

## ANALISIS PREDIKSI HARGA BERLIAN MENGUNAKAN METODE REGRESI LINIER

Almi Agustina, Yoana Nabilah Putri, Hoirun Nisah,  
Shafa Dhiya Raihanah, Ahmad Najib, Vitri Tundjungsari

Teknik Informatika, Universitas Esa Unggul  
Jalan Arjuna Utara No.9, Kebon Jeruk 11510, Jakarta Barat, Indonesia  
yoananabilahp@student.esaunggul.ac.id

### ABSTRAK

Prediksi harga berlian merupakan permasalahan penting dalam analisis data karena dipengaruhi oleh berbagai karakteristik fisik yang bersifat numerik. Namun, kontribusi relatif masing-masing variabel terhadap harga belum sepenuhnya dipahami secara kuantitatif. Penelitian ini bertujuan untuk menganalisis pengaruh variabel *carat*, *depth*, *table*, serta dimensi fisik *x*, *y*, dan *z* terhadap harga berlian menggunakan metode regresi linier. Data penelitian diperoleh dari *Seaborn Diamonds Dataset* dengan pengambilan sampel acak sebanyak 50 observasi, yang dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Kinerja model dievaluasi menggunakan metrik *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, dan koefisien determinasi ( $R^2$ ). Hasil penelitian menunjukkan nilai *MAE* sebesar 1.200,69, *MSE* sebesar 2.582.825,77, dan  $R^2$  sebesar 0,944, yang mengindikasikan bahwa model regresi linier mampu menjelaskan sebagian besar variasi harga berlian berdasarkan variabel numerik yang digunakan. Dengan demikian, penelitian ini menunjukkan bahwa regresi linier efektif digunakan dalam estimasi harga berlian berbasis data numerik. Pengembangan model melalui penambahan atribut kategorikal dan penerapan metode regresi non-linear berpotensi meningkatkan akurasi serta kemampuan generalisasi model dalam analisis prediktif.

**Kata kunci :** Regresi Linier, Prediksi Harga, Berlian, Data Mining, Evaluasi Model, Variabel Numerik

### 1. PENDAHULUAN

Perkembangan era *big data* telah mengubah cara manusia memanfaatkan informasi, di mana data berukuran besar dan kompleks kini dapat diolah untuk menghasilkan wawasan yang bernilai. Metode analitik seperti regresi linier tetap memiliki peran penting karena kemampuannya mengolah data numerik dalam jumlah besar secara efisien dan memberikan interpretasi yang mudah dipahami [1], serta telah terbukti memberikan tingkat kesalahan prediksi yang rendah dibandingkan metode lain seperti *K-Nearest Neighbor* dan *Support Vector Machine* dalam kasus prediksi harga komoditas [2]. Dalam konteks ini, pendekatan statistik menjadi dasar dalam mengidentifikasi hubungan antarvariabel, terutama pada data yang memiliki pola linear.

Salah satu cabang penting dalam analisis *big data* adalah *Data Mining*, yaitu proses mengekstraksi pola dan informasi bermakna dari kumpulan data besar yang tersimpan dalam basis data [3]. *Data Mining* berperan besar dalam mendukung pengambilan keputusan berbasis data di berbagai bidang, mulai dari ekonomi hingga manufaktur [4], serta telah banyak diterapkan dalam konteks prediksi harga, analisis tren, dan peramalan permintaan menggunakan metode regresi linier berganda [5]. Melalui penerapan teknik seperti regresi linier, *Data Mining* mampu mengubah data mentah menjadi pengetahuan yang aplikatif dan bernilai strategis [6].

Dalam industri perdagangan berlian, teknik *Data Mining* dapat digunakan untuk menganalisis hubungan antara karakteristik fisik berlian dan harga jualnya. Faktor-faktor seperti berat karat (*carat*), kedalaman

potongan (*depth*), proporsi meja (*table*), serta dimensi panjang, lebar, dan tinggi (*x*, *y*, *z*) merupakan indikator penting yang memengaruhi harga [7]. Dengan menganalisis variabel-variabel tersebut, peneliti dapat membangun model prediksi harga berlian yang lebih akurat. Namun, hubungan antarvariabel tidak selalu dapat dipahami secara langsung, sehingga diperlukan model analitik yang mampu menjelaskan pengaruh masing-masing variabel secara sistematis. Selain itu, kontribusi relatif setiap variabel fisik terhadap harga berlian dalam konteks data numerik yang dianalisis belum sepenuhnya teridentifikasi secara jelas.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menerapkan metode regresi linier dalam memprediksi harga berlian berdasarkan enam variabel numerik, yaitu *carat*, *depth*, *table*, *x*, *y*, dan *z*, dengan *price* sebagai variabel target. Penelitian ini juga bertujuan untuk menganalisis hubungan antara atribut fisik berlian dan harga jualnya, serta mengevaluasi efektivitas metode regresi linier dalam konteks analisis data numerik.

Untuk mencapai tujuan tersebut, penelitian ini dilakukan melalui beberapa tahapan, yaitu pengumpulan data dari *Seaborn Diamonds Dataset* sebanyak 50 sampel, pembangunan model regresi linier, serta pengujian menggunakan metrik *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, dan  $R^2$ . Melalui tahapan tersebut, penelitian diharapkan dapat memberikan kontribusi dalam penerapan *Data Mining* dan analitik prediktif pada era *big data*, serta menghasilkan model prediksi harga berlian yang lebih objektif dan terukur.

