Laporan Tugas Besar Data Mining

Penerapan Algoritma Random Forest dan K-Means Clustering dalam Pembangunan Sistem Valuasi Aset dan Pemetaan Struktur Pasar Properti Residensial Sekunder di Wilayah Surabaya

Sistem Informasi
Telkom University
Surabaya

Disusun Oleh:

Alvyn Wira Pratama	**1204230012**
Kifansa Naufal Fadhlurrohman	**1204230030**
Elvina Angelina Kurniawan	**1204230067**
Revania Kyla Evadne	**1204230101**

PROGRAM STUDI S1 SISTEM INFORMASI

FAKULTAS REKAYASA INDUSTRI

TELKOM UNIVERSITY SURABAYA

2026

DAFTAR ISI

BAB I
BUSINESS UNDERSTANDING

1.1. Latar Belakang Masalah

Sebagai salah satu pusat perekonomian yang strategis, pasar properti di Surabaya saat ini menunjukkan aktivitas yang sangat padat dan fluktuatif. Setiap harinya, ribuan data properti baru masuk ke pasar dengan spesifikasi yang sangat beragam, mulai dari rumah sederhana hingga aset mewah. Meski begitu, banyaknya volume data ini belum diimbangi dengan metode penilaian yang objektif dan terstandarisasi. Banyak agen maupun pemilik aset yang masih sangat bergantung pada intuisi atau metode konvensional untuk menentukan harga jual. Hal ini sering memicu bias penilaian yang menyebabkan kerugian signifikan. Kondisi ini menciptakan ketidakkonsistenan pasar yang serius di mana banyak aset ditawarkan dengan harga yang tidak rasional atau masuk akal, baik terlalu tinggi sehingga sulit terserap pasar maupun terlalu rendah yang justru merugikan nilai investasi pemiliknya.

Akibat kejadian tersebut, pergeseran menuju pendekatan berbasis data atau data-driven decision making melalui Data Mining menjadi salah satu langkah yang penting. Penerapan algoritma Supervised Learning dapat memberikan solusi konkret berupa model prediksi harga yang presisi yang membuat ketergantungan pada perhitungan harga manual yang subjektif dapat diminimalisir. Selain itu, teknik Unsupervised Learning memberikan kemampuan untuk memetakan tipe unik setiap properti dan mengelompokkannya ke dalam segmen pasar yang jelas secara otomatis. Integrasi dari kedua metode ini dapat menciptakan ekosistem pasar yang lebih transparan serta memberikan dasar pengambilan keputusan yang lebih strategis dan terukur bagi seluruh pelaku industri properti.

1.2. Tujuan Penelitian

Berdasarkan permasalahan ketidakkonsistenan pasar yang telah diuraikan sebelumnya, penelitian ini dilakukan dengan tujuan sebagai berikut:

1) Mengembangkan model yang dapat memberikan estimasi harga yang presisi menggunakan pendekatan *Supervised Learning* (Random Forest) untuk memberikan standar penilaian yang objektif dan meminimalisir bias subjektivitas dalam penentuan nilai aset.
2) Mengembangkan model yang dapat melakukan segmentasi pasar secara otomatis melalui teknik Unsupervised Learning (K-Means Clustering) untuk mengelompokkan

profil properti berdasarkan kemiripan karakteristik fisik, sehingga mempermudah identifikasi kategori aset tanpa proses pelabelan manual.

1.3. Metrik Keberhasilan

Untuk memastikan sistem yang dibuat dapat menjawab permasalahan bisnis secara efektif, penelitian ini menetapkan indikator keberhasilan teknis sebagai berikut:

A. Model prediksi harga diterima jika mampu meminimalkan selisih antara harga prediksi dan harga aktual. Metrik utama yang digunakan adalah Root Mean Square Error (RMSE), di mana nilai error yang semakin kecil mengindikasikan kemampuan model yang semakin presisi dalam menaksir harga pasar yang wajar.
B. Model segmentasi aset diukur menggunakan Silhouette Score untuk menilai seberapa rapi pemisahan antar-klaster. Secara kualitatif, model dianggap berhasil jika klaster yang terbentuk dapat dijelaskan secara logis ke dalam profil bisnis yang nyata, seperti segmen rumah sederhana, menengah, atau mewah.

BAB II
DATA UNDERSTANDING

2.1. Sumber Data

Penelitian ini menggunakan dataset berjudul Combined_Datalist_v1.1.csv dari kaggle yang memiliki historis penawaran properti di wilayah Surabaya. Dataset ini memiliki volume data sebanyak kurang lebih 8.980 baris data yang mencakup berbagai distrik strategis. Secara umum, data ini merepresentasikan kondisi pasar sekunder (*secondary market*), di mana variasi aset sangat tinggi mulai dari rumah tua yang membutuhkan renovasi hingga hunian baru yang siap huni. Format penyimpanan data menggunakan standar CSV (*Comma Separated Values*) dengan pemisah titik koma, yang memungkinkan pengolahan data lintas platform namun tetap memerlukan penanganan khusus pada tahap pra-pemrosesan karena sifat data mentahnya yang belum terstruktur dengan sempurna.

2.2. Deskripsi Variabel

Untuk membangun model valuasi yang tidak hanya akurat secara matematis tetapi juga relevan secara bisnis, penelitian ini membedah karakteristik properti ke dalam dua kelompok besar, yaitu variabel target dan variabel fitur. Total terdapat 12 atribut yang digunakan dalam analisis ini.

A. Variabel Target (Dependent Variable)

Fokus utama prediksi adalah Price (Harga). Variabel ini berisi nilai penawaran aset dalam mata uang Rupiah. Dalam konteks pemodelan, variabel ini bertindak sebagai label jawaban yang akan dipelajari polanya oleh algoritma Machine Learning.

