

ANALISIS KOMPARATIF STACKED REGRESSION BERBASIS RANDOM FOREST, XGBOOST, DAN LIGHTGBM DENGAN REGRESI LINEAR BERGANDA PADA PREDIKSI HARGA BERLIAN

COMPARATIVE ANALYSIS OF STACKED REGRESSION BASED ON RANDOM FOREST, XGBOOST, AND LIGHTGBM WITH MULTIPLE LINEAR REGRESSION FOR DIAMOND PRICE PREDICTION

Nur Aulina^{1†}, Diani Raras Puspita², Anna Maria Restho
Yulianuti³, Irsyad Faris Sena Pratama⁴, and Septian
Rahardiantoro⁵

^{1,2,3,4,5}Department of Statistics, IPB University, Indonesia

[†]corresponding author: nuraulinanur@apps.ipb.ac.id

Copyright © 2026 Nur Aulina, Diani Raras Puspita, Anna Maria Restho Yulianuti, Irsyad Faris Sena Pratama and Septian Rahardiantoro. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Diamond price prediction is a challenging regression problem due to the influence of multiple physical and quality attributes, including the 4C characteristics (*carat*, *cut*, *color*, and *clarity*) as well as dimensional variables. This study compares the performance of Multiple Linear Regression with Box–Cox transformation and Stacked Regression for diamond price prediction using the *Diamonds Dataset* obtained from Kaggle. After preprocessing, 53,920 observations were used for modeling. The preprocessing stage included removing invalid observations, categorical variable encoding, and Box–Cox transformation on the target variable to reduce skewness and stabilize variance. The dataset was divided into *training* and *testing* sets using an 80:20 stratified split. Stacked Regression was constructed using Random Forest, XGBoost, and LightGBM as base learners, while Linear Regression was used as the meta-learner. Model performance was evaluated using MAE, RMSE, MAPE, and R^2 . The results showed that Stacked Regression achieved the best performance with MAE of 266.70, RMSE of USD 552.67, MAPE of 6.13%, and R^2 of 0.9809.

Keywords: Box–Cox Transformation, Diamond Price Prediction, Ensemble Learning, LightGBM, Stacked Regression

Abstrak

Prediksi harga berlian merupakan permasalahan regresi yang kompleks karena dipengaruhi oleh karakteristik fisik dan kualitas berlian termasuk atribut 4C (*carat*, *cut*, *color*, dan *clarity*) serta dimensi fisik berlian. Penelitian ini membandingkan performa Regresi Linear Berganda dengan transformasi Box-Cox dan Stacked Regression dalam memprediksi harga berlian menggunakan *Diamonds Dataset* dari Kaggle. Setelah tahap *preprocessing*, sebanyak 53.920 amatan digunakan dalam pemodelan. Tahapan *preprocessing* meliputi penghapusan data tidak valid, *encoding* peubah kategorik, dan transformasi Box-Cox pada peubah respon. Dataset dibagi menjadi data *training* dan *testing* menggunakan *stratified split* 80:20. Model Stacked Regression menggunakan Random Forest, XGBoost, dan LightGBM sebagai *base learner*, serta Regresi Linear sebagai *meta learner*. Evaluasi dilakukan menggunakan MAE, RMSE, MAPE, dan R^2 . Hasil penelitian menunjukkan bahwa Stacked Regression menghasilkan performa terbaik dengan MAE sebesar 266,70, RMSE USD 552,67, MAPE 6,13%, dan R^2 sebesar 0,9809.