Selain aspek metodologis, pemilihan regresi linier dalam penelitian ini juga didasarkan pada kebutuhan akan model yang transparan dan mudah diinterpretasikan. Berbeda dengan beberapa metode pembelajaran mesin yang bersifat *black box*, regresi linier memungkinkan peneliti dan praktisi untuk memahami arah serta besar pengaruh setiap variabel independen terhadap variabel *dependent*. Hal ini menjadi penting dalam konteks akademik maupun industri, karena hasil analisis tidak hanya ditujukan untuk memperoleh nilai prediksi semata, tetapi juga untuk memberikan pemahaman konseptual mengenai faktor-faktor yang paling dominan dalam menentukan harga berlian. Dengan demikian, model yang dihasilkan dapat dijadikan dasar pertimbangan dalam pengambilan keputusan, baik oleh pelaku industri maupun oleh peneliti selanjutnya.

Di sisi lain, penggunaan dataset berlian yang telah banyak digunakan dalam penelitian sebelumnya memberikan keuntungan berupa validitas dan keterbandingan hasil. *Seaborn Diamonds Dataset* dikenal memiliki struktur data yang rapi serta variabel numerik yang relevan dengan karakteristik fisik berlian. Namun demikian, dalam penelitian ini dilakukan pembatasan jumlah sampel menjadi 50 data untuk menyesuaikan dengan tujuan analisis dan ruang lingkup penelitian. Pendekatan ini diharapkan tetap mampu merepresentasikan pola hubungan antarvariabel, sekaligus memudahkan proses analisis dan interpretasi hasil model regresi linier yang dibangun.

Pendekatan kuantitatif yang digunakan dalam penelitian ini menempatkan data numerik sebagai dasar utama dalam penarikan kesimpulan. Melalui analisis statistik dan evaluasi performa model, penelitian ini berupaya menilai sejauh mana regresi linier mampu menangkap pola hubungan antara atribut fisik berlian dan harga jualnya. Evaluasi menggunakan *MAE*, *MSE*, dan  $R^2$  dipilih karena ketiga metrik tersebut dapat memberikan gambaran yang komprehensif mengenai tingkat kesalahan prediksi serta kemampuan model dalam menjelaskan variasi data. Dengan adanya evaluasi ini, hasil penelitian diharapkan tidak hanya bersifat deskriptif, tetapi juga memiliki dasar kuantitatif yang kuat.

Pemilihan metode regresi linier dalam penelitian ini juga mempertimbangkan karakteristik data yang dianalisis, yaitu data numerik dengan kecenderungan hubungan linear antarvariabel. Regresi linier memiliki keunggulan dalam kesederhanaan model serta kemudahan interpretasi hasil, sehingga memungkinkan peneliti untuk mengidentifikasi pengaruh masing-masing variabel *independent* secara kuantitatif. Dalam konteks analisis harga berlian, kemampuan ini menjadi penting karena setiap atribut fisik memiliki kontribusi yang berbeda terhadap nilai jual, dan hubungan tersebut perlu dijelaskan secara sistematis berdasarkan pendekatan statistik.

Selain itu, penggunaan variabel *carat*, *depth*, *table*, serta dimensi *x*, *y*, dan *z* memungkinkan analisis

yang lebih objektif karena variabel-variabel tersebut bersifat terukur dan konsisten. Pendekatan ini mengurangi ketergantungan pada faktor subjektif yang sulit dikuantifikasi, seperti persepsi visual atau preferensi pasar. Dengan memfokuskan analisis pada variabel numerik, model regresi linier yang dibangun diharapkan mampu merepresentasikan hubungan matematis yang mendasari pembentukan harga berlian secara lebih akurat dan rasional.

Dalam perspektif *Data Mining*, regresi linier berfungsi sebagai metode dasar yang efektif untuk mengeksplorasi pola hubungan antarvariabel sebelum diterapkan pendekatan yang lebih kompleks. Model yang dihasilkan tidak hanya memberikan nilai prediksi, tetapi juga informasi mengenai arah dan besarnya pengaruh setiap variabel terhadap harga. Informasi ini dapat dimanfaatkan untuk memahami struktur data serta sebagai dasar pengambilan keputusan yang berbasis data, baik dalam konteks akademik maupun praktis.

Penggunaan jumlah sampel yang terbatas dalam penelitian ini juga mencerminkan pendekatan eksploratif dalam analisis data. Fokus utama penelitian diarahkan pada pemahaman hubungan antarvariabel dan evaluasi kinerja metode regresi linier terhadap data numerik yang tersedia. Meskipun skala data belum mencerminkan *big data* secara penuh, hasil penelitian tetap diharapkan mampu menggambarkan pola hubungan linear yang relevan serta memberikan gambaran awal mengenai efektivitas metode yang digunakan.

## 2. TINJAUAN PUSTAKA

### 2.1. Regresi Linear

Regresi linier merupakan metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel *dependent* dengan satu atau lebih variabel *independent* dalam bentuk persamaan garis lurus [8]. Metode ini bertujuan untuk memprediksi nilai variabel target berdasarkan kombinasi linier dari variabel *input*. Keunggulan regresi linier terletak pada kemudahan interpretasi, kestabilan model, serta efisiensi komputasi, sehingga banyak digunakan dalam pengolahan data numerik berskala besar [9].