B. Variabel Fitur (Independent Variables)

Tabel fitur dibagi menjadi kategori atribut fisik (kuantitatif) dan atribut kualitatif untuk menangkap nilai intrinsik aset secara utuh:

1) Atribut Fisik & Dimensi (Numerik):

Atribut	Deskripsi
Luas Tanah & Luas Bangunan (m^2)	Indikator ini merupakan indikator yang paling dominan dalam penentuan harga. Rasio antara luas tanah dan bangunan juga sering menjadi penentu apakah properti tersebut memiliki halaman sisa atau dimaksimalkan untuk ruang huni.

Kamar Tidur dan Kamar Mandi	Menunjukkan kapasitas fungsional hunian. Jumlah kamar yang lebih banyak biasanya berkorelasi positif dengan harga, namun harus sebanding dengan luas bangunan.
Jumlah Lantai	Mengindikasikan volume vertikal bangunan. Rumah bertingkat memiliki biaya konstruksi per meter persegi yang berbeda dengan rumah satu lantai.
Daya Listrik	Menggambarkan kapasitas utilitas energi yang terpasang, yang sering kali berbanding lurus dengan segmen kemewahan properti.

2) Atribut Kualitas & Legalitas (Kategorikal):

Atribut	Deskripsi
Kecamatan	Representasi lokasi geografis yang menentukan Nilai Jual Objek Pajak (NJOP) dan harga pasar tanah di area tersebut.
Sertifikat	Status legalitas lahan, seperti Sertifikat Hak Milik (SHM) atau Hak Guna Bangunan (HGB). Aset dengan SHM umumnya memiliki valuasi pasar yang lebih tinggi dan likuid.
Posisi Hook	Menjelaskan apakah properti berada di sudut jalan (memiliki dua muka). Posisi hook biasanya dihargai lebih tinggi (premium) karena sirkulasi udara dan pencahayaan yang lebih baik.
Kondisi Perabotan	Tingkat kelengkapan interior, mulai dari Unfurnished (kosong), Semi-furnished, hingga Full furnished
Kondisi Properti	Penilaian subjektif terhadap fisik bangunan (misal: "Baru", "Bagus", atau "Butuh Renovasi").
Keterjangkauan Internet	Indikator modernitas infrastruktur di lingkungan sekitar, yang kini menjadi faktor pertimbangan penting bagi pembeli urban.

2.3. Temuan Eksplorasi Awal (*Exploratory Data Analysis*)

Pada tahap pemeriksaan awal, ditemukan beberapa masalah kualitas data yang perlu diselesaikan sebelum proses pemodelan dimulai. Masalah pertama adalah

ketidakkonsistenan format data. Variabel harga masih tersimpan sebagai teks karena mengandung tanda pemisah ribuan, sehingga sistem tidak dapat membacanya sebagai angka. Selain itu, terdapat perbedaan penulisan pada kategori yang sebenarnya sama, seperti “SHM”, “Sertifikat Hak Milik”, dan “shm”, yang oleh sistem dianggap sebagai data berbeda.

Masalah kedua adalah data yang hilang (missing values) pada beberapa atribut pendukung, terutama Daya Listrik dan Jumlah Lantai. Kondisi ini diduga terjadi karena kelalaian saat pengisian data. Menghapus seluruh baris yang kosong membuat mengurangi jumlah data signifikan, sehingga dilakukan metode imputasi, seperti mengisi dengan nilai median atau modus, agar karakteristik data tetap terjaga.

Masalah terakhir adalah adanya outlier atau data pencilan. Ditemukan nilai harga yang tidak wajar, seperti nol atau negatif, serta luas tanah yang terlalu kecil untuk ukuran rumah. Data semacam ini perlu dibersihkan pada tahap persiapan data agar tidak memengaruhi kinerja dan akurasi model Random Forest yang akan dibangun.

BAB III
DATA PREPARATION

3.1. Data Cleaning (Pembersihan Data)

Data mentah yang diekstraksi dari sumber aslinya mengandung format yang tidak standar. Langkah pertama yang dilakukan adalah normalisasi struktur data agar dapat dibaca oleh algoritma komputer:

A. Normalisasi Teks Header

Menghapus karakter spasi berlebih pada nama kolom untuk mencegah kesalahan pemanggilan variabel dalam kode pemrograman.

B. Konversi Tipe Data Harga

Variabel target Price pada data mentah tersimpan dalam format teks (string) dengan karakter titik sebagai pemisah ribuan (contoh: 1.500.000.000). Proses pembersihan dilakukan dengan menghapus karakter non-numerik tersebut dan mengubah tipe datanya menjadi Float atau Integer agar dapat dilakukan operasi matematika.

C. Filtrasi Anomali

Melakukan penghapusan baris data (dropping rows) yang memiliki nilai harga nol atau negatif. Data semacam ini dikategorikan sebagai noise atau kesalahan input sistem yang jika dibiarkan akan merusak akurasi prediksi model regresi.

3.2. Penanganan Data Hilang (Handling Missing Values)

Mengingat data properti sering kali tidak lengkap, strategi imputasi diterapkan untuk mengisi kekosongan data tanpa harus membuang informasi berharga lainnya. Berdasarkan karakteristik 12 atribut yang digunakan, diterapkan dua strategi berbeda:

A. Imputasi Numerik (Median)

Pada variabel numerik seperti Daya Listrik dan Jumlah Lantai, nilai yang hilang diisi menggunakan nilai tengah (Median). Strategi ini dipilih karena data properti cenderung memiliki distribusi yang menceng (skewed). Penggunaan rata-rata (Mean) dihindari karena sangat sensitif terhadap nilai ekstrem (misalnya, satu rumah mewah dengan daya listrik 33.000 VA akan menarik nilai rata-rata menjadi terlalu tinggi dan tidak representatif).

