Chapter 2 : Summary

Diedrick Darrell Darmadi - 1103223031

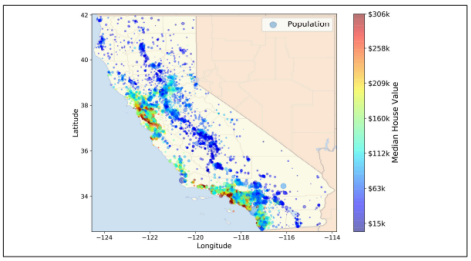

Figure 2-1. California housing prices

Bab 2 menjelaskan bagaimana membangun sebuah proyek machine learning dari awal hingga akhir menggunakan contoh dataset California Housing. Proses dimulai dengan mendefinisikan tujuan bisnis secara jelas, yaitu memprediksi harga rumah berdasarkan berbagai fitur seperti lokasi, jumlah kamar, dan kepadatan penduduk. Setelah tujuan ditentukan, langkah berikutnya adalah memahami data dengan melakukan eksplorasi awal, menampilkan beberapa contoh data mentah, dan memeriksa struktur dataset. Pada tahap ini biasanya dilakukan visualisasi distribusi nilai serta identifikasi potensi masalah seperti missing values atau nilai ekstrem.

Setelah itu dataset diunduh, dibaca, dan dilakukan pemeriksaan struktur awal, termasuk jumlah baris, tipe fitur, serta lima baris pertama untuk melihat bentuk data mentah.

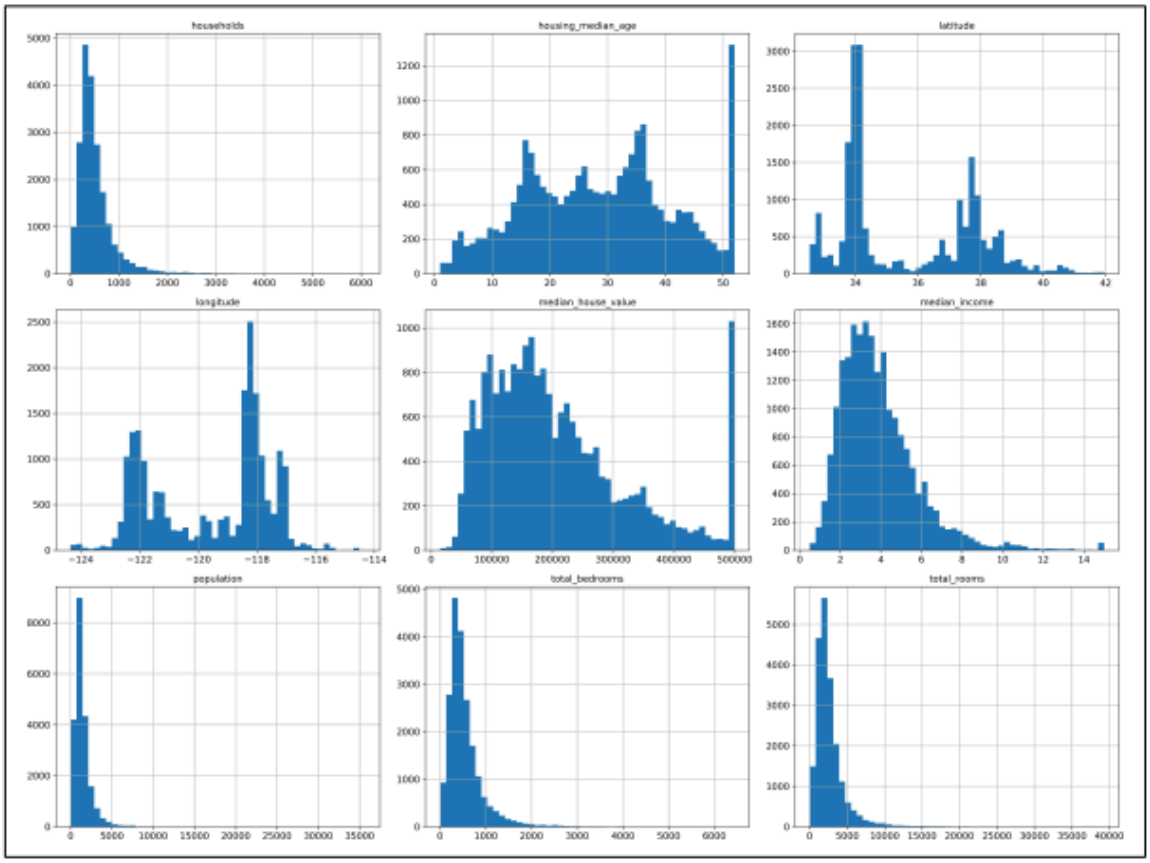

Figure 2-8. A histogram for each numerical attribute

Setelah pembagian data, dilakukan eksplorasi data (EDA) untuk memahami pola dan struktur dataset. Scatter plot berdasarkan koordinat geografis digunakan untuk memvisualisasikan distribusi harga rumah dan mengidentifikasi pola spasial. Korelasi antar fitur dihitung untuk mengetahui fitur mana yang paling berhubungan dengan harga, dan beberapa kombinasi fitur baru dibuat untuk meningkatkan kualitas informasi.

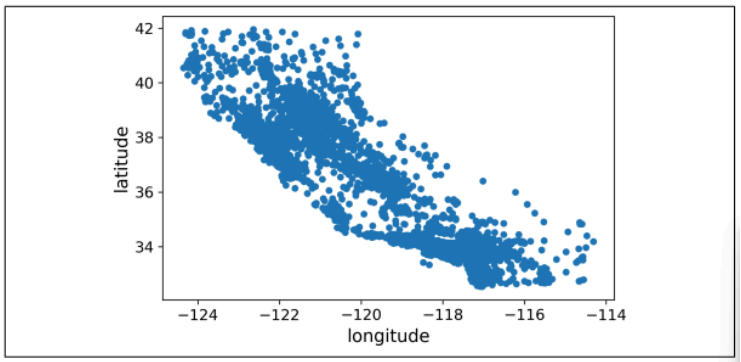

Figure 2-11. A geographical scatterplot of the data

Pada tahap preprocessing, data dibersihkan dari missing values, fitur numerik dinalarisasi menggunakan imputer dan scaling, dan fitur kategorikal diencoding. Seluruh langkah preprocessing digabungkan ke dalam sebuah *pipeline* agar prosesnya konsisten, terstruktur, dan dapat direproduksi dengan mudah.

Model awal kemudian dilatih menggunakan beberapa algoritma seperti Linear Regression, Decision Tree, dan Random Forest. Masing-masing model diuji pada data training menggunakan *cross-validation* untuk melihat performa dan mendeteksi overfitting. Model seperti Decision Tree menunjukkan overfitting berat, sementara Random Forest memberikan performa yang lebih stabil dan akurat

Tahap berikutnya adalah *hyperparameter tuning* menggunakan Grid Search dan Randomized Search untuk menemukan konfigurasi terbaik bagi model. Proses ini mencoba banyak kombinasi parameter dan mengevaluasi performanya secara sistematis. Model terbaik kemudian dievaluasi terhadap test set untuk mendapatkan estimasi performa dunia nyata.

Bab ini ditutup dengan proses menyimpan model beserta pipeline preprocessing menggunakan `joblib` agar dapat digunakan kembali pada data baru tanpa mengulang pelatihan. Dengan demikian, seluruh siklus proyek machine learning terselesaikan dari awal hingga akhir.