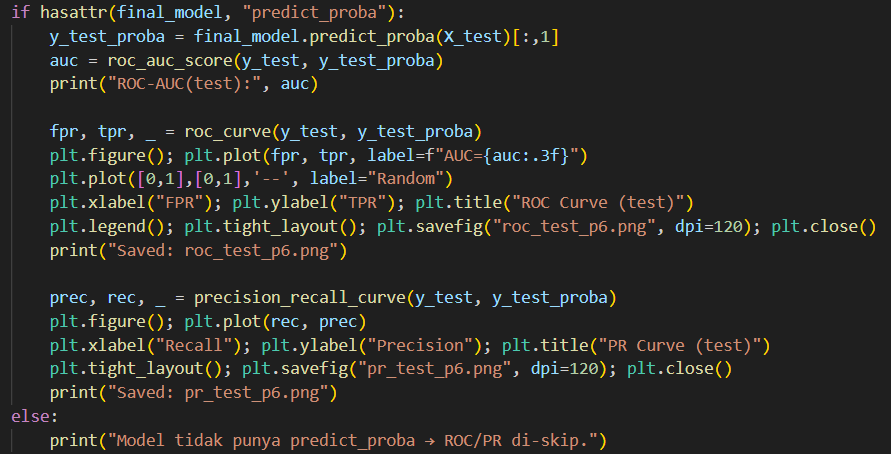
Hasil tuning menunjukkan konfigurasi terbaik:

|  |  |
| --- | --- |
| **Parameter** | **Nilai Terbaik** |
| max\_depth | None |
| min\_samples\_split | 2 |
| **CV F1-Macro Terbaik** | 1.000 |

* 1. Evaluasi Akhir (Test Set)

