Chapter 2 – Rahmanda Afebrio Yuris Soesatyo

End-to-End Machine Learning Project

1. Perumusan Masalah dan Pemahaman Data

Proses dimulai dengan menentukan tujuan bisnis secara jelas.

- Tujuannya adalah memprediksi harga rumah berdasarkan fitur-fitur seperti lokasi, jumlah kamar, dan kepadatan penduduk.
- Selanjutnya, dilakukan eksplorasi awal data untuk memahami isinya.
- Langkah ini mencakup mengunduh dan membaca data , memeriksa struktur dataset (jumlah baris, tipe fitur) , dan melihat beberapa contoh data mentah.
- Visualisasi dilakukan untuk melihat distribusi nilai (contohnya menggunakan histogram untuk setiap atribut numerik) dan mengidentifikasi masalah seperti *missing values* atau nilai ekstrem.

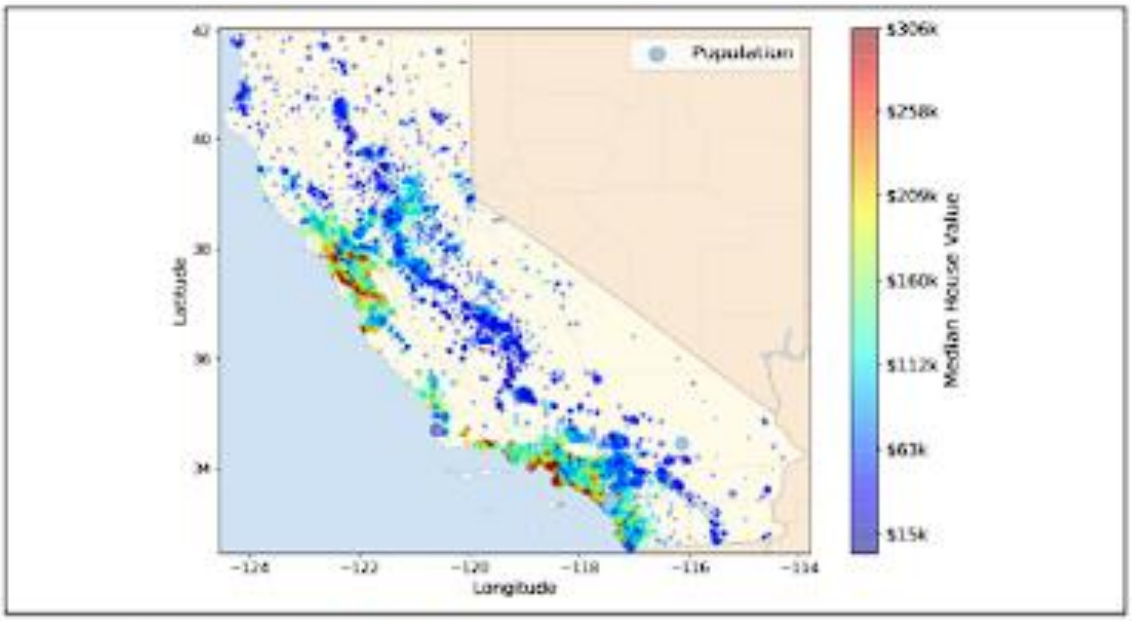

Figure 2-1. California housing prices

2. Eksplorasi Data (EDA) dan *Feature Engineering*

Setelah membagi data (menjadi *training set* dan *test set*), eksplorasi data yang lebih mendalam dilakukan.

- Visualisasi Spasial: *Scatter plot* berdasarkan koordinat geografis (lintang dan bujur) digunakan untuk memvisualisasikan distribusi harga rumah dan menemukan pola spasial. * Korelasi: Korelasi antar fitur dihitung untuk mengetahui atribut mana yang paling berhubungan dengan harga rumah.
- *Feature Engineering*: Kombinasi fitur baru dibuat untuk meningkatkan kualitas informasi yang akan digunakan model.