Pada regresi linier berganda, lebih dari satu variabel *independent* digunakan untuk menjelaskan variabel *dependent*. Pendekatan ini memungkinkan analisis kontribusi masing-masing atribut terhadap nilai target, sehingga tidak hanya berfungsi sebagai alat prediksi, tetapi juga sebagai sarana untuk memahami hubungan antarvariabel [4].

### 2.2. Data Mining dan Analitik Prediktif

*Data Mining* merupakan proses ekstraksi pola, hubungan, dan informasi bermakna dari kumpulan data besar yang tersimpan dalam basis data. Teknik ini memiliki peran penting dalam mendukung pengambilan keputusan berbasis data di berbagai bidang, seperti ekonomi, manufaktur, dan perdagangan digital [2].

Dalam analitik prediktif, perencanaan dan pengambilan keputusan yang efektif membutuhkan prediksi kuantitatif yang andal terhadap variabel yang dianalisis [10]. Untuk memenuhi kebutuhan tersebut, *Data Mining* memanfaatkan metode statistik dan pembelajaran mesin untuk memperkirakan nilai di masa depan berdasarkan data historis. Regresi linier merupakan salah satu metode dasar yang sering digunakan untuk memprediksi nilai kontinu karena mampu memodelkan hubungan numerik secara langsung dan mudah diinterpretasikan [11].

### 2.3. Karakteristik Fisik Berlian

Harga berlian dipengaruhi oleh berbagai atribut fisik, antara lain berat karat (*carat*), kedalaman potongan (*depth*), proporsi meja (*table*), serta dimensi geometris yang meliputi panjang (*x*), lebar (*y*), dan tinggi (*z*). Atribut-atribut tersebut memiliki hubungan kuantitatif terhadap harga jual, sehingga sesuai dianalisis menggunakan pendekatan regresi linier [3].

Dataset *Diamonds* yang tersedia pada pustaka *Seaborn* banyak digunakan dalam penelitian karena menyediakan atribut numerik tersebut secara lengkap beserta harga aktual sebagai variabel target. Beberapa penelitian menunjukkan bahwa variabel *carat* memiliki pengaruh paling dominan terhadap harga berlian, diikuti oleh dimensi fisik lainnya, sementara *depth* dan *table* memberikan kontribusi yang relatif lebih kecil [12].

### 2.4. Evaluasi Model

Kinerja model regresi umumnya dievaluasi menggunakan beberapa metrik, yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan koefisien determinasi ( $R^2$ ). MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual, sedangkan MSE menghitung rata-rata kuadrat selisih prediksi sehingga lebih sensitif terhadap kesalahan besar [5].

Sementara itu, nilai  $R^2$  digunakan untuk menunjukkan proporsi variasi data target yang dapat dijelaskan oleh model. Semakin mendekati nilai satu, semakin baik kemampuan model dalam merepresentasikan hubungan antara variabel *independent* dan variabel *dependent* [1].

### 2.5. Penelitian Terdahulu

Beberapa penelitian terdahulu telah menerapkan regresi linier berganda dalam berbagai kasus prediksi harga. Metode ini terbukti mampu memberikan tingkat kesalahan yang relatif rendah dibandingkan *K-Nearest Neighbor* dan *Support Vector Machine* pada prediksi harga komoditas. Selain itu, regresi linier juga banyak digunakan dalam analisis tren dan peramalan permintaan karena kesederhanaan serta transparansi modelnya [7].

Dalam konteks prediksi harga berlian, penelitian sebelumnya menunjukkan bahwa atribut fisik seperti *carat* dan dimensi geometris memiliki pengaruh

signifikan terhadap harga jual. Pendekatan regresi linier dinilai efektif untuk memodelkan hubungan linier antaratribut numerik tersebut dan menghasilkan prediksi yang cukup akurat. Berdasarkan hasil-hasil penelitian tersebut, regresi linier dipilih dalam penelitian ini sebagai metode utama untuk memprediksi harga berlian menggunakan enam variabel numerik, yaitu *carat*, *depth*, *table*, *x*, *y*, dan *z* [6].

## 3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif karena berfokus pada pengolahan data numerik dan pengukuran hubungan antarvariabel melalui model statistik. Pendekatan yang diterapkan adalah eksperimen komputasional (*computational experiment*), di mana proses analisis dilakukan menggunakan perangkat lunak Python dengan pustaka *Scikit-Learn* untuk membangun dan mengevaluasi model prediksi.

Metode utama yang digunakan adalah Regresi Linier Berganda (*Multiple Linear Regression*), yaitu metode statistik yang digunakan untuk menganalisis pengaruh beberapa variabel *independent* terhadap satu variabel *dependent* [13]. Dalam penelitian ini, variabel *dependent* adalah harga berlian (*price*), sedangkan variabel *independent* mencakup *carat*, *depth*, *table*, *x*, *y*, dan *z*, yang merepresentasikan atribut fisik berlian.

Proses penelitian mengikuti tahapan dalam metodologi *CRISP-DM* (*Cross Industry Standard Process for Data Mining*), yang dipilih karena merupakan standar industri yang fleksibel, sistematis, serta banyak digunakan dalam proyek *Data Mining* [14]. Pendekatan ini membantu menjaga alur analisis tetap terstruktur dan terarah, mulai dari tahap pemahaman bisnis hingga evaluasi hasil model [11]. Adapun enam langkah utama dalam metodologi *CRISP-DM* dijelaskan sebagai berikut:



Gambar 1. Cross Industry Standard Process for Data Mining

### 3.1. Business Understanding

Tahap ini bertujuan untuk memahami tujuan utama analisis, yaitu memprediksi harga berlian berdasarkan karakteristik fisiknya. Model prediksi ini diharapkan dapat membantu dalam memahami faktor-faktor numerik yang memengaruhi harga berlian dan memberikan landasan bagi pengembangan sistem penilaian harga otomatis di masa mendatang.

