Prediksi Harga Diamond dengan Machine Learning

PROYEK
DATA MINING

01

Amelia Renata Kumalanisa

4111422026

02

Aditya Pratama Juliyawan

4111422029

03

Daffa Syauqi Raihan

4111422034

04

Angga Yulian Adi Pradana

41114220243

Business Understanding

Diamond memiliki pengaruh yang sangat luas dan beragam di dunia, mencakup berbagai aspek seperti ekonomi, budaya, sosial, dan teknologi. Dengan nilai investasi yang tinggi dan peran yang luas dalam industri, penelitian, dan bidang medis, diamond memiliki potensi yang luar biasa. Mengetahui harga diamond ke depannya memungkinkan kita membuat keputusan yang cerdas dan tepat waktu, baik dalam bisnis maupun kehidupan sehari-hari.

Tujuan :

Memprediksi harga diamond sehingga perusahaan, investor, dan pelaku pasar mampu membuat keputusan yang lebih baik terkait investasi, produksi, strategi pemasaran, dan manajemen risiko untuk mengoptimalkan kinerja bisnis dalam jangka panjang.

Data Understanding

Dataset

- **carat**: berat fisik diamond. Nilai berkisar dari 0.2 hingga 5.01 carat.
- **cut**: kualitas potongan berlian, semakin presisi potongannya, maka nilainya semakin tinggi.
- **color**: warna berlian, mulai dari J (terburuk) hingga D (terbaik).
- **clarity:** tingkat kejernihan berlian, yang berkisar dari I1 (terburuk) hingga IF (terbaik).
- **depth**: persentase kedalaman total berlian, dihitung sebagai z / mean(x, y) = 2 * z / (x + y). Semakin tinggi nilainya, semakin dalam berlian tersebut.
- **table:** persentase lebar bagian atas berlian (meja) sebagai persentase dari lebar rata-rata. Meja berlian yang ideal berkontribusi pada kecemerlangan berlian.
- **Price:** Harga berlian dalam dolar AS. Ini adalah kolom target dalam dataset ini.
- **x, y, z:** Dimensi panjang, lebar, dan kedalaman berlian dalam milimeter.

Data preparation

Data Preprocessing

Menghapus Kolom Tidak Berguna:

- Menghapus kolom Unnamed: 0 yang tidak relevan untuk analisis.

Menghapus Nilai Tidak Masuk Akal:

- Menghapus baris yang memiliki nilai 0 pada kolom x, y, atau z karena tidak mungkin ada berlian dengan dimensi 0.

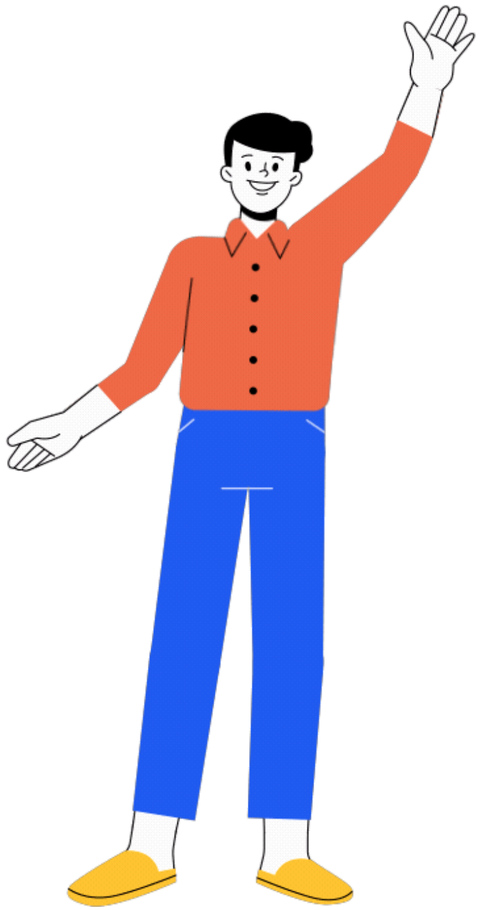
Mengevaluasi Fitur Kategorikal

1. Distribusi Kolom Kategorikal:
   - Cut: Distribusi harga berdasarkan kualitas potongan berlian.
   - Color: Distribusi harga berdasarkan warna berlian.
   - Clarity: Distribusi harga berdasarkan kejernihan berlian.
2. Visualisasi dan Temuan:
   - Menggunakan violin plot untuk melihat distribusi harga berlian terhadap cut, color, dan clarity.
   - Temuan: Berlian dengan potongan "Ideal" dan warna "D" serta kejernihan "IF" cenderung memiliki harga lebih tinggi.
   - Distribusi harga bervariasi di antara kategori, menunjukkan pengaruh fitur kategorikal terhadap harga.

