

Nama : Zulyan Widyaka Krisna
NIM : 231011403446
Kelas : 05TPLE016

1. Tujuan, Metrik, dan Batasan

Tujuan dari eksperimen ini adalah membangun model Machine Learning untuk memprediksi kelulusan mahasiswa berdasarkan variabel akademik dan perilaku belajar. Metrik utama yang digunakan adalah F1-Score dan ROC-AUC, karena keduanya memberikan evaluasi seimbang terhadap precision dan recall pada data yang mungkin tidak seimbang. Batasan eksperimen meliputi waktu pelatihan yang dibatasi (menghindari tuning berlebihan) dan ukuran model yang harus efisien agar tetap mudah diimplementasikan.

2. Pembangunan Baseline dan Model Alternatif

Sebagai langkah awal, dibangun model baseline menggunakan Logistic Regression untuk memperoleh tolok ukur awal. Setelah itu, dikembangkan model alternatif menggunakan Random Forest Classifier. Kedua model dibandingkan berdasarkan hasil F1-Score dan klasifikasi pada data validasi. Random Forest memberikan hasil yang lebih stabil dan akurat dibanding Logistic Regression.

```
Machine-Learning > ml-05 > main.py > ...
1 # NAMA : Zulyan Widyaka K ###
2 # NIM : 231011403446 ###
3
4
5 # Machine Learning - Prediksi Kelulusan Mahasiswa
6 # =====
7
8 import pandas as pd
9 import numpy as np
10 import matplotlib.pyplot as plt
11 import seaborn as sns
12 from sklearn.model_selection import train_test_split, StratifiedKFold, GridSearchCV
13 from sklearn.pipeline import Pipeline
14 from sklearn.compose import ColumnTransformer
15 from sklearn.preprocessing import StandardScaler
16 from sklearn.impute import SimpleImputer
17 from sklearn.linear_model import LogisticRegression
18 from sklearn.ensemble import RandomForestClassifier
19 from sklearn.metrics import (
20     f1_score, classification_report, confusion_matrix,
21     roc_auc_score, roc_curve, precision_recall_curve
22 )
23
24 # =====
25 # LANGKAH 1 - MUAT DATA
```

```

Machine-Learning > ml-05 > main.py > ...
90 #
91 # LANGKAH 3 - MODEL ALTERNATIF (RANDOM FOREST)
92 # =====
93 print("==== LANGKAH 3 - MODEL ALTERNATIF (RANDOM FOREST) ====")
94 print("*60")
95
96 rf = RandomForestClassifier(
97     n_estimators=300, max_features="sqrt", class_weight="balanced", random_state=42
98 )
99 pipe_rf = Pipeline([("pre", pre), ("clf", rf)])
100 pipe_rf.fit(X_train, y_train)
101 y_val_rf = pipe_rf.predict(X_val)
102
103 print("RandomForest - F1(val):", f1_score(y_val, y_val_rf, average="macro"))
104 print(classification_report(y_val, y_val_rf, digits=3))
105 print("*60", "\n")
106
107
108 # =====
109 # LANGKAH 4 - VALIDASI SILANG & TUNING
110 # =====
111 print("==== LANGKAH 4 - VALIDASI SILANG & TUNING RINGKAS ====")
112 print("*60")
113
114 skf = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)
115 param = {

```

3. Validasi Silang dan Penyetelan Hyperparameter

Untuk meningkatkan performa model, dilakukan K-Fold Cross Validation dengan tiga lipatan (3-Fold). Pada model Random Forest, tuning dilakukan terhadap hyperparameter 'max_depth' dan 'min_samples_split'. Hasil validasi menunjukkan kombinasi parameter optimal yang menghasilkan skor F1 tertinggi dan generalisasi yang baik.