Metode regresi linier dipilih karena mampu menjelaskan hubungan antara variabel *independent* dan *dependent* secara sederhana, transparan, dan interpretatif. Pendekatan ini sesuai untuk analisis prediktif yang menekankan pada hubungan proporsional antarvariabel numerik tanpa memerlukan kompleksitas komputasi tinggi.

### 3.2. Data Understanding

Dataset diambil dari *Seaborn Dataset Repository*, yaitu kumpulan data publik yang sering digunakan untuk analisis statistik dan pembelajaran mesin. Dataset berlian ini berisi lebih dari 53.000 baris data dengan berbagai atribut. Pengunduhan data dilakukan menggunakan kode berikut:

```
# 1. Unduh dataset
url_reg = "https://raw.githubusercontent.com/mwaskom/seaborn-data/master/diamonds.csv"
data_reg = pd.read_csv(url_reg)
```

Gambar 2. Source Code Mengunduh Dataset

Atribut yang digunakan dalam penelitian ini difokuskan pada data numerik yang secara langsung merepresentasikan karakteristik fisik berlian, yaitu *carat*, *depth*, *table*, *x*, *y*, dan *z*. Pemilihan atribut numerik ini dilakukan karena variabel-variabel tersebut memiliki hubungan kuantitatif dengan harga berlian dan memungkinkan penerapan metode regresi linier untuk mengukur pengaruhnya secara statistik terhadap variabel *dependent*, yaitu *price*.

- *Carat* menggambarkan berat berlian, yang secara umum memiliki korelasi positif dengan harga, di mana semakin besar berat berlian maka nilainya cenderung meningkat.
- *Depth* dan *table* merepresentasikan proporsi potongan berlian yang dapat memengaruhi pantulan cahaya serta kualitas visual, sehingga turut memengaruhi nilai jualnya.
- Sementara itu, *x*, *y*, dan *z* menunjukkan dimensi panjang, lebar, dan kedalaman fisik berlian yang juga menjadi indikator penting dalam menentukan ukuran dan proporsi berlian secara keseluruhan.

```
# 2. Pilih beberapa kolom numerik (data yang akan diolah)
data_reg = data_reg[['carat', 'depth', 'table', 'price', 'x', 'y', 'z']]
```

Gambar 3. Source Code Kolom Numerik

Dengan memilih atribut-atribut numerik ini, penelitian berfokus pada hubungan linier antara faktor-faktor fisik berlian dan harga jualnya. Pendekatan ini juga meminimalkan kompleksitas yang dapat muncul dari variabel kategorikal seperti *cut*, *color*, dan *clarity*,

sehingga memungkinkan model untuk menghasilkan interpretasi yang lebih sederhana namun tetap akurat dalam konteks analisis prediktif berbasis data numerik.

Untuk mempermudah proses analisis dan visualisasi, dilakukan pengambilan sampel sebanyak 50 data acak menggunakan:

```
data_reg = data_reg.sample(50, random_state=42)
```

Gambar 4. Source Code Pengambilan Sample Data

Pemilihan sampel acak ini bertujuan menjaga keragaman data agar tetap representatif terhadap populasi asli, serta mempercepat proses komputasi tanpa mengurangi kualitas representasi data, terutama saat melakukan pelatihan model pada tahap berikutnya [15].

### 3.3. Data Preparation

Pada tahap ini, data disiapkan agar siap digunakan dalam pemodelan. Proses meliputi pemisahan antara variabel *independent* (*features*) dan *dependent* (*target*) serta pembagian data menjadi data latih (80%) dan data uji (20%):

```
# 4. Pisahkan fitur dan target
X = data_reg[['carat', 'depth', 'table', 'x', 'y', 'z']]
y = data_reg['price']

# 5. Split data menjadi training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 5. Source Code Pemisahan Variabel

Tahapan ini penting untuk menghindari *overfitting* dan memastikan model dapat melakukan generalisasi dengan baik terhadap data baru.

### 3.4. Modelling

Pemodelan dilakukan menggunakan algoritma *Linear Regression* dari pustaka *Scikit-Learn*. Model dilatih menggunakan data latih untuk menemukan koefisien regresi terbaik yang menggambarkan hubungan linear antara atribut fisik berlian dan harga.

Secara matematis, regresi linier berganda memodelkan hubungan antara satu variabel *dependent*  $y$  dan beberapa variabel *independent*  $x_1, x_2, \dots, x_n$  dalam bentuk persamaan [14]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon_i$$

dimana  $\beta_0$  merupakan *intercept* (kontanta),  $\beta_i$  adalah koefisien regresi yang menunjukkan besarnya pengaruh setiap variabel *independent* terhadap variabel *dependent*, dan  $\varepsilon$  adalah *error* atau *residual* yang merepresentasikan selisih antara nilai actual dan nilai prediksi.

Nilai koefisien  $\beta$  dapat diperoleh menggunakan pendekatan solusi normal (*normal equation*), yaitu metode matematis untuk mencari parameter regresi yang meminimalkan total galat kuadrat, dengan rumus sebagai berikut:

$$\beta = (X^T X)^{-1} X^T y$$

Dimana  $X$  adalah matriks data *input* (fitur),  $y$  adalah *vector target* (harga berlian), dan  $(X^T X)^{-1} X^T$  berfungsi untuk menentukan nilai koefisien regresi *optional* berdasarkan hubungan linier antara variabel.

Setelah proses pelatihan, model menghasilkan nilai prediksi harga berdasarkan kombinasi linier dari atribut numerik yang digunakan. Hasil prediksi tersebut kemudian dibandingkan dengan data actual untuk mengukur tingkat akurasi model.