Kata Kunci: Ensemble Learning, LightGBM, Prediksi Harga Berlian, Stacked Regression, Transformasi Box–Cox

1. Pendahuluan

Industri perhiasan dan investasi berlian memerlukan prediksi harga yang akurat untuk mendukung valuasi aset dan pengambilan keputusan. Harga berlian dipengaruhi oleh karakteristik 4C (*carat*, *color*, *clarity*, dan *cut*) serta karakteristik fisik lain yang membentuk hubungan kompleks dan nonlinier (Kigo *et al.*, 2023). Regresi Linear Berganda (RLB) banyak digunakan karena sederhana dan mudah diinterpretasikan (Montgomery *et al.*, 2021). Namun, data harga berlian umumnya mengalami right-skewed, heteroskedastisitas, dan korelasi tinggi antarprediktor sehingga asumsi klasik sulit terpenuhi dan performa model menurun (James *et al.*, 2021).

Transformasi Box-Cox merupakan salah satu pendekatan yang umum digunakan untuk mengatasi pelanggaran asumsi pada model regresi karena mampu memperbaiki distribusi data dan menstabilkan ragam (Box dan Cox, 1964). Meskipun efektif dalam memperbaiki karakteristik distribusi data, transformasi ini hanya bekerja pada peubah respon dan belum mampu menangkap hubungan nonlinier yang kompleks antarpubah. Oleh karena itu, berbagai penelitian menunjukkan bahwa metode *machine learning*, khususnya *ensemble learning*, mampu memberikan akurasi prediksi yang lebih baik dibandingkan regresi klasik (Kigo *et al.*, 2023).

Stacked Regression merupakan salah satu teknik *ensemble learning* yang dirancang untuk menggabungkan prediksi dari beberapa model melalui *meta learner* guna meningkatkan akurasi dan kemampuan

generalisasi model (Wolpert, 1992). Dibandingkan model tunggal seperti Random Forest, XGBoost, atau LightGBM, Stacked Regression dapat menggabungkan informasi dari beberapa algoritma yang berbeda sehingga memanfaatkan keunggulan masing-masing *base learner* secara adaptif. Kemampuan tersebut membuat Stacked Regression berpotensi menghasilkan prediksi yang lebih stabil dan akurat pada data dengan pola hubungan yang kompleks (Breiman, 1996).

Penelitian yang membandingkan RLB dengan transformasi Box-Cox dan Stacked Regression secara setara masih relatif terbatas. Perbedaan *preprocessing*, transformasi target, dan pembagian data pada penelitian sebelumnya menyebabkan perbandingan performa yang kurang objektif. Oleh karena itu, penelitian ini menerapkan skema evaluasi yang seragam pada seluruh model agar perbandingan dapat dilakukan secara lebih adil. Penelitian ini bertujuan mengimplementasikan dan mengevaluasi model RLB dengan transformasi Box-Cox serta Stacked Regression dengan *base learner* berupa Random Forest, XGBoost, dan LightGBM, yang performanya dibandingkan menggunakan metrik MAE, RMSE, MAPE, dan R^2 .

2. Metodologi

2.1 Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh melalui platform penyedia dataset terbuka yakni Kaggle dengan menggunakan *Diamonds Dataset* (<https://www.kaggle.com/datasets/shivam2503/diamonds>). Dataset ini terdiri atas 53.940 amatan dengan sepuluh peubah yaitu satu peubah respon dan sembilan peubah penjelas yang menggambarkan karakteristik fisik dan kualitas berlian. Peubah yang digunakan dalam penelitian disajikan pada Tabel 1.

Tabel 1: Peubah yang digunakan.

Kode	Nama Peubah	Keterangan
Y	<i>Price</i>	Harga berlian dalam satuan dolar Amerika Serikat (USD)
X1	<i>Carat</i>	Berat berlian dalam satuan karat
X2	<i>Cut</i>	Kualitas potongan berlian (Fair, Good, Very Good, Premium, Ideal)
X3	<i>Color</i>	Warna berlian dengan skala J (terburuk) hingga D (terbaik)
X4	<i>Clarity</i>	Tingkat kejernihan berlian
X5	<i>Depth</i>	Persentase total kedalaman berlian (%)
X6	<i>Table</i>	Lebar bagian atas berlian relatif terhadap titik terlebar (%)
X7	x	Panjang berlian (mm)
X8	y	Lebar berlian (mm)
X9	z	Kedalaman berlian (mm)