Memeriksa Outlier

Visualisasi Outlier:

- Price vs Y: Plot garis menunjukkan beberapa outliers pada y.
- Price vs Z: Plot garis menunjukkan beberapa outliers pada z.
- Price vs Depth: Plot garis menunjukkan beberapa outliers pada depth.
- Price vs Table: Plot garis menunjukkan beberapa outliers pada table.



Data Preparation

Menghapus outlier

Outlier dihapus dengan mengidentifikasi dan menghilangkan baris dengan nilai 0 pada kolom dimensi (x, y, z), serta baris dengan depth di luar rentang 45-75 dan table di luar rentang 40-80. Selain itu, berlian dengan nilai x atau y lebih dari 40, serta z lebih dari 40 atau kurang dari 2, juga dihapus. Proses ini memastikan dataset bersih dari nilai tidak masuk akal, meningkatkan akurasi dan kinerja model prediksi harga berlian.

Encoding Variabel Kategorik:

- Mengubah fitur kategorikal (cut, color, clarity) menjadi numerik menggunakan LabelEncoder untuk persiapan pemodelan.

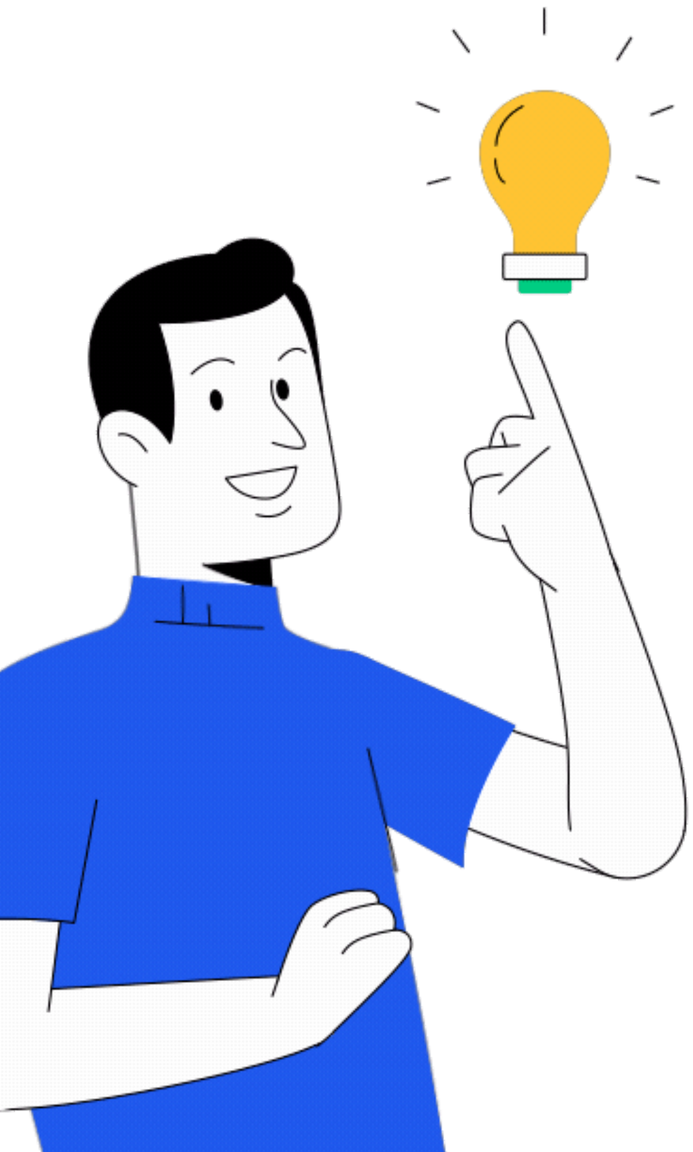
Matriks Korelasi

Pemeriksaan Korelasi:

Korelasi Tinggi: Fitur carat, x, y, z menunjukkan korelasi tinggi dengan harga berlian.

Korelasi Rendah: Fitur cut, clarity, dan depth memiliki korelasi rendah dengan harga, tetapi tetap digunakan dalam model karena datasetnya kecil.