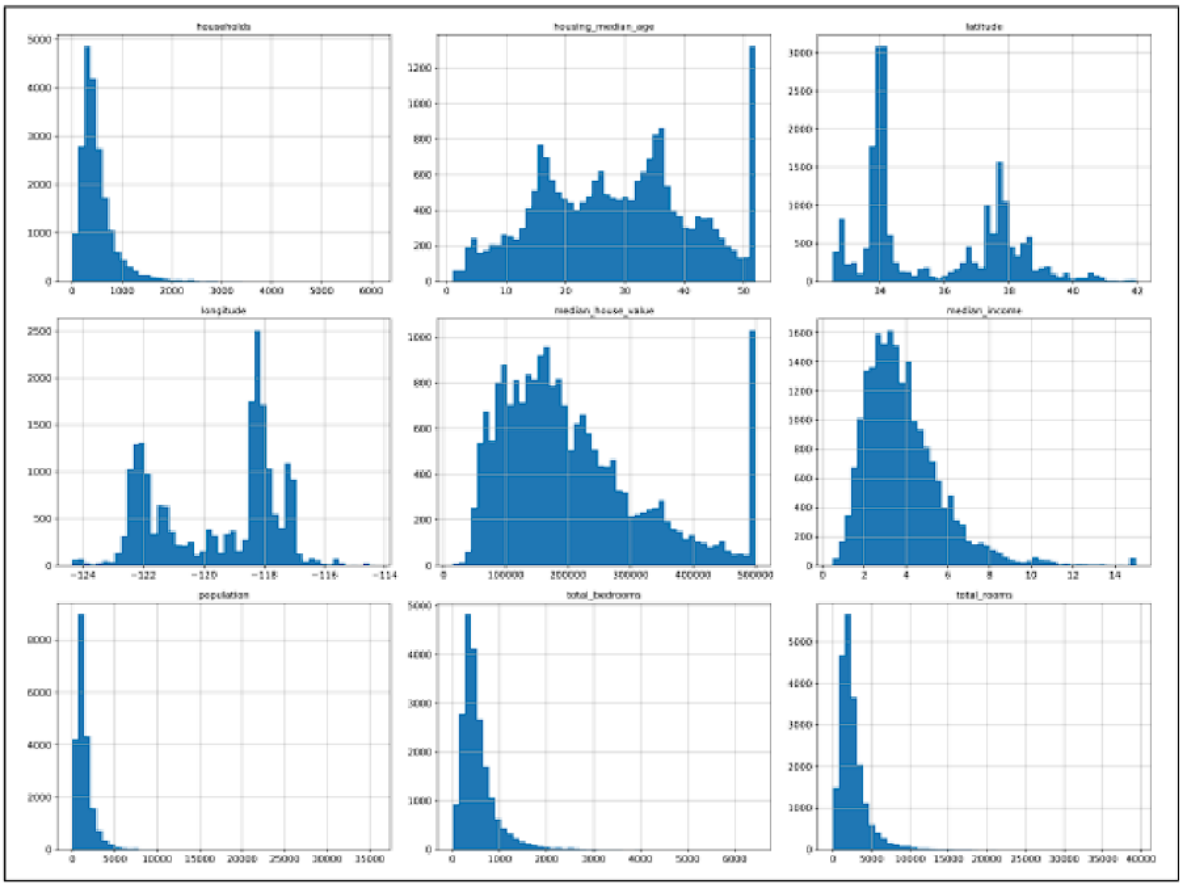

Figure 2-8. A histogram for each numerical attribute

3. *Preprocessing* Data

Tahap ini berfokus pada pembersihan dan transformasi data agar siap digunakan oleh model.

- Pembersihan Data: Menangani missing values.
- Transformasi Numerik: Fitur numerik dinormalisasi menggunakan teknik seperti *imputer* dan *scaling*.
- Transformasi Kategorikal: Fitur kategorikal di-*encoding*.
- *Pipeline*: Semua langkah *preprocessing* ini digabungkan ke dalam sebuah pipeline untuk memastikan prosesnya konsisten, terstruktur, dan dapat direproduksi dengan mudah.

4. Pelatihan, Evaluasi, dan *Tuning* Model

Beberapa model awal dilatih untuk perbandingan kinerja.

- Model Awal: Melatih algoritma seperti Linear Regression, Decision Tree, dan Random Forest.
- Evaluasi: Setiap model diuji pada data *training* menggunakan *cross-validation* untuk mengukur performa dan mendeteksi *overfitting*.

  Contohnya, Decision Tree menunjukkan *overfitting* yang berat, sedangkan Random Forest memberikan performa yang lebih stabil dan akurat.

- *Hyperparameter Tuning*: Menggunakan Grid Search dan Randomized Search untuk mencoba banyak kombinasi parameter dan menemukan konfigurasi terbaik bagi model yang sudah dipilih

5. Evaluasi Akhir dan Peluncuran

- Evaluasi Akhir: Model terbaik dievaluasi terhadap *test set* untuk mendapatkan estimasi yang realistis tentang performa di dunia nyata.
- Penyimpanan Model: Model terbaik, beserta *pipeline preprocessing*-nya, disimpan menggunakan joblib agar dapat digunakan kembali untuk memprediksi data baru tanpa perlu mengulang pelatihan.

Dengan langkah-langkah ini, seluruh siklus proyek *machine learning* terselesaikan dari awal hingga akhir