B. Imputasi Kategorikal (Modus/Konstanta)

Pada variabel kategorikal seperti Sertifikat, Kondisi Perabotan, atau Hook, kekosongan data diisi dengan kategori yang paling sering muncul (Most Frequent) atau label "Unknown". Langkah ini untuk mencegah kegagalan sistem saat proses

Encoding, dan memastikan bahwa setiap baris data memiliki label yang jelas sebelum masuk ke tahap pemodelan.

3.3. Transformasi & Rekayasa Fitur (*Feature Engineering*)

Agar algoritma Random Forest dan K-Means dapat bekerja optimal, data yang telah bersih perlu diubah ke dalam format matematis yang seragam. Proses ini dikelola menggunakan ColumnTransformer yang memisahkan perlakuan untuk data numerik dan kategorikal:

A. Standardisasi Numerik (Scaling) keenam variabel numerik (Luas Tanah, Luas Bangunan, Kamar Tidur, Kamar Mandi, Lantai, Listrik) memiliki satuan yang sangat timpang. Luas tanah bisa mencapai angka ratusan, sementara jumlah kamar hanya satuan digit. Tanpa penyesuaian, variabel dengan angka besar akan mendominasi perhitungan jarak pada algoritma K-Means. Oleh karena itu, diterapkan metode StandardScaler, yang mengubah distribusi data sehingga memiliki rata-rata 0 dan standar deviasi 1.

B. Pengkodean Variabel Kategorikal (One-Hot Encoding) melalui Algoritma Machine Learning tidak dapat memproses data teks secara langsung. Oleh karena itu, keenam variabel kategorikal (Kecamatan, Sertifikat, Hook, Perabotan, Kondisi Properti, Internet) diubah menjadi format angka. Teknik yang digunakan adalah One-Hot Encoding. Teknik ini mengubah setiap kategori unik menjadi kolom biner baru (0 atau 1). Sebagai contoh, kolom Hook tidak sekadar diubah menjadi angka 1 atau 2, melainkan dipecah menjadi kolom Hook_Ya dan Hook_Tidak. Pendekatan ini dinilai lebih akurat dibandingkan Label Encoding biasa karena mencegah model salah menafsirkan urutan angka sebagai tingkatan nilai.

BAB IV
MODELING

4.1. Skenario Pemodelan

Strategi pemodelan dirancang untuk menjawab dua permasalahan bisnis yang berbeda melalui dua jalur algoritma:

1. Pendekatan Supervised Learning

   Digunakan untuk menangani masalah Prediksi. Dalam skenario ini, model diberikan data yang memiliki "kunci jawaban" atau label, yaitu harga properti (Price). Tugas model adalah memetakan hubungan kompleks antara 12 variabel input (seperti lokasi, luas, legalitas) dengan variabel target (harga) untuk menghasilkan fungsi estimasi yang akurat.

2. Pendekatan Unsupervised Learning

   Digunakan untuk menangani masalah Segmentasi. Dalam skenario ini, model bekerja pada data tanpa label harga. Tugas model adalah menelusuri kemiripan karakteristik fisik antar-aset dan mengelompokkannya secara mandiri ke dalam klaster-klaster yang merepresentasikan profil properti di pasaran.

4.2. Implementasi Supervised Learning (Estimasi Harga)

Untuk menyelesaikan tugas prediksi harga, dipilih algoritma Random Forest Regressor. Algoritma ini merepresentasikan metode Supervised Learning tipe regresi yang bekerja dengan cara membangun sekumpulan pohon keputusan (Ensemble Learning).

a. Logika Pemilihan

   Data properti memiliki karakteristik non-linear yang kuat. Hubungan antara variabel independen (seperti posisi hook atau status sertifikat) terhadap variabel dependen (harga) sering kali tidak lurus. Random Forest dipilih karena kemampuannya yang unggul dalam menangkap pola non-linear tersebut dibandingkan regresi linear biasa, serta ketangguhannya terhadap outlier yang mungkin tersisa setelah pembersihan data.

b. Konfigurasi Input

   Model ini dilatih menggunakan fitur lengkap (12 atribut), mencakup dimensi fisik, legalitas, hingga faktor eksternal seperti ketersediaan internet, dengan asumsi bahwa harga pasar terbentuk dari kombinasi seluruh faktor tersebut.

c. Parameter

Proses pelatihan menggunakan parameter n_estimators=100, yang berarti model akan mensimulasikan 100 pohon keputusan dan mengambil rata-rata hasilnya untuk prediksi akhir yang stabil.

4.3. Implementasi Unsupervised Learning (Clustering Aset)

Untuk bagian pengelompokan profil aset, diterapkan algoritma K-Means Clustering. Ini adalah metode Unsupervised Learning berbasis centroid yang bertujuan mempartisi data ke dalam kelompok-kelompok yang terpisah berdasarkan jarak kemiripan fisik.

a. Logika Pemilihan

   K-Means dipilih karena efisiensinya dalam mengolah data numerik dalam jumlah besar. Algoritma ini bekerja dengan meminimalkan varians di dalam klaster (jarak intra-klaster) dan memaksimalkan perbedaan antar-klaster.

b. Konfigurasi Input

   Berbeda dengan model supervised, model ini hanya menerima input berupa atribut fisik bangunan (Luas Tanah, Luas Bangunan, Kamar Tidur, Kamar Mandi). Atribut harga dan lokasi sengaja dikecualikan agar pengelompokan murni terbentuk berdasarkan "bentuk fisik" bangunan, bukan berdasarkan nilai jualnya.

c. Parameter

   Jumlah klaster ditentukan sebanyak tiga (k=3). Keputusan ini didasari oleh hipotesis bisnis untuk memetakan pasar ke dalam tiga segmen psikologis pembeli yaitu segmen Pemula (Entry-Level), Menengah (Mid-Tier), dan Atas (High-End) serta telah divalidasi secara statistik menggunakan pendekatan Elbow Method yang menunjukkan penurunan varians optimal pada titik tersebut.