Data yang sudah diproses siap untuk dikenakan model regressi



Model Building

Persiapan:

- Variabel Independen (X): Semua fitur kecuali 'harga'
- Variabel Dependen (y): 'harga'
- Pembagian Data: Train-Test Split (80% Train, 20% Test)

Pipeline model:

- StandardScaler: Menstandarkan data
- Model:
  - Linear Regression
  - Lasso Regression
  - Decision Tree Regressor
  - Random Forest Regressor
  - K-Nearest Neighbors Regressor
  - XGBoost Regressor

Implementasi pipeline

Evaluasi Model:

- Cross-Validation: 12-fold CV dengan RMSE scoring
- Hasil:
  - Linear Regression: 1383.85
  - Lasso: 1366.99
  - Decision Tree: 738.92
  - Random Forest: 548.62
  - K-Neighbors: 816.56
  - XGBoost: 548.35

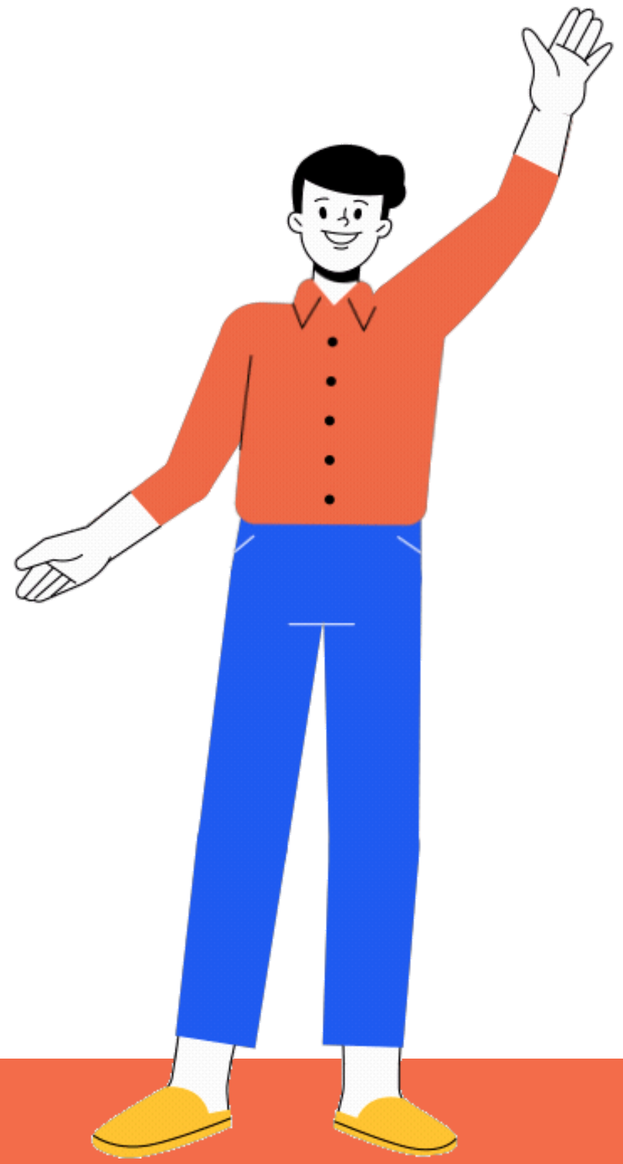
Evaluation

Berdasarkan nilai RMSE, model XGBoost dipilih karena memiliki nilai RMSE paling rendah. Kemudian, prediksi pada data pengujian dengan model XGBoost diperoleh:

- R^2 (Koefisien Determinasi): 0.9821
- Adjusted R^2: 0.9821

Tabel Perbandingan Harga Asli, Harga Prediksi, dan Selisih:

	Actual	Predicted	Difference
31712	771	865.756165	-94.756165
19865	8419	8786.837891	-367.837891
42610	505	522.002380	-17.002380
29785	709	710.018433	-1.018433
20340	8739	10010.160156	-1271.160156
...	...	...	...
50799	2306	2285.798828	20.201172
40238	1124	1236.950806	-112.950806
23860	11951	11353.575195	597.424805
11809	5090	4685.851562	404.148438
39776	1094	948.306824	145.693176



Summary

Kesimpulan:

- Model XGBoost Regresi menunjukkan kinerja terbaik dengan RMSE terendah di antara model lainnya.
- Nilai R^2 dan adjusted R^2 pada data uji mengkonfirmasi kemampuan prediksi yang sangat baik dari model XGBoost.

Saran:

- Rekomendasi Utama: Mengingat kinerja unggul model XGBoost, disarankan untuk menggunakan model ini untuk prediksi harga diamond (berlian).
- Rekomendasi Tambahan: Random Forest juga menunjukkan nilai RMSE yang rendah, sehingga dapat dipertimbangkan sebagai alternatif.