

Laporan Singkat: Analisis Mendalam Model (Pertemuan 6)

Tanggal: 25 Oktober 2025 Disusun oleh: [Nama Anda] Mata Kuliah: [Mata Kuliah Anda]

1. Pendahuluan

Laporan ini mendokumentasikan langkah-langkah yang diambil dalam Pertemuan 6, yang berfokus pada analisis mendalam, validasi silang (*cross-validation*), dan interpretasi model *machine learning*.

Melanjutkan dari Pertemuan 5, laporan ini menggunakan model `RandomForestClassifier` sebagai studi kasus. Tujuannya adalah untuk memvalidasi robustitas model, menganalisis kinerjanya secara lebih detail (termasuk kurva Precision-Recall), dan menginterpretasi fitur apa yang paling berpengaruh dalam prediksi.

2. Langkah 1: Pemuatan Data dan Pemisahan (Splitting)

- Pemuatan Data:** Dataset `processed_kelulusan.csv` (hasil dari Pertemuan 4) berhasil dimuat.
- Pemisahan Data:** Dataset dibagi menjadi tiga bagian menggunakan `train_test_split` dengan strategi `stratify=y` (identik dengan Pertemuan 5):
 - Train Set (70%)
 - Validation Set (15%)
 - Test Set (15%)

3. Langkah 2 & 3: Pipeline, Baseline, dan Validasi Silang

3.1. Pipeline dan Baseline

Sebuah *pipeline* standar (`SimpleImputer` + `StandardScaler`) dan model `RandomForestClassifier` (baseline) dibuat. Model ini dilatih pada *train set* dan dievaluasi pada *validation set* untuk mendapatkan skor F1-Macro awal.

- Hasil (Validation Set):** [Isi skor F1-Macro baseline RF dari output Anda]

3.2. Validasi Silang (Cross-Validation)

Untuk mengukur seberapa baik model dapat digeneralisasi pada data baru, **Validasi Silang Stratified K-Fold (k=5)** dilakukan pada **train set**.

- **Tujuan:** Menguji model pada 5 "lipatan" data latih yang berbeda untuk mendapatkan estimasi performa yang lebih stabil dan mengurangi bias dari satu kali *split*.
- **Hasil:** Model *baseline* RF mencatat skor CV F1-Macro rata-rata.
 - **CV F1-Macro (train):** [Isi skor rata-rata \pm std. dev. dari output Anda, misal: 0.969... \pm 0.060...]

4. Langkah 4: Tuning Model (GridSearch)

Model `RandomForestClassifier` dioptimasi menggunakan `GridSearchCV` (dengan `cv=skf` dari langkah 3) untuk menemukan kombinasi parameter terbaik (`max_depth` dan `min_samples_split`).

- **Hasil Tuning (GridSearch):**
 - **Best Params:** [Isi parameter terbaik dari output Anda, misal: `{'clf__max_depth': None, 'clf__min_samples_split': 2}`]
 - **Best CV F1:** [Isi skor F1 terbaik dari output Anda, misal: 0.969...]
- **Hasil (Validation Set):** Model terbaik hasil *tuning* dievaluasi pada *validation set* dan menunjukkan performa yang kuat.
 - **Best RF F1 (val):** [Isi skor F1-Macro validasi dari output Anda]

5. Langkah 5: Evaluasi Akhir pada Test Set

Model terbaik (`best_model`) dari hasil `GridSearchCV` dipilih sebagai model final dan dievaluasi pada **test set**.

- **Metrik Kinerja (Test Set):**
 - **F1 Macro:** [Isi skor F1-Macro test dari output Anda]
 - **ROC-AUC:** [Isi skor ROC-AUC test dari output Anda]
 - **Confusion Matrix:** (Salin hasil *Confusion Matrix (test)* dari output Anda di sini)
 - **Classification Report:** (Salin hasil *Classification Report (test)* dari output Anda di sini)
- **Visualisasi Kinerja:**
 - **p6_roc_test.png (Kurva ROC):** Plot ini menunjukkan kemampuan model membedakan kelas. Nilai AUC yang tinggi mengkonfirmasi kinerja model yang sangat baik.
 - **p6_pr_test.png (Kurva Precision-Recall):** Plot ini sangat penting untuk data *imbalanced* (meskipun data kita seimbang). Kurva yang "tinggi dan ke kanan" menunjukkan model mampu mempertahankan *Precision* tinggi seiring dengan meningkatnya *Recall*.

6. Langkah 6: Pentingnya Fitur (Feature Importance)

Analisis *feature importance* (Gini) dilakukan pada model final untuk mengidentifikasi fitur mana yang paling berkontribusi pada keputusan model.

Hasil: (Salin 3-5 fitur teratas dari output Anda di sini, misal:)

Top feature importance (Gini):

IPK_x_Study: 0.3541

IPK: 0.2879

Waktu_Belajar_Jam: 0.1450

...

-
- **Implikasi:** Fitur turunan **IPK_x_Study** (interaksi antara IPK dan Waktu Belajar) teridentifikasi sebagai prediktor paling penting, diikuti oleh **IPK**. Ini memberikan wawasan bahwa bukan hanya IPK atau waktu belajar saja, tetapi kombinasi keduanya adalah prediktor kelulusan yang paling kuat.

7. Langkah 7 & 8: Penyimpanan Model dan Cek Inference

1. **Penyimpanan:** Model final yang telah dianalisis disimpan sebagai **rf_model_p6.pkl**.
2. **Inference Check:** Dilakukan tes pemuatan model dan prediksi pada satu data fiktif (IPK 3.4).
 - **Hasil Prediksi:** [Isi hasil prediksi dari output Anda, misal: 1 (Lulus)]
 - Tes ini berhasil, mengkonfirmasi model dapat disimpan dan digunakan kembali dengan benar.