

PREDICTING USED CAR PRICES IN SAUDI ARABIA

CAPSTONE MODUL 3
MUHAMMAD SATRIA YUDHA MAHENDRA



PROJECT OVERVIEW



- 1. Rumusan Masalah**
- 2. Tujuan**
- 3. Deskripsi Data**
- 4. Asumsi dan Limitasi**
- 5. Pendekatan Analitik**
- 6. Pemilihan model**
- 7. Kesimpulan & Rekomendasi**

RUMUSAN MASALAH

Masalah Utama:

- Penentuan harga mobil bekas sering **bergantung** pada **intuisi** dan **pengalaman**, sehingga rentan terhadap **bias** manusia.

Dampak Negatif Jika Tidak Diatasi:

- Negosiasi **kurang efisien**.
- **Risiko** kesepakatan harga di bawah atau di atas nilai pasar sebenarnya.

Kebutuhan:

- **Sistem prediksi harga** yang konsisten dan berbasis data.
- **Mempertimbangkan karakteristik** mobil (merek, model, tahun, kondisi, dll.) untuk memberikan rekomendasi harga yang lebih akurat.

TUJUAN

Predictive Model:

- Membangun **model machine learning** untuk **memprediksi rentang harga ideal** mobil bekas berdasarkan fitur seperti merek, model, ukuran mesin, tahun, dan kondisi.

Optimize Workflow:

- **Mengotomatiskan** proses inspeksi dan negosiasi harga, mengurangi bias manusia, dan meningkatkan akurasi keputusan.

Support Negotiation:

- Memberikan **rekomendasi** harga **berbasis data** untuk membantu Tim Penilaian & Negosiasi menentukan harga akhir dengan penjual.

Consistency & Transparency:

- Menjadikan proses **penetapan harga** lebih **konsisten, transparan, dan adil**.

DESKRIPSI DATA

Overview:

- Dataset berisi data mobil bekas yang dijual di Saudi Arabia melalui **Syarah.com**.

Key Columns:

- **Type**: Tipe mobil bekas.
- **Region**: Lokasi penjualan mobil bekas.
- **Make**: Merek mobil.
- **Gear_Type**: Tipe transmisi mobil.
- **Origin**: Asal mobil.
- **Options**: Fitur atau opsi mobil.
- **Year**: Tahun pembuatan.
- **Engine_Size**: Kapasitas mesin.
- **Mileage**: Jarak tempuh mobil.
- **Negotiable**: Apakah harga bisa dinegosiasi.
- **Price**: Harga mobil bekas.

ASUMSI DAN LIMITASI

Assumptions

- Data diperoleh hingga **20 Desember 2022**.
- Referensi harga pasar dari platform: **Syarah, CarSemsar, dan Hatla2ee**.
- Model menggunakan dataset terkini yang mencerminkan **tren pasar saat itu**.

Limitations

- Tidak mencakup **perubahan pasar di masa depan**.
- Dataset **tidak mencakup semua fitur mobil** atau semua **wilayah** di Saudi Arabia.
- Data **terbatas** pada fitur dasar seperti **merek, tipe, dan kapasitas mesin**.
- Proses menggunakan **random state = 1001** untuk konsistensi.

PENDEKATAN ANALITIK

Stages:

1. Data Understanding

- Analisis struktur dataset dan identifikasi fitur penting.
- Cek missing values, outliers, dan duplikasi data.

2. Data Exploration

- Lakukan EDA untuk memahami distribusi fitur dan hubungan dengan harga.
- Identifikasi pola anomali dan nilai ekstrem.

3. Data Cleaning

- Tangani nilai hilang, outliers, dan encoding fitur kategorikal.
- Pilih fitur relevan untuk meningkatkan performa model.

4. Modeling

- Uji algoritma: XGBoost, Random Forest, Linear Regression, dll.
- Optimasi menggunakan cross-validation & hyperparameter tuning.

PENDEKATAN ANALITIK (LANJUTAN)

5. Model Evaluation

- Evaluasi performa model melalui residual analysis.

6. Interpretability

- Gunakan SHAP untuk menjelaskan pengaruh fitur terhadap prediksi.

7. Conclusion & Recommendations

- Ringkasan kekuatan, keterbatasan, dan insights dari model.
- Rekomendasi untuk peningkatan data dan analisis di masa depan.

PEMILIHAN MODEL

	RMSE	MAE	MAPE	R2 Score
XGB Regressor	16535.633335	11218.903345	0.186000	0.792073
RandomForest	17075.569520	11582.015896	0.199472	0.778273
KNN Regressor	18234.836701	12779.686424	0.228487	0.747145

Evaluasi Model: XGBoost, RandomForest, & KNN Regressors

1.XGBoost Regressor

Model Terbaik:

- R^2 : 0.7819 (tertinggi)
- Kesalahan terendah (RMSE, MAE, MAPE)
- Memberikan prediksi yang paling akurat.

2.RandomForest Regressor

Alternatif Kuat:

- R^2 : 0.7770
- Kesalahan sedikit, tetapi masih memberikan hasil yang baik.

3. K-Nearest Neighbors (KNN) Regressor

- Kinerja Kurang Baik:
 - R^2 : 0.7390 (terendah)
 - Kesulitan dalam menangkap kompleksitas data, dengan kesalahan tertinggi.
 -

Kesimpulan

XGBoost Regressor adalah model yang optimal untuk tugas ini, memberikan kinerja superior di semua metrik.

KESIMPULAN

Model prediksi harga mobil bekas di Syarah.com menunjukkan **hasil yang menjanjikan namun masih perlu perbaikan.**

Wawasan Utama:

- **Akurasi:** 50% prediksi akurat, 8% jauh dari target, dan 42% di luar rentang yang diharapkan.
- **Fitur Utama:** Ukuran mesin dan tahun mempengaruhi harga mobil dengan kuat.
- **Data Latih:** Dataset kecil dan ketidakseimbangan data membatasi generalisasi model.
- **Tantangan:** Kesulitan menangani nilai ekstrem dan kasus batas.

Poin Perhatian:

- **Ukuran Mesin:** Inkonsistensi pada ukuran mesin di bawah 2000cc dan di atas 7000cc dapat menyebabkan prediksi harga yang tidak realistik.
- **Kesalahan Data:** Potensi kesalahan entri atau varian yang tidak tercatat.

REKOMENDASI

Beberapa langkah yang dapat dilakukan untuk meningkatkan model:

- **Perluas Dataset:** Tingkatkan jumlah data pelatihan dengan menambah variasi atribut mobil, terutama yang terkait dengan kondisi mobil dan rentang harga, agar model dapat lebih baik dalam generalisasi dan mengatasi kasus ekstrem.
- **Perbaiki Teknik Fitur:** Tambahkan fitur-fitur penting seperti kondisi mobil, modifikasi, dan riwayat kendaraan, serta pastikan data ukuran mesin lebih konsisten untuk meningkatkan akurasi model.
- **Tanggulangi Inkonsistensi Ukuran Mesin:** Periksa dan bersihkan data ukuran mesin untuk merek premium seperti BMW, Mercedes, dan Lexus, dengan memastikan ukuran mesin yang sangat kecil atau sangat besar ditangani dengan benar untuk menghindari prediksi yang tidak realistik.
- **Pantau Kinerja Model:** Lakukan evaluasi berkala terhadap kinerja model, terutama pada prediksi untuk kasus ekstrem atau outliers, guna memastikan model mampu beradaptasi dengan perubahan data dan ketidaksesuaian yang ada.

THANK YOU