```
# 6. Buat model regresi linear
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Gambar 6. Source Code Modelling

Model ini mengestimasi harga berlian berdasarkan kombinasi linier dari enam variabel numerik yang digunakan.

### 3.5. Evaluation

Evaluasi model dilakukan untuk menilai seberapa baik model dalam memprediksi harga berlian. Metrik evaluasi yang digunakan antara lain:

Tabel 1. Metrik Evaluasi

| Metrik Evaluasi           | Rumus                                                                                                       | Deskripsi                                                                    |
|---------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Mean Absolute Error (MAE) | $MAE = \frac{1}{n} \sum_{i=1}^n  y_i - \hat{y}_i $                                                          | Menghitung rata-rata kesalahan absolut.                                      |
| Mean Squared Error (MSE)  | $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$                                                        | Menghitung rata-rata Kuadrat dari selisih prediksi dan nilai aktual.         |
| $R^2$ Score               | $R^2 = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$ | Menyajikan seberapa besar variasi harga berlian dapat dijelaskan oleh model. |

Implementasi perhitungan metrik dilakukan dengan kode berikut:

```
# 7. Evaluasi model
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Gambar 7. Source Code Evaluasi Model

Selain itu, hasil evaluasi juga divisualisasikan dalam bentuk *scatter plot* untuk membandingkan harga aktual dan prediksi.

```
# 8. Visualisasi hasil
plt.figure(figsize=(7,5))
plt.scatter(y_test, y_pred, alpha=0.7, color='royalblue', label='Data Aktual vs Prediksi')

# * Garis linear (ideal y=x)
min_val = min(y_test.min(), y_pred.min())
max_val = max(y_test.max(), y_pred.max())
plt.plot([min_val, max_val], [min_val, max_val], color='red', linewidth=2, label='Garis Linear (y=x)')

plt.xlabel('Actual Price')
plt.ylabel('Predicted Price')
plt.title('Regression: Actual vs Predicted Price (50 Data)')
plt.legend()
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```

Gambar 8. Source Code Visualisasi Hasil

Visualisasi hasil dilakukan menggunakan *scatter plot* antara nilai aktual dan prediksi. Selain itu, ditambahkan garis linear ( $y=x$ ) sebagai garis referensi ideal untuk menunjukkan tingkat kedekatan antara hasil prediksi model dan nilai sebenarnya.

### 3.6. Deployment

Karena penelitian ini bersifat eksperimen komputasional, tahap implementasi sistem tidak dilakukan. Namun demikian, hasil penelitian ini dapat dijadikan landasan pengembangan aplikasi prediksi harga berlian berbasis pembelajaran mesin pada tahap penelitian selanjutnya.

## 4. HASIL DAN PEMBAHASAN

### 4.1. Hasil Pengolahan Data

Dari dataset awal, 50 baris data berlian dengan atribut numerik yaitu *carat*, *depth*, *table*, *price*,  $x$ ,  $y$ , dan  $z$ . Contoh data yang diolah ditunjukkan pada Gambar 9.