2.2 Metode Penelitian

1. Melakukan *preprocessing* data untuk mempersiapkan data dan memastikan kualitas data sebelum proses pemodelan. Tahapan ini diterapkan secara konsisten pada seluruh model yang meliputi penanganan data tidak valid, *encoding* peubah kategorik, serta transformasi Box-Cox pada peubah respon. Transformasi Box-Cox diterapkan pada peubah respon *price* dan dinyatakan sebagai:

$$y^{(\lambda)} = \frac{y^{(\lambda)} - 1}{\lambda}, \lambda \neq 1 \quad (1)$$

dengan y merupakan nilai *price* asli dan λ merupakan parameter transformasi yang ditentukan secara empiris menggunakan prosedur *maximum likelihood* pada data *training*. Prediksi akhir kemudian dikembalikan ke satuan USD menggunakan invers transformasi Box-Cox sebagai berikut:

$$\hat{y} = (\hat{y}_{BC} \times \lambda + 1)^{1/\lambda} \quad (2)$$

2. Melakukan eksplorasi data untuk memahami karakteristik dataset sebelum proses pemodelan. Tahap ini meliputi analisis statistik deskriptif, identifikasi *missing value*, visualisasi distribusi data menggunakan histogram dan *boxplot*, analisis hubungan antarpeubah menggunakan matriks korelasi, serta deteksi *outlier* menggunakan metode *Interquartile Range* (IQR).
3. Melakukan pembagian data menjadi 80% data *training* dan 20% data *testing* menggunakan *stratified sampling* berdasarkan kuantil peubah respon *price* dengan *random seed* sebesar 42 untuk menjaga representativitas distribusi harga pada kedua subset. Validasi model dilakukan menggunakan *5-Fold Cross Validation* sedangkan *tuning* hiperparameter dilakukan menggunakan *Grid Search Cross Validation* dengan konfigurasi terbaik dipilih berdasarkan nilai CV-RMSE terkecil (Basha & Oveis, 2024). Kombinasi hiperparameter yang diuji disajikan pada Tabel 2.

Tabel 2: Kombinasi hiperparameter yang diuji pada setiap model.

Model	Hiperparameter	Nilai yang Diuji
<i>Random Forest</i>	n_estimators	100, 200, 300
	max_depth	None, 10, 20
	min_samples_split	2, 5, 10
<i>XGBoost</i>	n_estimators	100, 200, 300
	learning_rate	0,01; 0,05; 0,1
	max_depth	3, 5, 7
<i>LightGBM</i>	n_estimators	100, 200, 300
	learning_rate	0,01; 0,05; 0,1
	num_leaves	31, 50, 70

Kombinasi hiperparameter dipilih berdasarkan parameter yang umum digunakan pada penelitian berbasis *ensemble learning* dan disesuaikan dengan karakteristik dataset berlian.

4. Melakukan pemodelan menggunakan Regresi Linear dengan transformasi Box-Cox sebagai model klasik pembandingan (*baseline model*) untuk mengukur peningkatan performa yang dihasilkan oleh pendekatan *machine learning* dibandingkan metode konvensional (Randhawa *et al.*, 2025). Estimasi parameter dilakukan menggunakan metode *Ordinary Least Squares* (OLS) pada skala Box-Cox dengan persamaan:

$$BC(\widehat{price}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_9 X_9 + \varepsilon \quad (3)$$

di mana $BC(\widehat{price})$ merupakan nilai prediksi *price* setelah transformasi Box-Cox, β_0 adalah *intercept*, β_1 hingga β_9 merupakan koefisien regresi masing-masing peubah penjelas, dan ε merupakan galat model. Nilai λ ditentukan dari data *training* menggunakan fungsi ``boxcox()`` pada *package* MASS di R.

