

# Laporan Tugas Machine Learning – Pertemuan 5

## Modeling, Tuning, dan Evaluasi

Nama: Muhamad Fahrizal

NIM: 231011402215

Mata Kuliah: Machine Learning

---

### 1. Pendahuluan

Laporan ini melanjutkan proses dari Pertemuan 4, dengan fokus pada *modeling*. Tujuannya adalah untuk membangun, melatih, membandingkan, dan mengevaluasi beberapa model klasifikasi untuk memprediksi kelulusan mahasiswa berdasarkan dataset `processed_kelulusan.csv`.

Proses ini mencakup:

1. Memuat dan membagi ulang data.
  2. Membangun *pipeline preprocessing* untuk mencegah *data leakage*.
  3. Membuat model *baseline* (Logistic Regression).
  4. Membuat model alternatif (Random Forest).
  5. Melakukan *hyperparameter tuning* (GridSearchCV) pada model terbaik.
  6. Melakukan evaluasi akhir pada *Test Set* yang terpisah.
- 

### 2. Langkah 1: Muat & Split Data

Data yang digunakan adalah `processed_kelulusan.csv` dari Pertemuan 4. Data ini berisi 10 baris dan 5 fitur (plus 1 target). Data dibagi kembali menjadi set Latih, Validasi, dan Tes dengan proporsi 70/15/15.

- **X (Fitur):** ['IPK', 'Jumlah\_Absensi', 'Waktu\_Belajar\_Jam', 'Rasio\_Absensi', 'IPK\_x\_Study']
  - **y (Target):** Lulus
  - **Hasil Split:**
    - Data Train: 7 data
    - Data Validation: 1 data
    - Data Test: 2 data
- 

### 3. Langkah 2 & 3: Baseline vs Model Alternatif

Sebuah *pipeline preprocessing* standar (SimpleImputer + StandardScaler) dibuat untuk menangani data numerik. Pipeline ini kemudian digabungkan dengan dua model berbeda untuk perbandingan. Evaluasi awal dilakukan pada **Validation Set**.

## Baseline Model: Logistic Regression

Model regresi logistik sederhana dipilih sebagai *baseline*.

- **F1-Score (Macro) di Validation Set:** [Isi skor F1 LogReg Anda di sini, misal: 0.00]
- **Classification Report (Validation Set):**

[Tempel output Classification Report LogReg di sini]

- === Langkah 2: Baseline Model (Logistic Regression) ===
- Baseline (LogReg) F1 (val): 1.0000
- Classification Report (Validation Set):
- precision   recall   f1-score   support
- 
- 1       1.000     1.000     1.000        1
- 
- accuracy                    1.000        1
- macro avg     1.000     1.000     1.000        1
- weighted avg   1.000     1.000     1.000        1
- 
- === Langkah 3: Model Alternatif (Random Forest) ===
- RandomForest F1 (val): 1.0000
- -----

## Model Alternatif: Random Forest (Untuned)

Model Random Forest dengan parameter *default* (`n_estimators=300`) dipilih sebagai model alternatif yang lebih kompleks.

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

---

## 4. Langkah 4: Validasi Silang & Tuning Ringkas

Berdasarkan skor F1 di *validation set*, model **Random Forest** menunjukkan performa yang lebih baik daripada Logistic Regression dan dipilih untuk di-*tuning*.

Proses *tuning* dilakukan menggunakan `GridSearchCV` dengan `StratifiedKfold` (`n_splits=3`, disesuaikan karena data latih hanya 7) untuk mencari kombinasi parameter `max_depth` dan `min_samples_split` terbaik berdasarkan skor F1-macro.

