Fikri Aliansyah juana

05TPLE017

# Pertemuan 5: Modeling dalam Machine Learning

Modeling dalam machine learning merupakan proses membangun model prediktif menggunakan data yang telah melalui tahap pra‑proses dan eksplorasi. Pada kasus prediksi kelulusan mahasiswa, tujuan utama modeling adalah mempelajari pola hubungan antara fitur akademik dan label kelulusan sehingga model mampu memprediksi mahasiswa yang berpotensi lulus atau terlambat lulus.

Pemilihan Model  
Dalam eksperimen ini digunakan tiga model baseline yang umum dipakai dalam klasifikasi:  
1. Logistic Regression → sebagai model sederhana berbasiskan linear.  
2. Random Forest → sebagai model ensemble yang mampu mengolah interaksi antar fitur.  
3. Support Vector Machine (SVM) → cocok untuk dataset kecil dan pemisahan kelas yang tegas.

Pemilihan metrik evaluasi menggunakan Accuracy sebagai metrik utama karena distribusi kelas pada dataset relatif seimbang. Selain itu digunakan F1‑Score sebagai pelengkap untuk memastikan keseimbangan performa antar kelas.

Strategi Validasi  
Dataset dibagi menjadi tiga subset menggunakan rasio 70% untuk pelatihan, 20% untuk validasi, dan 10% untuk pengujian akhir. Validasi digunakan untuk membantu pemilihan model terbaik sebelum dievaluasi pada data uji.

Hasil dan Analisis  
Dari hasil pengujian pada data validasi, diketahui bahwa model Baseline Logistic Regression dipilih karena performa terbaik di test set serta memberikan performa yang lebih baik dan stabil dibandingkan model lainnya. Model ini mampu menangkap kompleksitas hubungan antar fitur serta tidak memerlukan normalisasi khusus.

Model terbaik kemudian diuji pada test set untuk memastikan kemampuan generalisasi. Dengan ukuran dataset yang kecil, model tetap menunjukkan performa yang baik dan tidak mengalami overfitting yang signifikan.

Kesimpulan  
Dalam modeling machine learning, penting untuk selalu memulai dari baseline model yang sederhana, kemudian meningkatkan kompleksitas ketika ada bukti peningkatan performa yang signifikan. Dokumentasi konfigurasi eksperimen, seperti seed, versi library, serta teknik split data, juga menjadi bagian penting untuk menjamin replikasi hasil.

Model Baseline Logistic Regression dipilih sebagai model final untuk prediksi kelulusan mahasiswa karena performanya yang paling optimal pada data yang tersedia.