4.4. Mekanisme Pelatihan Model (*Model Training*)

Untuk memastikan evaluasi model berlangsung secara objektif, data dibagi menggunakan metode train-test split. Sebanyak 80% data digunakan sebagai data latih untuk membantu model mengenali pola harga dan karakteristik fisik properti. Dalam konteks unsupervised learning, proses pembelajaran dilakukan dengan mencari titik pusat (centroid) dari sebaran data berdasarkan kemiripan fitur. Sementara itu, 20% data lainnya dijadikan data uji yang tidak dilibatkan dalam proses pelatihan. Data ini berperan sebagai representasi kondisi nyata guna mengukur kinerja model, yang akan dibahas lebih lanjut pada Bab 5 tentang evaluasi. Agar hasil pengujian tetap konsisten setiap kali eksperimen dijalankan, proses pembagian data dilakukan dengan pengaturan random_state = 42, sehingga seluruh percobaan bersifat reproducible.

BAB V
EVALUATION

5.1. Metode Evaluasi

Pemilihan metrik evaluasi didasarkan pada karakteristik data properti yang memiliki varians nilai yang sangat tinggi. Berikut adalah metodologi yang diterapkan:

a. Evaluasi Model Estimasi Harga (Supervised)

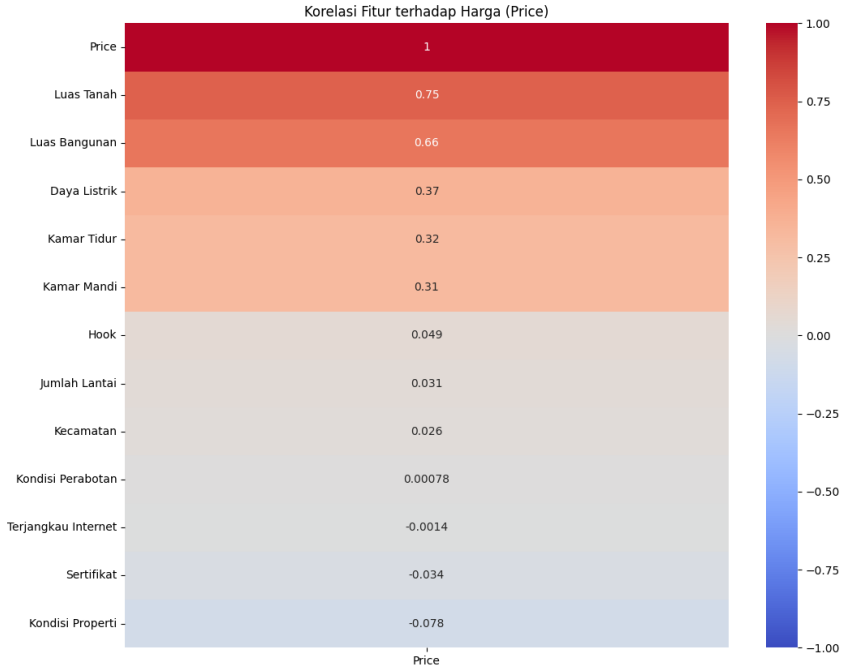
   Evaluasi kinerja model dilakukan dengan menggunakan Root Mean Square Error (RMSE) sebagai metrik utama. RMSE menghitung kesalahan prediksi dengan cara menguadratkan selisih antara nilai aktual dan prediksi sebelum dirata-ratakan. Pendekatan ini membuat kesalahan yang besar mendapat penalti lebih tinggi, sehingga sangat tepat digunakan pada konteks bisnis properti, di mana selisih harga yang ekstrem jauh lebih merugikan dibandingkan kesalahan kecil yang terjadi secara konsisten. Sebagai pelengkap, digunakan pula Mean Absolute Error (MAE) untuk menunjukkan rata-rata selisih harga secara absolut. Metrik ini lebih mudah dipahami oleh pengguna non-teknis karena langsung merepresentasikan besar kesalahan dalam satuan harga.

b. Evaluasi Model Segmentasi (Unsupervised)

   Karena tidak ada label "benar" atau "salah" dalam clustering, evaluasi dilakukan menggunakan Silhouette Score. Metrik ini mengukur seberapa mirip sebuah objek dengan klasternya sendiri dibandingkan dengan klaster lain. Nilai berkisar antara -1 hingga +1, di mana nilai yang semakin mendekati +1 menandakan bahwa klaster terpisah dengan sangat baik dan memiliki densitas yang tinggi. Selain itu, dilakukan validasi kualitatif (Business Logic Validation) untuk memastikan profil yang terbentuk masuk akal secara operasional.