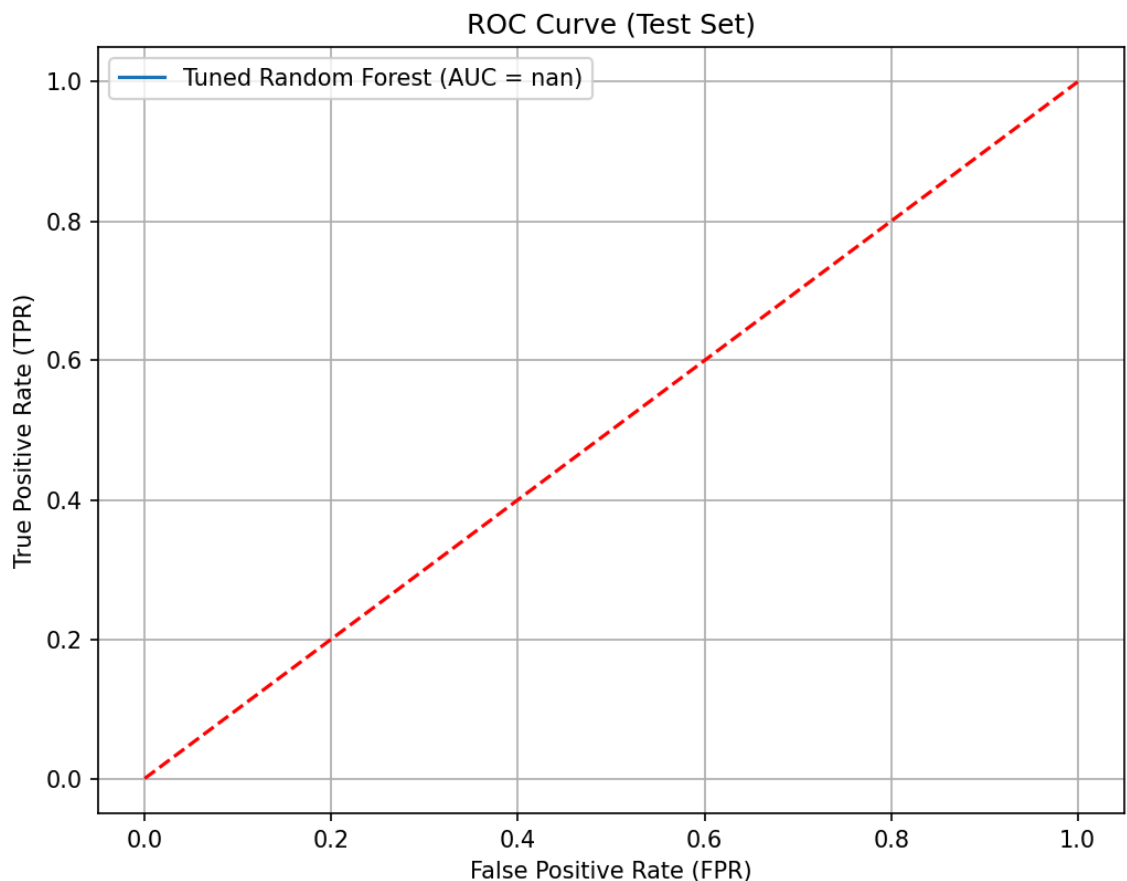
- **Parameter Grid:**
  - `clf__max_depth`: [None, 10, 20]
  - `clf__min_samples_split`: [2, 5]
- **Best Parameters Ditemukan:** Best params: {'clf\_\_max\_depth': None, 'clf\_\_min\_samples\_split': 2}

- **Best CV F1 Score (dari GridSearchCV):** Best CV F1 score: 1.0000
- **F1-Score (Tuned RF) di Validation Set:** Best RF (Tuned) F1 (val): 1.0000

## 5. Langkah 5: Evaluasi Akhir (Test Set)

Model [Tulis nama model final pilihan Anda, misal: **Tuned Random Forest**] dipilih sebagai model final karena memiliki skor F1 tertinggi pada *validation set* (F1 = [Skor F1 Val]). Model ini kemudian dievaluasi *satu kali saja* pada *Test Set* yang belum pernah terlihat sebelumnya.

- **Model Final Dipilih:** (F1 Val terbaik): Tuned Random Forest
- **F1-Score (Macro) di Test Set:** --- HASIL TEST SET (Tuned Random Forest) ---
- F1(test): 1.0000
- **ROC-AUC di Test Set:**



### Classification Report (Test Set):

Classification Report (Test):