|       | carat | depth | table | price | x    | y    | z    |
|-------|-------|-------|-------|-------|------|------|------|
| 1388  | 0.24  | 62.1  | 56.0  | 559   | 3.97 | 4.00 | 2.47 |
| 50052 | 0.58  | 60.0  | 57.0  | 2201  | 5.44 | 5.42 | 3.26 |
| 41645 | 0.40  | 62.1  | 55.0  | 1238  | 4.76 | 4.74 | 2.95 |
| 42377 | 0.43  | 60.8  | 57.0  | 1304  | 4.92 | 4.89 | 2.98 |
| 17244 | 1.55  | 62.3  | 55.0  | 6901  | 7.44 | 7.37 | 4.61 |
| 1608  | 1.00  | 55.4  | 62.0  | 3011  | 6.63 | 6.59 | 3.66 |
| 46398 | 0.51  | 60.2  | 56.0  | 1765  | 5.22 | 5.24 | 3.15 |
| 45493 | 0.52  | 62.0  | 56.0  | 1679  | 5.17 | 5.19 | 3.21 |
| 49385 | 0.62  | 60.0  | 59.0  | 2102  | 5.58 | 5.56 | 3.34 |
| 10460 | 1.14  | 60.3  | 57.0  | 4789  | 6.79 | 6.85 | 4.11 |
| 34243 | 0.40  | 62.8  | 56.0  | 855   | 4.73 | 4.70 | 2.96 |
| 24772 | 1.83  | 62.8  | 56.0  | 13134 | 7.76 | 7.82 | 4.89 |
| 42987 | 0.60  | 55.3  | 63.0  | 1367  | 5.67 | 5.61 | 3.12 |
| 51711 | 0.70  | 63.5  | 56.0  | 2401  | 5.58 | 5.66 | 3.57 |
| 43777 | 0.53  | 61.2  | 65.0  | 1440  | 5.16 | 5.27 | 3.19 |
| 22232 | 1.55  | 61.2  | 55.0  | 10286 | 7.49 | 7.47 | 4.58 |
| 1319  | 0.98  | 61.6  | 66.0  | 2958  | 6.46 | 6.24 | 3.92 |
| 43219 | 0.40  | 62.6  | 56.0  | 1397  | 4.73 | 4.70 | 2.95 |
| 21843 | 1.50  | 62.8  | 56.0  | 9904  | 7.26 | 7.33 | 4.58 |
| 53626 | 0.35  | 60.7  | 62.0  | 552   | 4.53 | 4.59 | 2.77 |
| 27497 | 2.07  | 61.2  | 58.0  | 18193 | 8.18 | 8.20 | 5.01 |
| 36554 | 0.40  | 60.9  | 60.0  | 945   | 4.77 | 4.75 | 2.90 |
| 21055 | 0.31  | 61.9  | 55.0  | 625   | 4.31 | 4.35 | 2.68 |
| 23028 | 0.30  | 61.7  | 54.2  | 631   | 4.32 | 4.35 | 2.67 |
| 5536  | 0.91  | 60.6  | 57.0  | 3854  | 6.30 | 6.26 | 3.81 |
| 43434 | 0.53  | 61.4  | 56.0  | 1415  | 5.21 | 5.24 | 3.21 |
| 38522 | 0.34  | 60.8  | 57.0  | 1033  | 4.53 | 4.49 | 2.74 |
| 24964 | 2.02  | 62.7  | 57.0  | 13453 | 8.07 | 7.93 | 5.03 |
| 19771 | 1.04  | 59.1  | 56.0  | 8341  | 6.65 | 6.59 | 3.91 |
| 40271 | 0.23  | 61.8  | 55.0  | 493   | 3.93 | 3.96 | 2.43 |
| 8473  | 1.01  | 62.5  | 55.0  | 4416  | 6.39 | 6.40 | 4.00 |
| 43496 | 0.54  | 61.4  | 60.0  | 1422  | 5.21 | 5.25 | 3.21 |
| 23558 | 2.01  | 63.4  | 57.0  | 11566 | 7.97 | 7.89 | 5.03 |
| 46750 | 0.51  | 63.3  | 58.0  | 1800  | 5.04 | 5.07 | 3.20 |
| 49727 | 0.50  | 62.8  | 55.0  | 2153  | 5.00 | 5.13 | 3.18 |
| 34922 | 0.30  | 62.6  | 53.0  | 373   | 4.28 | 4.32 | 2.69 |
| 31849 | 0.30  | 60.9  | 56.0  | 776   | 4.35 | 4.32 | 2.64 |
| 47712 | 0.50  | 63.3  | 53.0  | 1885  | 5.06 | 5.09 | 3.21 |
| 48446 | 0.51  | 63.1  | 58.0  | 1974  | 5.11 | 5.07 | 3.21 |

Gambar 9. Dataset

Dataset ini menjadi dasar untuk pelatihan model regresi linear guna memprediksi harga berdasarkan karakteristik fisik berlian.

Tabel 2. Statistik Deskriptif Data Berlian

| Var.  | Min. | Max.  | Avg.   | Standar Deviasi |
|-------|------|-------|--------|-----------------|
| Carat | 0.23 | 2.52  | 0.98   | 0.64            |
| Depth | 55.3 | 63.7  | 61.5   | 1.9             |
| Table | 53.0 | 66.0  | 57.1   | 3.1             |
| Price | 373  | 18443 | 5912.4 | 5230.5          |
| x     | 3.93 | 8.57  | 6.1    | 1.4             |
| y     | 3.96 | 8.52  | 6.08   | 1.45            |
| z     | 2.43 | 5.44  | 3.73   | 0.88            |

Tabel 2. menyajikan ringkasan statistik dari 50 data berlian yang digunakan dalam penelitian ini. Nilai-nilai pada tabel tersebut memberikan gambaran umum mengenai sebaran data secara keseluruhan sebagaimana ditampilkan sebagian pada Gambar 9.

#### 4.2. Hasil Regresi Linear

Setelah dilakukan proses pelatihan model menggunakan algoritma *Linear Regression*, diperoleh hasil evaluasi sebagai berikut:

```
===== HASIL REGRESI LINEAR =====
Mean Absolute Error (MAE): 1200.69
Mean Squared Error (MSE): 2582825.77
R2 Score (Akurasi Model): 0.944
```

Gambar 10. Hasil Regresi Linear

*Mean Absolute Error (MAE)* sebesar 1.200,69 menunjukkan bahwa secara rata-rata, prediksi model menyimpang sekitar 1.200,69 satuan harga dari nilai sebenarnya. *Mean Squared Error (MSE)* sebesar 2.582.825,77 menghitung rata-rata kuadrat selisih antara prediksi dan nilai aktual, sehingga kesalahan besar akan lebih berpengaruh dan metrik ini sensitif terhadap outlier.  $R^2$  sebesar 0,944 menunjukkan bahwa sekitar 94,4% variasi harga berlian dapat dijelaskan oleh atribut fisik yang digunakan dalam model, menandakan bahwa model *Linear Regression* yang dibangun cukup baik dalam merepresentasikan data.