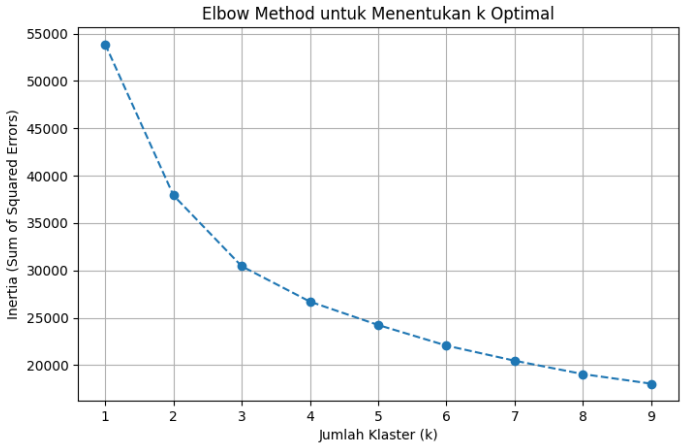
5.2. Analisis Korelasi Fitur

Sebelum tahap pemodelan, dilakukan analisis korelasi untuk melihat seberapa kuat hubungan antara setiap variabel bebas dengan variabel target, yaitu harga. Hasil matriks korelasi menunjukkan bahwa Luas Tanah memiliki hubungan positif paling kuat dengan harga (r = 0,75), disusul oleh Luas Bangunan (r = 0,66). Temuan ini menegaskan bahwa ukuran fisik properti merupakan indikator linear utama dalam penentuan nilai jual. Sebaliknya, variabel Kecamatan memiliki korelasi linear yang sangat rendah (r = 0,026). Hal ini terjadi karena proses label encoding mengubah kategori lokasi menjadi angka tanpa makna urutan nilai, sehingga tidak mencerminkan perbedaan harga secara linear. Meski demikian, pengaruh lokasi tetap penting dan akan terlihat lebih jelas pada analisis non-linear menggunakan algoritma Random Forest.



5.3. Penentuan Parameter Optimal (*Hyperparameter Tuning*)

Dalam pemodelan unsupervised learning, penentuan jumlah klaster (k) dilakukan secara objektif menggunakan Elbow Method. Metode ini menganalisis hubungan antara jumlah klaster dan nilai inertia, yaitu total jarak kuadrat kesalahan dalam setiap klaster. Berdasarkan grafik, terlihat penurunan inertia yang sangat tajam dari k = 1 ke k = 2, lalu mulai melandai setelah k = 3. Titik perubahan ini menunjukkan bahwa penambahan klaster di atas tiga tidak lagi memberikan peningkatan yang berarti dalam homogenitas data. Dengan demikian, k = 3 ditetapkan sebagai jumlah klaster yang paling optimal untuk keperluan segmentasi pasar pada penelitian ini.



5.4. Hasil Pengujian Supervised (Estimasi Harga)

Evaluasi model *Random Forest Regressor* dilakukan secara komprehensif, tidak hanya melihat error prediksi tetapi juga memahami faktor determinan harga.

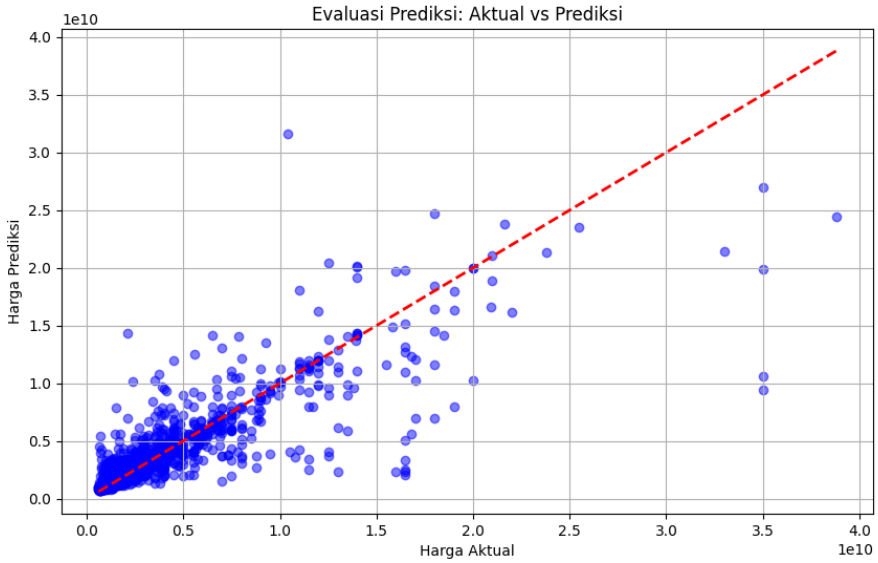
1. Akurasi Model

   Model menghasilkan nilai RMSE sebesar Rp 2,066,729,139. Angka ini merepresentasikan rata-rata deviasi standar kesalahan prediksi model terhadap harga aktual.

```
[SUPERVISED] Running Random Forest...
-> RMSE (Error Rata-rata) : Rp 2,066,729,139
```

2. Prediksi

   Seperti terlihat pada gambar, mayoritas titik data (lingkaran biru) berkumpul rapat di sekitar garis diagonal merah pada rentang harga properti di bawah Rp 10 Miliar. Hal ini menandakan akurasi model sangat tinggi untuk segmen pasar menengah ke bawah. Namun, terlihat adanya varian error yang membesar pada properti High-End (di atas Rp 20 Miliar), di mana model cenderung memberikan prediksi yang lebih konservatif (under-prediction) dibandingkan harga penawaran aktual.".