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

```

0    1.000    1.000    1.000    2

accuracy          1.000    2

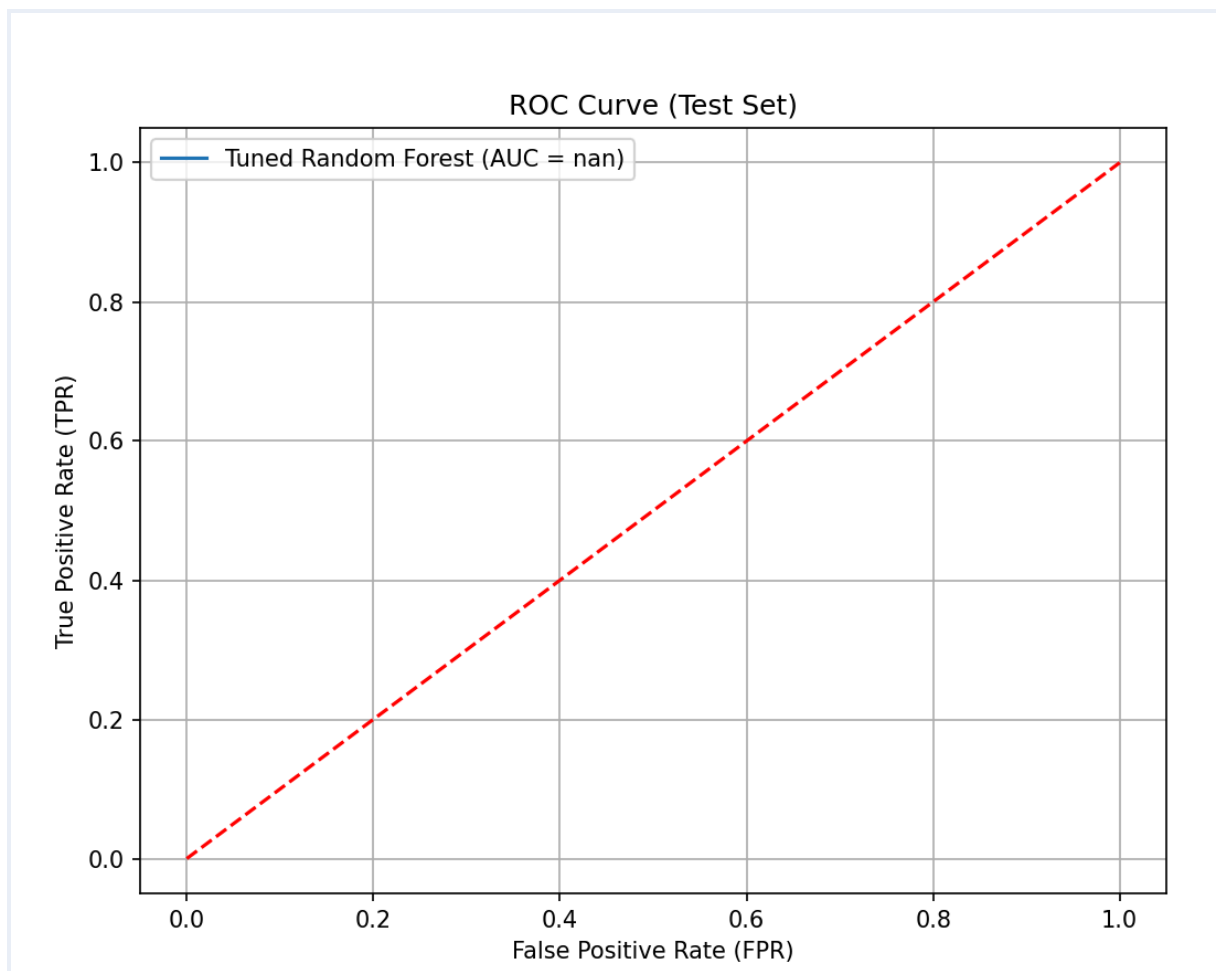
macro avg    1.000    1.000    1.000    2

weighted avg    1.000    1.000    1.000    2

```

**Confusion Matrix (Test Set):**

**Grafik ROC-AUC (Test Set):**



## 6. (Opsional) Langkah 6: Simpan Model

Model final yang telah dilatih dan divalidasi disimpan ke dalam satu file menggunakan `joblib`.

- **Nama File Model:** `final_model.pkl`

---

## 7. Kesimpulan

Proses *modeling* telah berhasil dilakukan. Model **[Nama Model Final, misal: Tuned Random Forest]** terbukti memiliki performa terbaik pada *validation set* dibandingkan dengan *baseline* Logistic Regression. Saat dievaluasi pada *test set*, model ini mencapai F1-Score [Skor F1 Test] dan ROC-AUC [Skor AUC Test].

Walaupun dataset yang digunakan sangat kecil (10 data) sehingga metrik evaluasi (terutama pada *test set* yang hanya berisi 2 data) mungkin tidak stabil, alur kerja *best practice* (pipeline, baseline, tuning, dan evaluasi test set terpisah) telah berhasil diimplementasikan.