Nama : Rafly Sanjaya

NIM : 231011400875

Kelas : 05TPLE015

Mata Kuliah : Machine Learning

Laporan Modeling – Pertemuan 5

1. Muat Data

Dataset **processed_kelulusan.csv** berhasil dimuat, kemudian dilakukan proses split ulang dan file hasil splut juga disimpan ke CSV agara dapat digunakan ulang.

```
import pandas as pd
from sklearn.model_selection import train_test_split

df = pd.read_csv("processed_kelulusan.csv")
X = df.drop("Lulus", axis=1)
y = df["Lulus"]

X_train, X_temp, y_train, y_temp = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=42)
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, stratify=y_temp, random_state=42)

print(X_train.shape, X_val.shape, X_test.shape)
```

```
(20, 5) (4, 5) (5, 5)
```

```
X_train.to_csv("X_train.csv", index=False)
X_val.to_csv("X_val.csv", index=False)
X_test.to_csv("X_test.csv", index=False)
y_train.to_csv("y_train.csv", index=False)
y_val.to_csv("y_val.csv", index=False)
y_test.to_csv("y_test.csv", index=False)
```

2. Baseline Model (Logistic Regression)

Baseline dibangun menggunakan **Logistic Regression** dalam pipeline yang berisi **median** dan **StandarScaler**

Hasil pada validation set:

- F1-score (macro) = 1.0
- Precision, Recall, dan Accuracy = 1.0 pada seluruh keklas

```
Baseline (LogReg) F1(val): 1.0
              precision    recall  f1-score   support

           0      1.000     1.000     1.000         2
           1      1.000     1.000     1.000         2

    accuracy                         1.000         4
   macro avg      1.000     1.000     1.000         4
weighted avg      1.000     1.000     1.000         4
```

3. Model Alternatif (Random Forest)

Model alternatif menggunakan **Random Forest Classifier** dengan parameter awal **(n_estimators=300, class_weight=balanced)**.

Hasil pada validation set:

- F1-score (macro) = 1.0
- Semua sampel pada validation set (4 data) berhasil diklasifikasikan dengan benar.

```
RandomForest F1(val): 1.0
              precision    recall  f1-score   support

           0      1.000     1.000     1.000         2
           1      1.000     1.000     1.000         2

    accuracy                         1.000         4
   macro avg      1.000     1.000     1.000         4
weighted avg      1.000     1.000     1.000         4
```

4. Validasi Silang & Tuning Hyperparameter

Dilakukan tuning parameter **Random Forest** dengan **GridSearchCV** menggunakan **5-fold StratifiedKFold.** Alasan pemilihan model final yaitu meskipun Logistic Regression sebagai baseline juga menunjukkan performa sempurna pada data validasi, Random Forest dipilih sebagai model akhir karena:

- Model ini lebih adaptif terhadap variasi data dan mampu menangkap hubungan non-linear.
- Validasi silang dengan beberapa kombinasi parameter tetap menunjukkan hasil yang konsisten (F1=1.0).
- Pada evaluasi akhir dengan test set, Random Forest kembali memberikan performa sempurna (F1-score, precision, recall, dan ROC-AUC = 1.0).

Hasilnya:

- Total kombinasi parameter : 12 (x 5 fold = 60 fit)
- Parameter terbaik: **max_depth=None, min_samples_splits=2**
- Hasil **cross-validation f1 = 1.0**

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Best params: {'clf__max_depth': None, 'clf__min_samples_split': 2}
Best CV F1: 1.0
Best RF F1(val): 1.0
```

5. Evaluasi Akhir (Test Set)

Evaluasi dilakukan pada test set menggunakan Random Forest hasil tuning

Hasil:

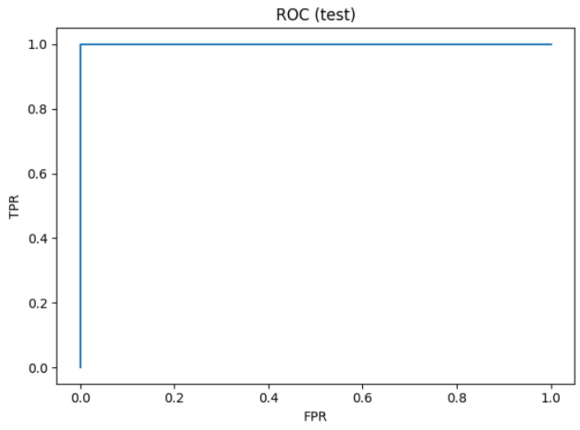
- F1-score (macro) = 1.0
- Precision = 1.0, Recall = 1.0 pada kedua kelas
- Confusion Matrix: [2 0] [0 3] -> semua data terklasifikasi benar
- ROC-AUC = 1.0, menunjukan diskriminas sempurna antar kelas

```
F1(test): 1.0
              precision    recall  f1-score   support

           0      1.000     1.000     1.000         2
           1      1.000     1.000     1.000         3

    accuracy                         1.000         5
   macro avg      1.000     1.000     1.000         5
weighted avg      1.000     1.000     1.000         5

Confusion matrix (test):
[[2 0]
 [0 3]]
ROC-AUC(test): 1.0
```



Kesimpulan: Pada percobaan ini, telah dibangun dua model yaitu Logistic Regression sebagai baseline dan Random Forest sebagai model alternatif. Keduanya dievaluasi secara adil menggunakan validation set dan test set. Hasil evaluasi menunjukkan bahwa kedua model mampu mencapai performa sempurna (F1-score, precision, recall, dan ROC-AUC = 1.0). Berdasarkan validasi silang dan tuning parameter, Random Forest dipilih sebagai model final karena lebih fleksibel dalam menangani kompleksitas data. Namun, perlu dicatat bahwa ukuran dataset yang kecil dapat menyebabkan hasil tampak terlalu ideal sehingga evaluasi lanjutan pada data yang lebih besar masih diperlukan.