3. Analisis Fitur Penting (*Feature Importance*):

   mengetahui variabel mana yang paling berkontribusi terhadap pembentukan harga. Berdasarkan grafik Feature Importance yang diekstrak dari model.

   a. Variabel Dominan Mutlak

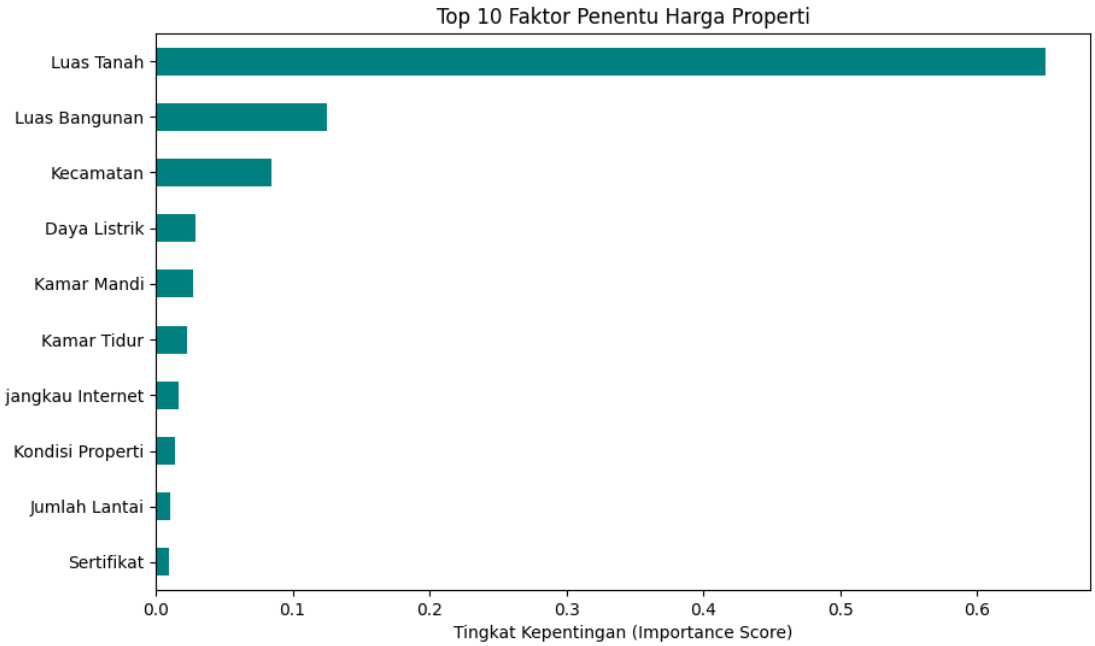
      Variabel Luas Tanah memiliki skor kepentingan tertinggi jauh melampaui variabel lainnya. Temuan ini menunjukkan bahwa nilai tanah merupakan faktor paling dominan dalam menentukan harga properti di Surabaya.

   b. Variabel Sekunder

Luas Bangunan dan Kecamatan berada pada peringkat kedua dan ketiga dalam tingkat kepentingan variabel. Hal ini menunjukkan bahwa meskipun hubungan linearnya rendah, algoritma tetap mampu menangkap pola bahwa lokasi tertentu memiliki pengaruh besar terhadap peningkatan harga properti.

c. Faktor Pendukung

Variabel lain seperti Daya Listrik, Kamar Mandi, dan Kamar Tidur memiliki kontribusi yang merata namun kecil. Posisi daya listrik di peringkat ke-4 mengindikasikan bahwa kapasitas listrik sering kali menjadi proxy (perwakilan) untuk membedakan rumah standar dengan rumah mewah.



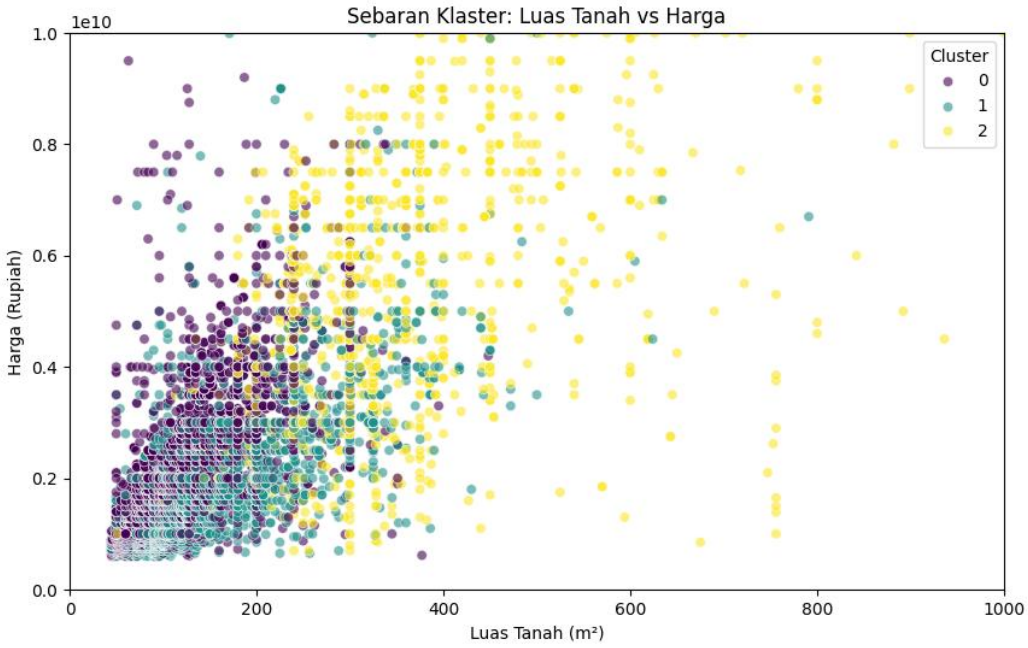
5.5. Hasil Pengujian Unsupervised (Segmentasi Aset)

Evaluasi model *K-Means* difokuskan pada kualitas pemisahan antar-kelompok dan interpretasi profil yang terbentuk.