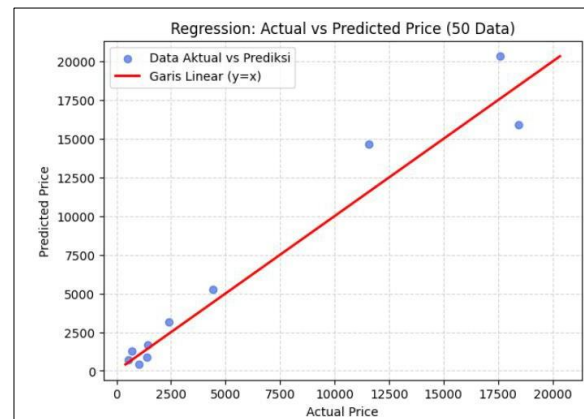
Dengan kombinasi  $R^2$  tinggi serta nilai *MAE* dan *MSE* yang masih dalam batas wajar, dapat disimpulkan bahwa model memiliki performa prediksi yang baik, meskipun masih terdapat beberapa prediksi yang kurang tepat.

#### 4.3. Visualisasi Hasil Prediksi

Hubungan antara nilai aktual dan nilai prediksi divisualisasikan pada Gambar 11 berikut, titik biru menunjukkan perbandingan nilai aktual terhadap nilai prediksi untuk setiap data, sedangkan garis merah menggambarkan garis ideal  $y = x$  yaitu kondisi ketika hasil prediksi sama dengan nilai sebenarnya.

Sebagian besar titik biru terlihat berada di sekitar garis merah, menandakan bahwa prediksi model mendekati nilai aktual. Namun, terdapat beberapa titik

yang menyimpang jauh dari garis, terutama pada area harga tinggi. Penyimpangan ini menandakan bahwa model tidak sepenuhnya mampu menangkap pola hubungan pada data ekstrem atau adanya variabel lain yang turut memengaruhi harga tetapi belum dimasukkan dalam model. Kondisi ini umum terjadi pada regresi linear karena asumsi hubungan antarvariabel bersifat linier, sementara kenyataan data bisa bersifat non-linear.



Gambar 11. Hasil Prediksi

Sebaran ini secara umum menunjukkan bahwa model regresi linear mampu menggambarkan pola utama hubungan antara variabel *input* dan harga berlian, meskipun masih terdapat ketidakakuratan pada nilai ekstrem.

#### 4.4. Pembahasan

Berdasarkan hasil evaluasi, nilai  $R^2$  sebesar 0.944 menunjukkan bahwa model regresi linear memiliki kemampuan prediksi yang sangat baik. Nilai ini berarti bahwa sekitar 94,4% variasi harga berlian dapat dijelaskan oleh atribut numerik seperti *carat*, *depth*, *table*, *x*, *y*, dan *z*. Semakin mendekati 1 nilai  $R^2$ , semakin baik kemampuan model dalam merepresentasikan hubungan linear antara variabel *input* dan *output* [15].

Sementara itu, nilai *MAE* sebesar 1,200.69 dan *MSE* sebesar 2,582,825.77 menunjukkan bahwa rata-rata kesalahan prediksi relatif kecil dibandingkan dengan rentang harga aktual yang mencapai puluhan ribu dolar. Hal ini menandakan bahwa model cukup efektif dalam mengestimasi harga berlian berdasarkan karakteristik fisik yang tersedia.

Visualisasi hasil pada Gambar 10 mendukung temuan ini, di mana sebagian besar data aktual dan prediksi menunjukkan kecenderungan yang searah. Pola linear yang muncul memperkuat bahwa hubungan antara atribut fisik dan harga bersifat hampir proporsional, khususnya pada variabel *carat* yang paling berpengaruh terhadap harga berlian.

Namun, penyimpangan pada beberapa titik, khususnya pada rentang harga tinggi, dapat disebabkan oleh faktor-faktor eksternal yang tidak disertakan dalam model, seperti *color* (warna), *clarity* (kejernihan), dan *cut quality* (kualitas potongan).

Atribut-atribut tersebut bersifat kategorikal dan memiliki dampak signifikan terhadap nilai jual berlian, sehingga ketidakhadirannya dapat menurunkan presisi prediksi.

Secara keseluruhan, model Regresi Linear berbasis atribut numerik ini sudah mampu memberikan estimasi harga dengan tingkat akurasi tinggi. Untuk pengembangan lebih lanjut, model dapat diperluas dengan menambahkan fitur kategorikal serta menerapkan metode regresi yang lebih kompleks seperti *Polynomial Regression* atau *Random Forest Regressor* untuk menangkap pola non-linear yang mungkin terdapat pada data.

## 5. KESIMPULAN

Berdasarkan hasil analisis dan pemodelan menggunakan algoritma *Linear Regression*, dapat disimpulkan bahwa model mampu memprediksi harga berlian dengan tingkat akurasi yang tinggi. Nilai  $R^2$  Score sebesar 0.944 menunjukkan bahwa model dapat menjelaskan sekitar 94,4% variasi harga berlian berdasarkan enam atribut numerik utama, yaitu *carat*, *depth*, *table*, *x*, *y*, dan *z*. Nilai *Mean Absolute Error (MAE)* sebesar 1,200.69 dan *Mean Squared Error (MSE)* sebesar 2,582,825.77 mengindikasikan bahwa rata-rata selisih antara harga aktual dan hasil prediksi masih dalam batas wajar terhadap rentang harga berlian yang besar.

