Chapter 2: End-to-End Machine Learning Project

Bab ini memberikan panduan langkah demi langkah untuk membangun sebuah proyek machine learning lengkap dari awal hingga akhir, mulai dari pengambilan data hingga evaluasi model. Fokusnya bukan pada teori algoritma, tapi pada workflow praktis.

A. Proyek yang Dikerjakan

- Proyek: Prediksi Harga Rumah di California
  Tujuan: Membangun model machine learning untuk memprediksi harga rumah berdasarkan fitur-fitur seperti:

  -Lokasi

  -Kepadatan penduduk

  -Jumlah kamar

  -Pendapatan rata-rata

  -Dataset: California Housing dataset dari 1990-an, tersedia melalui sklearn.datasets.fetch_california_housing().

B. Langkah-Langkah End-to-End Machine Learning Project

1. Definisikan Masalah

Tujuan bisnis → prediksi harga rumah (regresi)

Supervised learning → regresi, karena output berupa angka

2. Dapatkan Data

Data diunduh dari Scikit-Learn

Disimpan secara lokal jika ingin digunakan berulang

3. Jelajahi dan Visualisasikan Data

Menggunakan Pandas, Matplotlib, Seaborn

Contoh analisis: histogram distribusi fitur, korelasi antar variabel

Lihat outlier, missing value, dsb

4. Buat Set Validasi

Data dibagi menjadi training set dan test set

Teknik:

Random split (dengan train_test_split)

Stratified sampling → memastikan proporsi data tetap (misal berdasar kategori pendapatan)

5. Membersihkan Data

Menangani nilai kosong (SimpleImputer)

Konversi data kategorikal ke numerik (OneHotEncoder)

Scaling fitur numerik (StandardScaler)

➞ Disatukan menggunakan Scikit-Learn Pipelines untuk proses preprocessing otomatis dan konsisten

6. Pilih dan Latih Model

Contoh model awal:

Linear Regression

Decision Tree Regressor

Random Forest Regressor

Evaluasi model dengan RMSE (Root Mean Squared Error)

7. Fine-Tune Model

Gunakan Grid Search atau Randomized Search untuk mencari hyperparameter terbaik

Validasi kinerja dengan cross-validation (misalnya cross_val_score)

8. Evaluasi pada Test Set

Setelah model terbaik dipilih → evaluasi di test set final

Pastikan tidak overfitting

9. Simpan dan Deploy Model

Simpan model dengan joblib atau pickle

Siapkan model untuk di-deploy dalam aplikasi (misalnya API)

C. Fitur Tambahan

Menambahkan fitur baru yang berguna, misalnya rooms_per_household

Menggunakan korelasi dan eksperimen untuk menentukan fitur-fitur yang berkontribusi besar

Visualisasi error distribusi untuk memahami model lebih dalam

D. Insight Kunci dari Bab Ini

Workflow machine learning lebih kompleks daripada sekadar melatih model.

Proyek nyata memerlukan banyak waktu untuk eksplorasi dan pembersihan data.

Pipeline dan modularitas sangat penting untuk menjaga kode tetap bersih dan konsisten.

Evaluasi model bukan hanya dari metrik angka, tapi juga dari pemahaman distribusi error.