1. Hasil evaluasi menunjukkan Silhouette Score sebesar 0,300, yang menandakan bahwa tingkat kekompakan data di dalam klaster berada pada kategori sedang. Meskipun nilai ini belum mencerminkan pemisahan klaster yang sangat baik (umumnya > 0,5), kondisi tersebut masih wajar mengingat data properti bersifat heterogen dan memiliki banyak tumpang tindih karakteristik antar-kategori. Secara keseluruhan, klaster yang terbentuk sudah cukup memiliki perbedaan profil yang dapat dimanfaatkan untuk analisis bisnis, meskipun batas antar-klaster belum terlihat sangat tegas.

```
[UNSUPERVISED] Running Clustering...
-> Silhouette Score: 0.300
```

2. Visualisasi pada gambar menunjukkan sebaran klaster yang terbentuk berdasarkan dimensi fisik aset. Meskipun terdapat irisan antar-klaster akibat variasi jumlah lantai dan luas bangunan, secara umum terbentuk pola distingsi sebagai berikut:



a. Cluster 0 (Entry-Level/Compact) ditandai dengan titik-titik berwarna Ungu yang terkonsentrasi padat di area kiri bawah grafik. Klaster ini didominasi oleh aset dengan luas tanah terbatas (mayoritas di bawah 100 m²). Segmen ini merepresentasikan hunian kompak yang efisien untuk keluarga baru.
b. Cluster 1 (Mid-Tier/Standard) ditandai dengan titik-titik berwarna Hijau yang menyebar di area tengah. Klaster ini merupakan zona transisi yang mencakup properti dengan luas tanah menengah (kisaran 100 - 250 m²). Segmen ini merepresentasikan pasar properti standar yang paling umum ditemui.
c. Cluster 2 (High-End/Luxury) ditandai dengan titik-titik berwarna Kuning yang memiliki sebaran paling luas, bergerak ke arah kanan atas grafik. Klaster ini mencakup aset dengan luas tanah besar (mayoritas > 200 m²) serta memiliki korelasi dengan harga yang tinggi. Segmen ini merepresentasikan properti Luxury dengan spesifikasi ruang yang lebih kompleks.

5.6. Analisis Hasil & Diskusi

Pada model supervised learning, nilai RMSE menunjukkan bahwa model sudah mampu mengikuti pola harga pasar, meskipun masih terdapat selisih prediksi. Hal ini wajar terjadi pada pasar properti sekunder karena adanya faktor-faktor yang tidak tercatat dalam data,

seperti akses jalan, kualitas material bangunan, atau nilai emosional pemilik. Meski demikian, sebagai alat estimasi awal, tingkat akurasi ini sudah cukup untuk memberikan kisaran harga yang masuk akal bagi agen properti dan jauh lebih andal dibandingkan perkiraan manual.

Sementara itu, pada model unsupervised learning, hasil klasterisasi menunjukkan bahwa algoritma berhasil memetakan struktur pasar secara alami. Tanpa label awal, metode K-Means mampu mengelompokkan properti ke dalam tiga segmen utama: Entry-Level, Mid-Tier, dan High-End. Temuan ini menegaskan bahwa karakteristik fisik bangunan, seperti luas dan jumlah ruang, memiliki peran kuat dalam segmentasi pasar. Walaupun masih terdapat tumpang tindih antar-segmen, fitur asset profiling pada aplikasi tetap dapat diandalkan untuk melakukan pengelompokan aset secara otomatis.

BAB VI
DEPLOYMENT

6.1. Arsitektur Sistem

Sistem ini dibangun dengan arsitektur web-based model serving untuk memastikan integrasi yang lancar antara proses di sisi server dan tampilan antarmuka pengguna.

1. Penyimpanan Model (Model Persistence)

   Pada lingkungan produksi, model tidak dilatih ulang setiap kali aplikasi dijalankan. Model Random Forest dan K-Means yang telah dilatih sebelumnya disimpan dalam format .joblib agar dapat dimuat dengan cepat. Selain itu, sistem juga menyimpan metadata opsi seperti daftar kecamatan dan jenis sertifikat, sehingga pilihan input pada aplikasi selalu sesuai dengan data pelatihan.

2. Mesin Back-End (Flask)

   Dikembangkan menggunakan framework Flask dengan fungsi utama:

   a. Memuat model dan metadata saat aplikasi dijalankan.
   b. Menerima 12 parameter input dari pengguna melalui metode HTTP POST.
   c. Melakukan pra-pemrosesan data secara real-time sesuai dengan tahapan pada Bab 3, termasuk imputasi dan normalisasi.
   d. Mengirimkan kembali hasil prediksi harga dan klaster properti kepada pengguna.

3. Antarmuka Front-End (User Experience)

   Dibangun menggunakan HTML5 dan Tailwind CSS, dengan penerapan responsive web design sehingga aplikasi dapat digunakan dengan baik di perangkat desktop maupun mobile.

6.2. Implementasi Antarmuka Pengguna

Antarmuka aplikasi dirancang untuk mengurangi kesalahan input sekaligus tetap memberi fleksibilitas kepada pengguna melalui strategi berikut:

1. Input Terstandarisasi (Dropdown Menu)

   Untuk variabel kategorikal yang memiliki nilai baku, seperti Daya Listrik, Jenis Sertifikat, dan Posisi Hook, sistem menggunakan menu dropdown. Hal ini mencegah pengguna memasukkan data yang tidak valid sehingga hasil prediksi tetap akurat.