Hasil tersebut membuktikan bahwa regresi linear merupakan metode yang efektif, sederhana, dan

mudah diinterpretasikan dalam memahami hubungan antara karakteristik fisik berlian dan nilai jualnya. Variabel *carat* terbukti memiliki pengaruh paling dominan terhadap harga, diikuti oleh dimensi fisik (*x*, *y*, *z*), sedangkan *depth* dan *table* memberikan kontribusi yang lebih kecil terhadap variasi harga. Temuan ini konsisten dengan karakteristik pasar berlian yang menempatkan berat dan ukuran sebagai faktor utama dalam penentuan nilai jual.

Secara umum, penelitian ini menunjukkan bahwa atribut numerik dapat dijadikan dasar yang kuat untuk membangun sistem prediksi harga berlian secara otomatis. Namun, terdapat keterbatasan pada model yang hanya menggunakan variabel numerik, tanpa mempertimbangkan faktor kategorikal seperti *color*, *clarity*, dan *cut quality* yang juga berpengaruh signifikan terhadap harga. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan atribut kategorikal serta menerapkan algoritma non-linear seperti *Polynomial Regression*, *Random Forest*, atau *Gradient Boosting* guna meningkatkan akurasi dan kemampuan generalisasi model.

Dengan demikian, hasil penelitian ini dapat menjadi landasan awal bagi pengembangan sistem penilaian harga berlian berbasis data yang dapat dimanfaatkan oleh industri perhiasan maupun *platform* perdagangan daring (*e-commerce*) dalam menilai dan menentukan harga berlian secara objektif dan efisien.

## DAFTAR PUSTAKA

- [1] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting : methods , applications , and research opportunities," *J. Big Data*, 2020, doi: 10.1186/s40537-020-00329-2.
- [2] M. Muharrom, "Analisis Komparasi Algoritma Data Mining Naive Bayes, K-Nearest Neighbors dan Regresi Linier Dalam Prediksi Harga Emas," *Bull. Inf. Technol.*, vol. 4, no. 4, pp. 430–438, 2023, doi: 10.47065/bit.v4i4.986.
- [3] E. Daniati and A. Ristiyawan, "4924-Article Text-18566-1-10-20240722," vol. 8, pp. 173–183, 2024.
- [4] H. Jahani, R. Jain, and D. Ivanov, "Data science and big data analytics : a systematic review of methodologies used in the supply chain and logistics research," *Ann. Oper. Res.*, 2023, doi: 10.1007/s10479-023-05390-7.
- [5] M. Sulistiyono, A. Sidauruk, B. Satria, R. Wardhana, and U. A. Yogyakarta, "RAINFALL PREDICTION USING MULTIPLE LINEAR REGRESSION," vol. 9, no. 1, pp. 17–22, 2023, doi: 10.33480/jitk.v9i1.4203.RAINFALL.
- [6] J. Riset *et al.*, "Multi Regresi Linear Untuk Prediksi Kepadatan Penduduk Berdasarkan Variable Kependudukan Di Indonesia," vol. 1, no. 2, pp. 65–69, 2024.
- [7] M. S. Sánchez, "Optimization of diamond price prediction strategies using machine learning techniques," vol. 16, no. September, pp. 28–38, 2024, doi: 10.5897/JEIF2024.1223.
- [8] "Dasar Machine Learning\_v.3.0\_FULL ISBN.pdf."
- [9] J. Matematika, A. Pasaribu, and M. R. Syahputra, "Comparison of Fuzzy Logic and Multiple Linear Regression in Forecasting Rice Production in Toba District Mathematics Undergraduate Study Program , Faculty of Mathematics and Natural," vol. 5, no. c, pp. 121–127, 2022.
- [10] M. A. Suardana and V. Tundjungsari, "Prediction of Waste Volume Based on Annual Trends Using Data Mining," vol. 1, no. 2, pp. 49–55, 2025.
- [11] E. P. A. Akhmad, "Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran," *J. Apl. Pelayaran Dan Kepelabuhanan*, vol. 10, no. 2, pp. 120–131, 2020, doi: 10.30649/japk.v10i2.13.
- [12] Z. Ouyang, "Research on the Diamond Price Prediction based on Linear Regression , Decision Tree and Random Forest," vol. 24, pp. 248–257, 2024.
- [13] P. Pendapatan, "Implementasi CRISP-DM pada Data Mining untuk Melakukan," vol. 30, no. 1, pp. 111–121, 2024, doi: 10.36309/goi.v30i1.266.
- [14] R. Riyandi, R. Roy Hakiki, Dealmus, Y. O. Reins Dima, and N. P., "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Metode Regresi Linear Berganda Berbasis Shiny R,"

- Inform. J. Ilmu Komput.*, vol. 21, no. 1, pp. 1–13, 2025, doi: 10.52958/iftk.v21i1.10563.
- [15] A. K. Khotimah, A. A. Rahman, M. Z. Alam, Y. H. Nur, and T. R. Aufo, “Analisis Regresi Linier Berganda Dalam Estimasi Indeks Pembangunan Manusia di Indonesia Multiple Linear Regression Analysis In Estimating The Human Development Index In Indonesia,” vol. 15, no. November, pp. 90–99, 2024, doi: 10.30872/eksponensial.v15i2.1318.