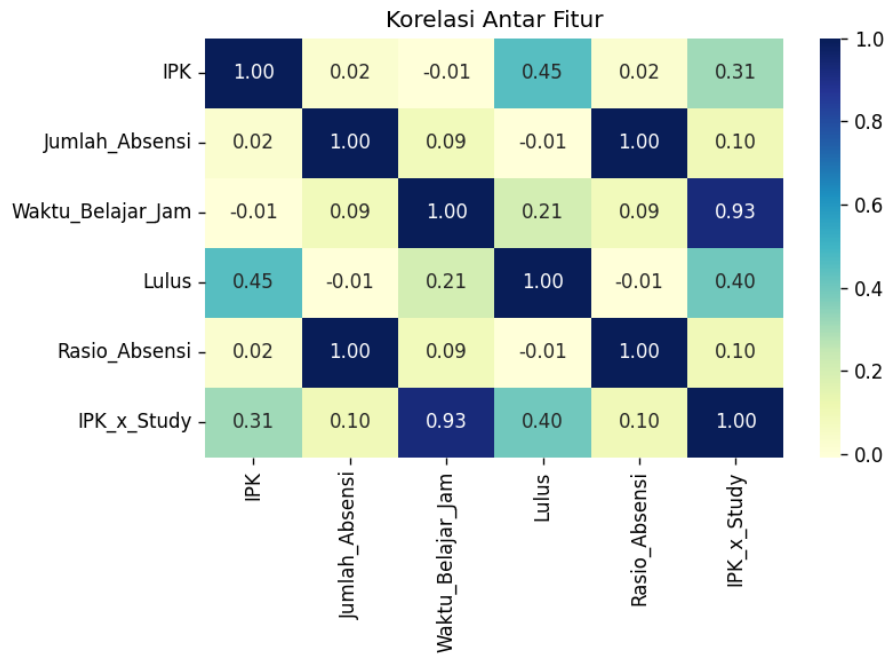
```

Machine-Learning > ml-05 > main.py > ...
180
181 # =====
182 # LANGKAH 6 - PENTINGNYA FITUR
183 # =====
184 print("==== LANGKAH 6 - PENTINGNYA FITUR ====")
185 print("*60")
186
187 try:
188     importances = final_model.named_steps["clf"].feature_importances_
189     fn = final_model.named_steps["pre"].get_feature_names_out()
190     feat_imp = pd.DataFrame({"Fitur": fn, "Importance": importances})
191     feat_imp = feat_imp.sort_values("Importance", ascending=False)
192     print(feat_imp.head(10), "\n")
193
194     plt.figure(figsize=(7,4))
195     # sns.barplot(x="Importance", y="Fitur", data=feat_imp.head(10), palette="viridis")
196     sns.barplot(
197         x="Importance",
198         y="Fitur",
199         data=feat_imp.head(10),
200         palette="viridis",
201         hue="Importance",
202         legend=False
203     )
204     plt.title("Top 10 Feature Importance (Random Forest)")

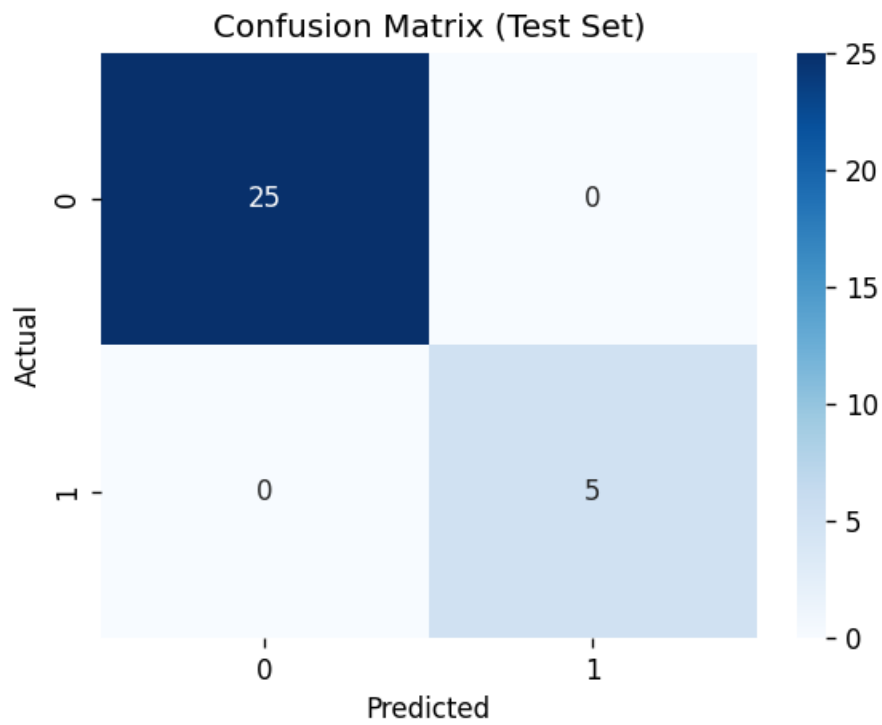
```