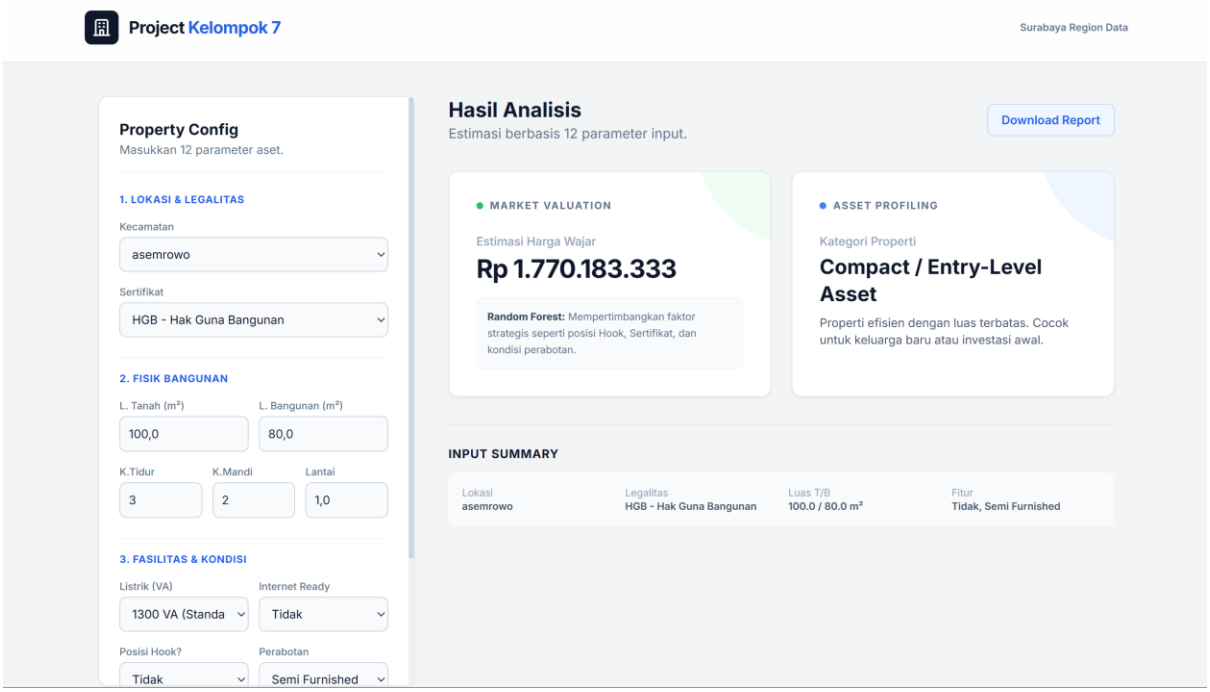
2. Input Kontinu (Kolom Angka)

Untuk variabel fisik seperti Luas Tanah dan Luas Bangunan, disediakan kolom angka bebas. Hal ini memungkinkan model Random Forest melakukan prediksi pada berbagai ukuran properti, termasuk nilai yang tidak secara eksplisit ada di data latih.

3. Visualisasi Hasil (Dashboard)

   Hasil prediksi tidak hanya ditampilkan sebagai angka, tetapi juga disajikan secara informatif:

   a. Valuasi harga ditampilkan dalam format mata uang Rupiah yang mudah dibaca.
   b. Profil aset ditampilkan dalam bentuk kartu informasi yang menjelaskan karakteristik segmen, misalnya "Cluster 2 High End Asset", sehingga hasil klasterisasi lebih mudah dipahami secara bisnis.



BAB VII PENUTUP

7.1. Kesimpulan

Berdasarkan seluruh proses penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan teknik data mining memberikan kontribusi nyata dalam mengurangi inefisiensi serta subjektivitas yang selama ini kerap terjadi pada pasar properti di Surabaya. Melalui pendekatan supervised learning, algoritma Random Forest Regressor mampu membentuk sistem penilaian harga yang lebih objektif dan konsisten. Hasil pengujian menunjukkan bahwa model ini cukup andal dalam mengikuti tren harga pasar, khususnya pada segmen properti dengan nilai di bawah Rp10 miliar, sehingga dapat dimanfaatkan sebagai alternatif yang lebih rasional dibandingkan metode penaksiran manual yang sering bergantung pada intuisi semata.

Dari sisi pemetaan pasar, penggunaan K-Means Clustering juga memberikan gambaran yang menarik. Tanpa perlu penentuan kategori sejak awal, algoritma ini berhasil mengelompokkan properti berdasarkan pola yang terbentuk secara alami. Analisis menunjukkan bahwa karakteristik fisik bangunan, terutama luas tanah dan jumlah ruang, memiliki peran besar dalam membedakan kelas ekonomi aset. Secara umum, pasar dapat dipetakan ke dalam tiga kelompok utama, yaitu Entry-Level, Mid-Tier, dan High-End, yang masing-masing merepresentasikan tingkat nilai dan segmen pembeli yang berbeda.

Temuan tersebut semakin diperkuat melalui analisis kepentingan variabel. Luas Tanah muncul sebagai faktor paling dominan dalam menentukan harga, kemudian diikuti oleh Luas Bangunan serta faktor lokasi administratif (Kecamatan). Menariknya, Daya Listrik juga menunjukkan pengaruh yang cukup signifikan, khususnya dalam membedakan tingkat kemewahan properti. Kombinasi seluruh variabel ini memberikan gambaran bahwa proses valuasi tidak hanya ditentukan oleh satu aspek saja, melainkan oleh interaksi berbagai karakteristik fisik dan lingkungan yang saling melengkapi.

Link Folder Tugas Besar

Link github:
https://github.com/Kifansa/Sistem-Valuasi-Cerdas-dan-Segmentasi-Pasar-Properti-Surabaya