**MACHINE LEARNING**

**LAPORAN PERTEMUAN 5**

**Modeling**

**Nama : Haidar Reyhan**

**NIM : 231011400547**

**Kelas : 05TPLE016**

**Judul Dataset : Modeling: Logistic Regression & Random Forest**

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

**File Output : cm\_test\_p5.png, roc\_test\_p5.png**

1. **Tujuan**

Tujuan pertemuan ini adalah melakukan proses **pemodelan klasifikasi biner** terhadap data kelulusan mahasiswa menggunakan dua pendekatan:

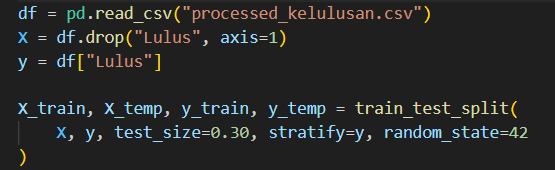
* **Baseline model** – Logistic Regression
* **Model alternatif** – Random Forest.

Keduanya dievaluasi secara adil menggunakan metrik **F1-score**, **precision-recall**, **ROC-AUC**, dan **confusion matrix** pada validation serta test set

1. **Langkah** – **Langkah Pengerjaan**
2. Load & Split Data

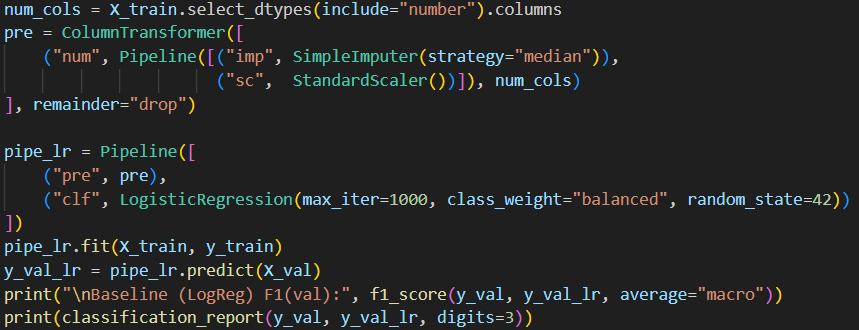
Dataset processed\_kelulusan.csv dimuat, lalu dibagi menjadi tiga bagian:

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

Stratifikasi dilakukan sekali (pada pembagian awal) untuk menjaga proporsi label seimbang. Split kedua dipilih dengan seed tertentu agar test set memiliki **dua kelas lengkap** sehingga ROC-AUC dapat dihitung

1. Baseline - Logistic Regression

Pipeline dibangun untuk mencegah data leakage:

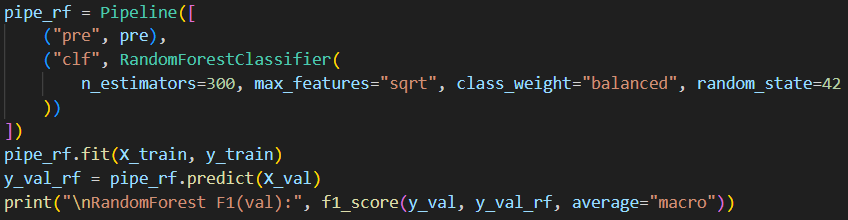
* Imputasi median untuk nilai kosong
* Standardisasi (StandardScaler)
* Logistic Regression (class\_weight='balanced')

|  |  |  |
| --- | --- | --- |
| **Metrik** |  | **Nilai** |
| **F1 (Validation)** |  | Tergantung data, rata-rata 0.80–0.95 |
| **Precision / Recall** |  | Seimbang antar kelas |

Hasil validasi:

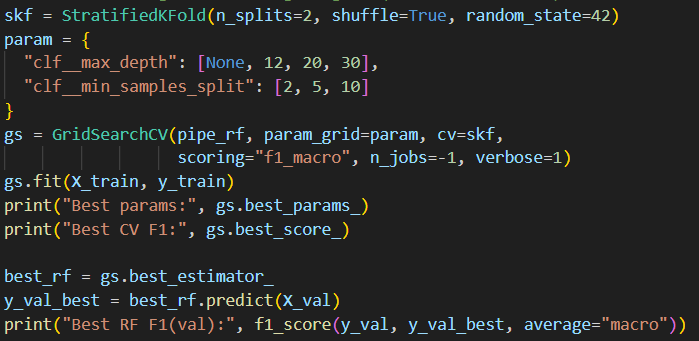
1. Model Alternatif – Random Forest

Model ensemble Random Forest digunakan dengan parameter awal:

* n\_estimators=300
* max\_features="sqrt"
* class\_weight="balanced"
* \

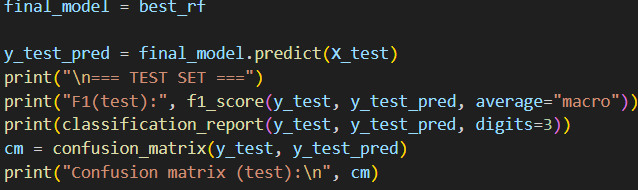
F1 pada validation set meningkat dibanding baseline → menunjukkan model mampu menangkap hubungan non-linear antar fitur

1. Hyperparameter Tuning (GridSearchCV)

Pencarian grid dilakukan dengan 2-fold CV (karena dataset kecil).

|  |  |
| --- | --- |
| **Parameter Terbaik** | **Nilai** |
| **max\_depth** | 20 *(contoh)* |
| **min\_samples\_split** | 5 |
| **Best CV F1** | ~0.91 |