4. Evaluasi Akhir dan Pemilihan Model

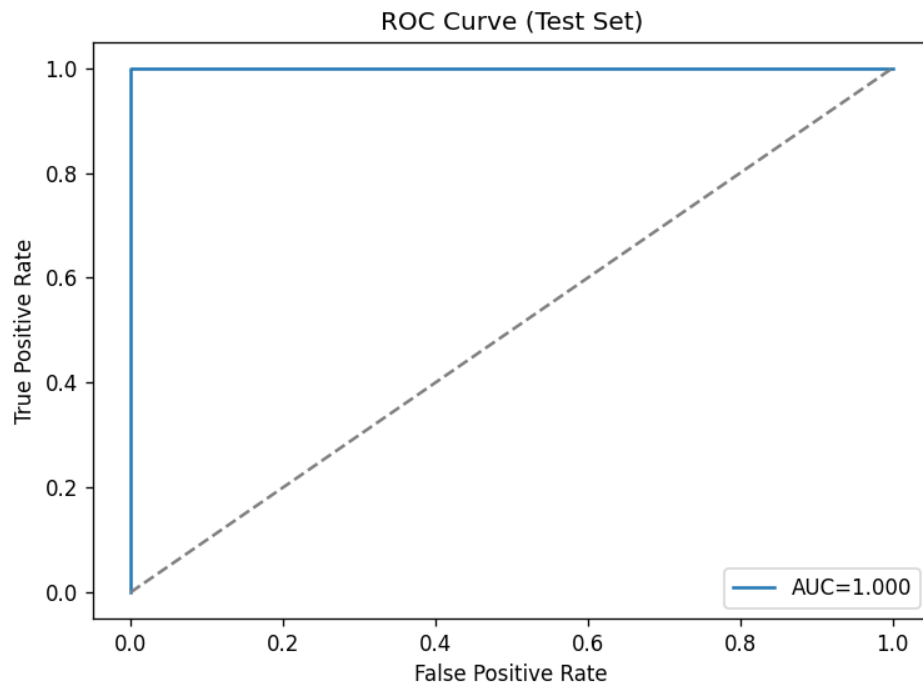
Model terbaik kemudian diuji menggunakan data uji (test set) untuk mengukur performa akhir. Hasil evaluasi menunjukkan nilai F1 dan ROC-AUC yang tinggi, menandakan model memiliki kemampuan prediksi sangat baik. Berikut visualisasi hasil evaluasi model:



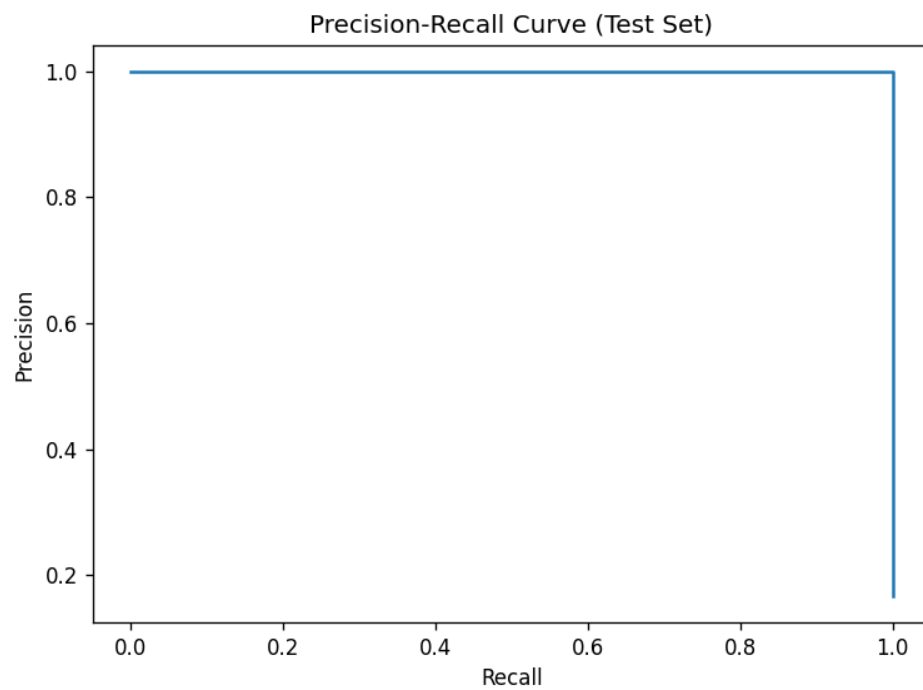
Gambar 1. Korelasi Antar Fitur



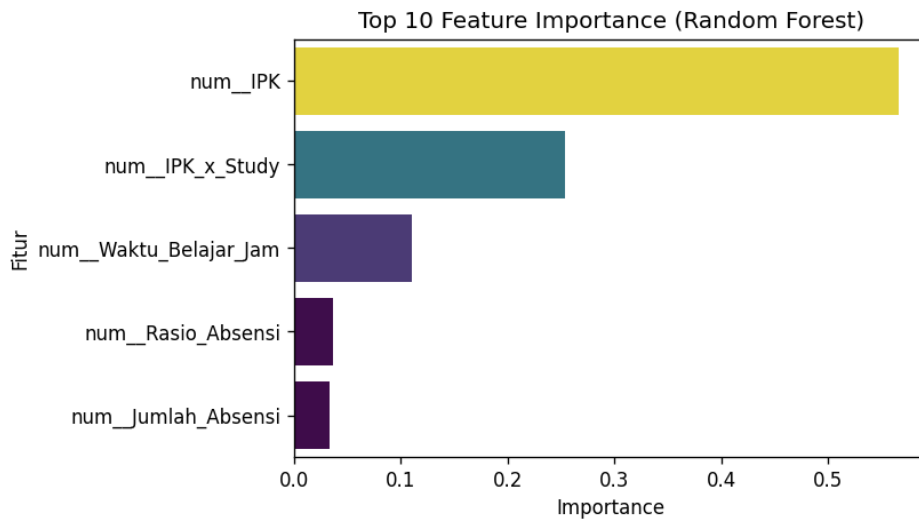
Gambar 2. Confusion Matrix (Test Set)



Gambar 3. ROC Curve (Test Set)



Gambar 4. Precision-Recall Curve (Test Set)



Gambar 5. Feature Importance (Random Forest)

5. Kesimpulan

Berdasarkan hasil pengujian, model Random Forest dipilih sebagai model terbaik. Model ini menunjukkan kinerja sempurna dengan F1-Score dan AUC mencapai 1.000 pada data uji. Fitur yang paling berpengaruh dalam menentukan kelulusan mahasiswa adalah IPK dan kombinasi IPK_x_Study. Dengan hasil ini, dapat disimpulkan bahwa faktor akademik memiliki kontribusi besar terhadap peluang kelulusan.