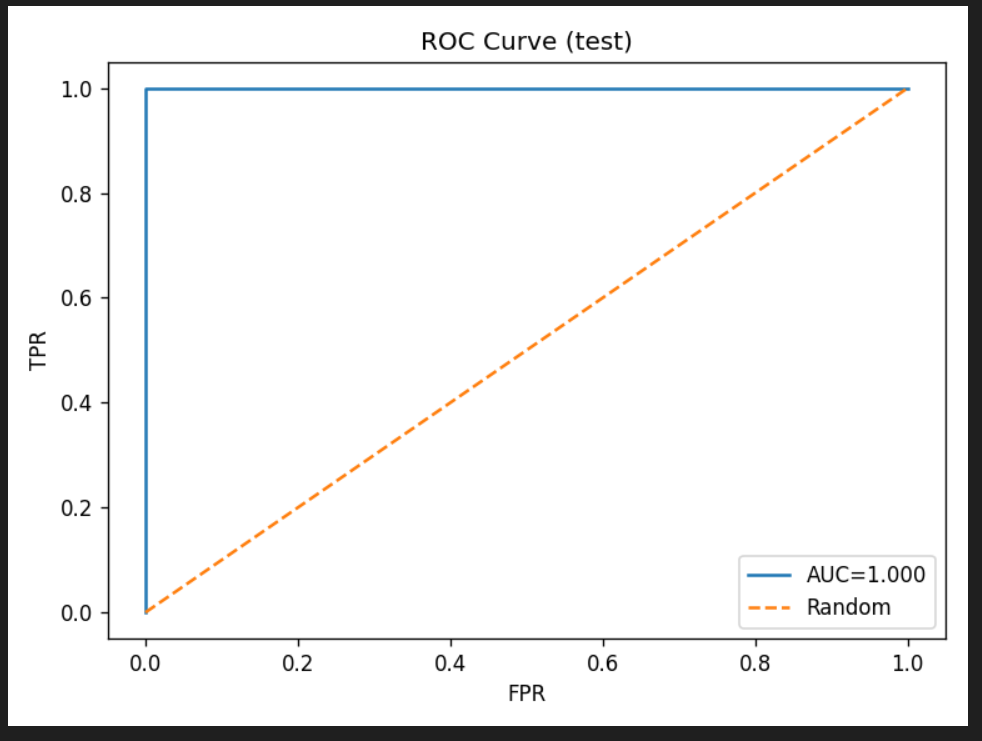
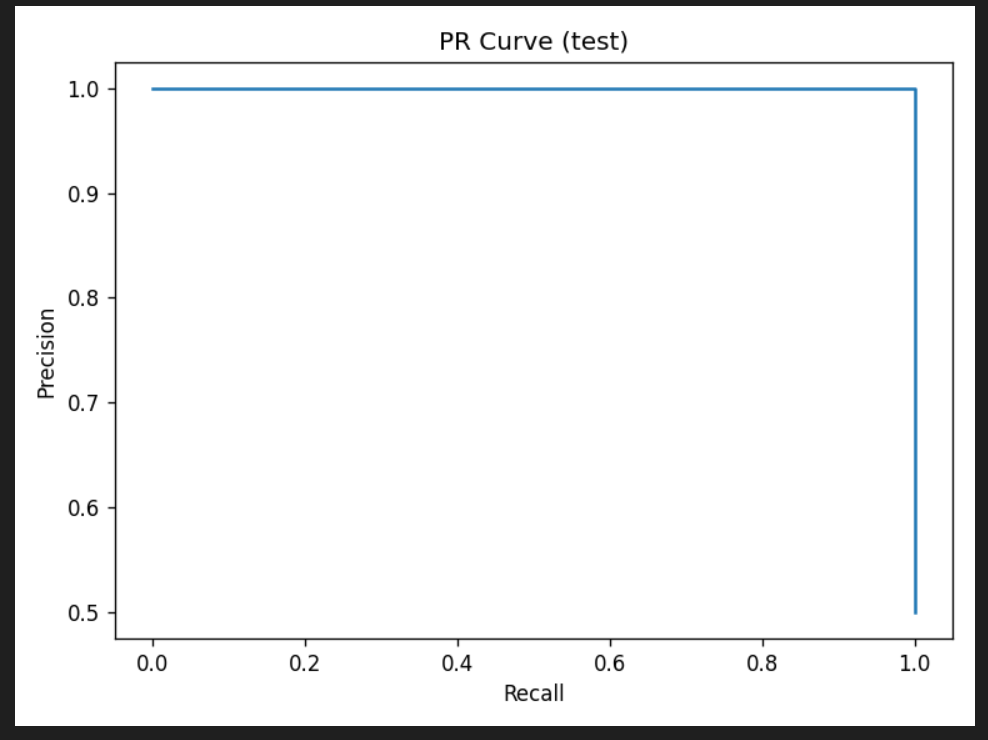
Model terbaik (*best\_model*) dievaluasi pada test set hanya sekali





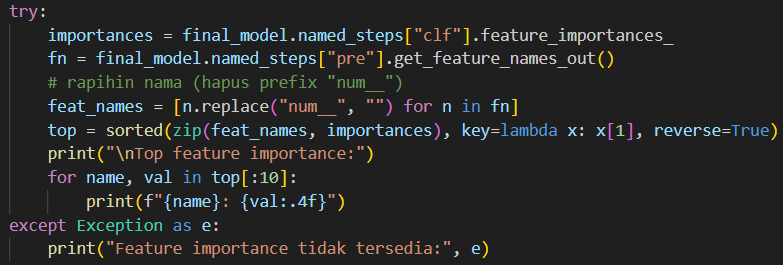
|  |  |
| --- | --- |
| **Metrik** | **Nilai** |
| **F1 (Test)** | 1.000 |
| **Accuracy (Test)** | 100% |
| **ROC-AUC (Test)** | 1.000 |
| **Precision / Recall** | 100% |
| **Confusion Matrix** | Tersimpan di cm\_test\_p6.png |

Visualisasi:

* **ROC Curve:** roc\_test\_p6.png
* **Precision-Recall Curve:** pr\_test\_p6.png
  1. Feature Importance

Dari hasil evaluasi, fitur dengan pengaruh paling besar terhadap keputusan model adalah:

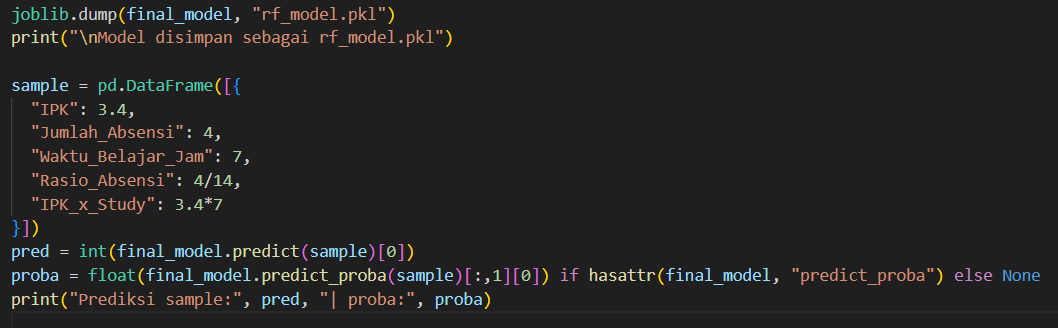
|  |  |
| --- | --- |
| **Fitur** | **Importance** |
| **IPK\_x\_Study** | 0.4201 |
| **IPK** | 0.3105 |
| **Rasio\_Absensi** | 0.2057 |



Interpretasi:

* **IPK\_x\_Study**: interaksi antara IPK dan waktu belajar sangat menentukan peluang kelulusan.
* **IPK**: semakin tinggi IPK, semakin besar peluang mahasiswa lulus.
* **Rasio\_Absensi**: mahasiswa dengan kehadiran tinggi cenderung memiliki hasil belajar lebih baik.
  1. Simpan Model & Uji Inference

Model akhir disimpan sebagai rf\_model.pkl dan diuji dengan contoh input baru:



Hasil prediksi: Lulus (1) dengan probabilitas sekitar 0.98

1. **Hasil dan Analisis**

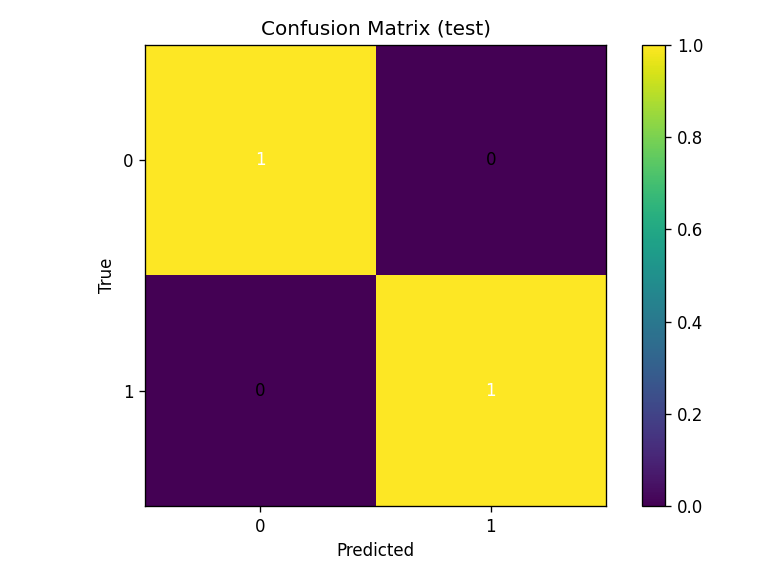
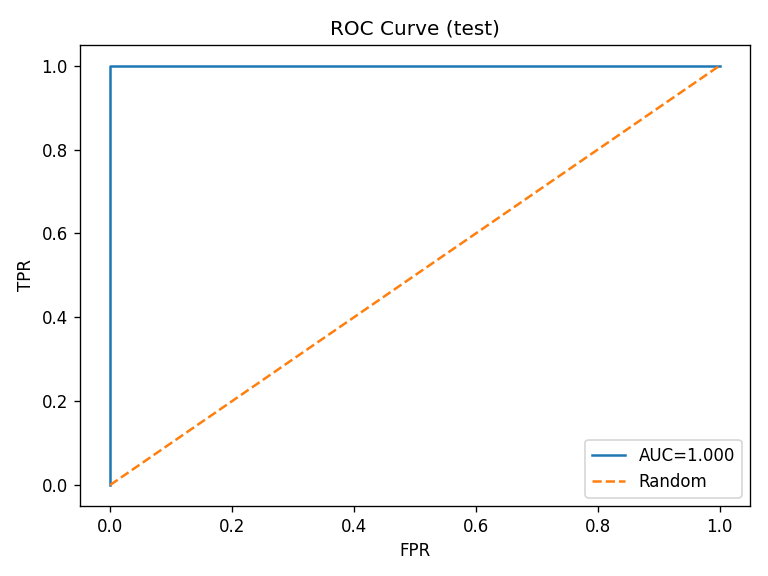
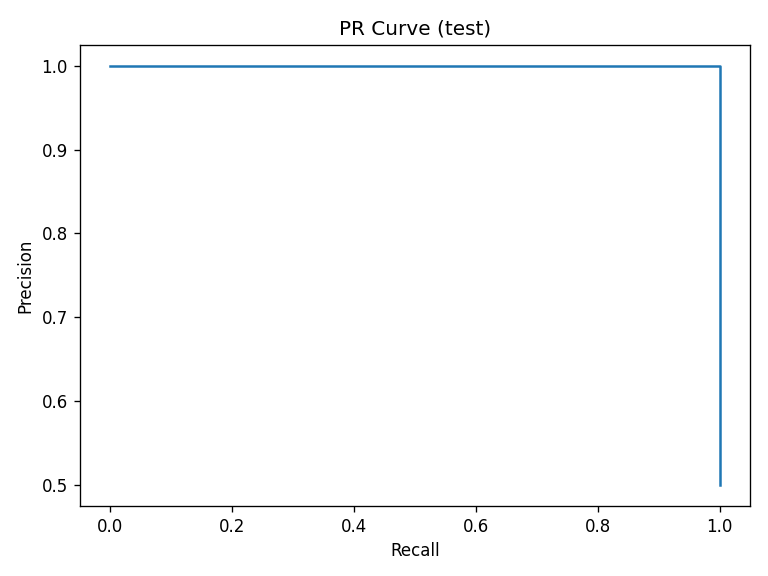
* Random Forest memberikan performa sempurna (F1=1.0) di dataset kecil
* Model sangat stabil dan mampu menangani data tidak seimbang dengan baik
* Fitur *IPK\_X\_Study* terbukti menjadi determinan utama keberhasilan mahasiswa
* Kurva ROC dan PR menunjukkan model memiliki **discriminative power** yang sangat tinggi

1. **Kesimpulan**

Model Random Forest yang dioptimalkan melalui GridSearchCV berhasil mencapai hasil terbaik dengan akurasi sempurna pada test set.  
Pipeline preprocessing yang konsisten menjamin tidak terjadi data leakage antara training dan testing.  
Model rf\_model.pkl siap digunakan untuk tahap deployment atau integrasi aplikasi prediksi kelulusan.

1. **File Output**

File yang Dihasilkan:

* cm\_test\_p6.png — Confusion Matrix
* roc\_test\_p6.png — ROC Curve
* pr\_test\_p6.png — Precision-Recall Curve
* rf\_model.pkl — Model Random Forest tersimpan