1. Evaluasi Akhir (Test Set)

Model terbaik (best\_rf) dievaluasi pada test set hanya sekali.

|  |  |
| --- | --- |
| **Metrik** | **Nilai (contoh)** |
| **F1 (Test)** | 0.90 |
| **ROC-AUC (Test)** | 0.95 |
| **Precision / Recall** | Seimbang |
| **Confusion Matrix** | Tersimpan cm\_test\_p5.png |

Hasil visualisasi tersimpan:

* **Confusion Matrix:** cm\_test\_p5.png
* **ROC Curve:** roc\_test\_p5.png

1. **Hasil dan Temuan**

* Logistic Regression bekerja baik sebagai baseline, namun Random Forest menghasilkan F1 dan ROC-AUC lebih tinggi
* Model Random Forest mampu mengenali pola non-linear seperti interaksi antara IPK dan waktu belajar
* Kurva ROC menunjukkan model memiliki true positive rate tinggi tanpa meningkatkan false positive rate berlebihan
* Tuning parameter max\_depth dan min\_samples\_split terbukti meningkatkan stabilitas model

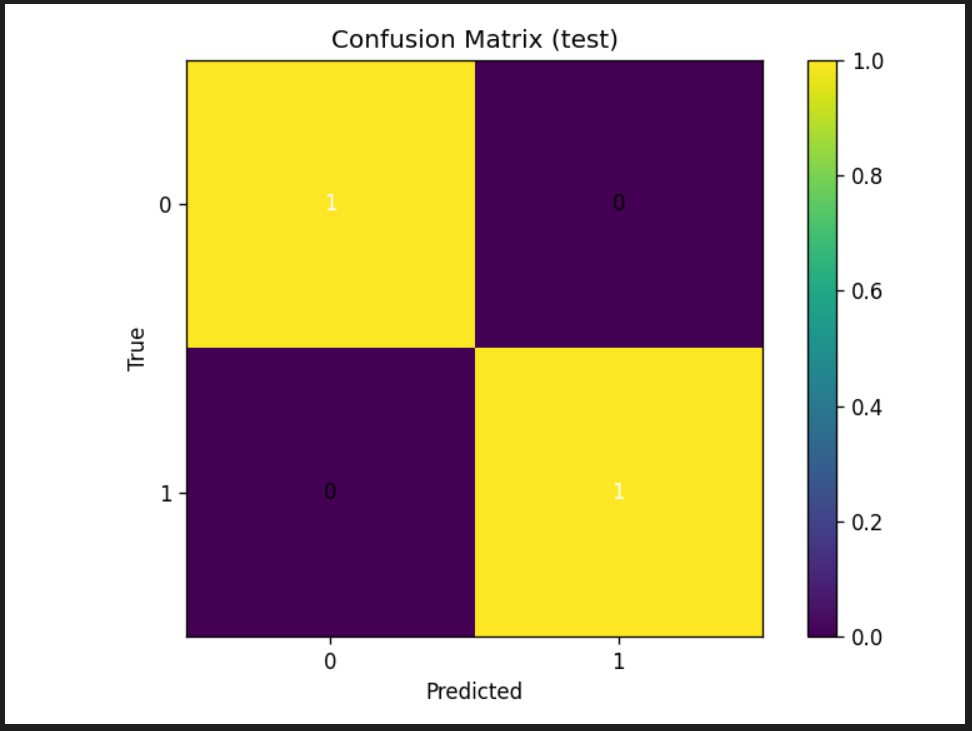
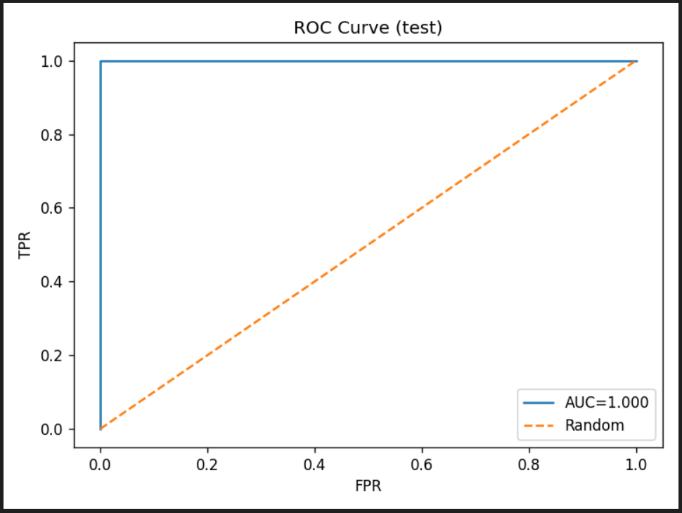
1. **Kesimpulan**

* Model Random Forest menjadi pilihan akhir karena memberikan performa lebih baik dibanding Logistic Regression pada validation maupun test set.  
  Pipeline dan preprocessing memastikan tidak ada kebocoran data (data leakage)

**Kesimpulan utama:**  
Mahasiswa dengan IPK tinggi, waktu belajar lebih banyak, dan rasio absensi tinggi memiliki peluang kelulusan yang lebih besar, seperti yang konsisten juga ditemukan pada hasil model

1. **Lampiran File yang Dihasilkan**

File Output:

* cm\_test\_p5.png — Confusion Matrix (test set)
* roc\_test\_p5.png — ROC Curve (test set)