DOKUMEN PROYEK

12S3205 - PENAMBANGAN DATA

Development of a Predictive Regression Model for House Prices Using Ensemble Stacking Techniques

Disusun Oleh:

12S22015	**Angelina Nadeak**
12S22029	**Jeremy Samosir**
12S22038	**Ade Siahaan**
12S22052	**Rosari Simanjuntak**

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO (FITE)
INSTITUT TEKNOLOGI DEL
TAHUN 2024/2025

DAFTAR ISI

BAB 1 BUSINESS UNDERSTANDING

1.1 Determine Business Objective

Industri properti merupakan salah satu sektor yang sangat dinamis dan memiliki dampak signifikan terhadap perekonomian suatu negara. Harga rumah menjadi indikator utama dalam transaksi jual-beli properti, baik untuk konsumen perorangan, agen properti, maupun perusahaan pengembang real estate. Namun, penentuan harga rumah seringkali bersifat subjektif dan sangat dipengaruhi oleh faktor-faktor eksternal yang sulit diprediksi seperti kondisi pasar, lokasi, dan tren ekonomi.

Tujuan bisnis dari proyek ini adalah menyediakan sistem prediksi harga rumah berbasis machine learning yang mampu mengurangi ketidakpastian dalam proses estimasi harga. Dengan sistem prediksi ini, diharapkan stakeholder properti seperti investor, penjual, pembeli, dan agen real estate dapat mengambil keputusan yang lebih cepat dan tepat.

Manfaat yang ingin dicapai dalam jangka panjang:

- Meminimalisir kesalahan estimasi harga.
- Memberikan insight berbasis data dalam proses negosiasi properti.
- Meningkatkan efisiensi waktu dan biaya dalam menentukan harga jual/beli.
- Meningkatkan daya saing perusahaan properti melalui adopsi teknologi AI.

1.2 Determine Project Goal

Tujuan teknis dari proyek ini adalah mengembangkan model prediksi harga rumah dengan pendekatan ensemble stacking, yang menggabungkan kekuatan dari beberapa algoritma machine learning seperti XGBoost, LightGBM, dan CatBoost. Model stacking ini bertujuan memaksimalkan akurasi prediksi dan meminimalkan error, terutama dalam kondisi data yang kompleks dan beragam.

Target pengembangan model:

- Memprediksi harga rumah berdasarkan fitur properti dengan akurasi tinggi.
- Mengurangi bias prediksi yang disebabkan oleh model tunggal.
- Menghasilkan model yang stabil dan generalisasi dengan baik terhadap data baru.

1.3 Produce Project Plan

Rencana pelaksanaan proyek:

- Tahap 1: Pengumpulan dan eksplorasi data properti.
- Tahap 2: Preprocessing, feature selection dan feature engineering.
- Tahap 3: Pengembangan model ensemble stacking.
- Tahap 4: Evaluasi model menggunakan beberapa metrik evaluasi, di antaranya:
    - RMSE (Root Mean Squared Error) untuk menghitung akar rata-rata kesalahan kuadrat prediksi.
    - MAE (Mean Absolute Error) untuk mengukur rata-rata selisih absolut antara nilai prediksi dan nilai aktual.
    - R^2 (Coefficient of Determination) untuk menilai seberapa besar variasi target yang dapat dijelaskan oleh model.
    - MAPE (Mean Absolute Percentage Error) untuk mengetahui rata-rata kesalahan dalam bentuk persentase.
- Tahap 5: Deployment model dalam bentuk aplikasi prediktif.

Waktu estimasi pengerjaan: 2 bulan Tim pelaksana: Data Scientist, Data Engineer, Business Analyst.

BAB 2 DATA UNDERSTANDING

2.1 Collecting Data

Dataset yang digunakan pada proyek ini diambil dari kompetisi "House Prices - Advanced Regression Techniques" di platform Kaggle. Dataset ini sangat kaya dan telah menjadi benchmark umum dalam pengembangan model regresi prediksi harga properti.

Jumlah data:

- Training set: 1.460 baris data.
- Testing set: 1.459 baris data.
- Total fitur: 80 fitur.

Data yang dikumpulkan mencakup berbagai aspek properti, antara lain:

- Karakteristik fisik rumah (luas bangunan, jumlah kamar tidur, jumlah kamar mandi, tahun pembangunan).
- Kualitas properti (rating konstruksi, kondisi bangunan).
- Lokasi properti (nama lingkungan, jarak ke fasilitas umum).
- Fitur eksternal (kondisi halaman, jenis garasi, tipe atap).

2.2 Describe Data

Setelah pengumpulan data, proses berikutnya adalah memahami distribusi dan karakteristik data. Dataset ini mencakup:

- Fitur numerik: LotArea, GrLivArea, YearBuilt, TotalBsmtSF, GarageArea.
- Fitur kategorikal: Neighborhood, HouseStyle, ExterQual, GarageType.

Sebagian besar variabel numerik memiliki distribusi yang skewed, terutama variabel SalePrice sebagai label target, yang menunjukkan kecenderungan outlier pada rumah-rumah mewah. Oleh karena itu, analisis statistik deskriptif seperti mean, median, standar deviasi, serta visualisasi boxplot dan histogram dilakukan untuk memahami sebaran data.

2.3 Validation Data

Agar model dapat dievaluasi secara objektif, data dibagi menjadi dua bagian:

- Training set (80%): Digunakan untuk membangun dan melatih model.
- Testing set (20%): Digunakan untuk menguji generalisasi model terhadap data baru.

Selain itu, digunakan teknik:

- Cross-validation (5-Fold): Memecah data training menjadi lima bagian untuk memastikan model tidak overfitting.
- Holdout validation: Menguji performa akhir pada data testing yang belum pernah dilihat model sebelumnya.

Proses validasi ini sangat penting untuk memastikan bahwa model yang dikembangkan tidak hanya menghafal data training, tetapi mampu memprediksi dengan baik untuk data real-world yang sebelumnya tidak dikenal.

BAB 3 DATA PREPARATION

3.1 Data Selection

3.2 Data Cleaning

3.3 Data Construct

3.4 Labeling Data

3.5 Data Integration

BAB 4 MODELLING

4.1 Build Model

BAB 5 EVALUATION

BAB 6 DEPLOYMENT

DAFTAR PUSTAKA