5. Melakukan pemodelan menggunakan Stacked Regression yaitu teknik *ensemble learning* yang menggabungkan prediksi beberapa model dasar (*base learner*) melalui model akhir (*meta learner*) yang mempelajari bobot optimal dari setiap prediksi *base learner* (Breiman, 1996). Arsitektur Stacked Regression dalam penelitian ini terdiri atas dua level.

- a. Level 0 (*Base Learner*)

Pada level ini digunakan tiga algoritma, yaitu Random Forest, XGBoost, dan LightGBM yang dilatih secara terpisah menggunakan data *training* dengan hiperparameter terbaik hasil *tuning*. Prediksi dari masing-masing model dihasilkan melalui *internal cross-validation* ($cv = 5$, *passthrough=False*) untuk menghindari *data leakage*. Random Forest merupakan metode *ensemble* berbasis *bagging* dan pemilihan fitur acak yang mampu menangani hubungan nonlinear tanpa asumsi distribusi tertentu (Kigo *et al.*, 2023). XGBoost merupakan implementasi *gradient boosting* sekuensial dengan regularisasi L1 dan L2 untuk mengurangi risiko *overfitting* (Basha & eis, 2024). Sementara itu, LightGBM merupakan varian *gradient boosting* berbasis *leaf-wise growth* dengan teknik GOSS dan EFB yang dirancang untuk meningkatkan efisiensi komputasi tanpa mengurangi akurasi model (Ke *et al.*, 2017).

- b. Pembentukan *Meta-Feature Dataset*

Prediksi dari ketiga *base learner* pada Level 0 kemudian dikombinasikan menjadi dataset baru yang terdiri atas tiga kolom, yaitu prediksi dari Random Forest, XGBoost, dan LightGBM. Dataset ini digunakan sebagai masukan pada Level 1. Tahap ini

bertujuan membentuk representasi baru yang menggabungkan informasi prediksi dari seluruh *base learner* sehingga *meta learner* dapat mempelajari kombinasi prediksi yang optimal.

c. Level 1 (*Meta Learner*)

Regresi Linear digunakan sebagai *meta learner* untuk mempelajari bobot optimal dari prediksi *base learner* karena interpretabilitasnya yang baik. Persamaan *meta learner* pada skala Box-Cox:

$$\hat{Y}_{BC} = \beta_0 + \beta_1 \hat{Y}_{RF} + \beta_2 \hat{Y}_{XGB} + \beta_3 \hat{Y}_{LGBM} \quad (4)$$

dengan $\hat{Y}_{BC}$ merupakan prediksi akhir pada skala Box-Cox, sedangkan $\hat{Y}_{RF}$, $\hat{Y}_{XGB}$, dan $\hat{Y}_{LGBM}$ masing-masing merupakan prediksi dari Random Forest, XGBoost, dan LightGBM.

Prediksi akhir kemudian dikembalikan ke satuan USD menggunakan invers transformasi Box-Cox agar hasil prediksi dapat diinterpretasikan pada skala harga asli.

6. Melakukan evaluasi performa model menggunakan metrik MAE, RMSE, MAPE, dan R^2 yang dihitung pada skala harga asli (USD). Metrik evaluasi yang digunakan dinyatakan sebagai:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (7)$$

$$R^2 = 1 - \frac{RSS}{TSS} \quad (8)$$

dengan y_i merupakan nilai aktual, $\hat{y}_i$ merupakan nilai prediksi, n merupakan jumlah amatan, RSS merupakan *Residual Sum of Squares*, TSS merupakan *Total Sum of Squares*, dan k merupakan jumlah peubah prediktor.

Performa model dikatakan lebih baik jika nilai MAPE, RMSE, dan MAE semakin kecil, sedangkan nilai R^2 semakin besar. Untuk mengevaluasi stabilitas dan kemampuan generalisasi pada data *training*, digunakan nilai CV-RMSE dengan metode *5-Fold Cross Validation*. Indikasi *overfitting* diperiksa dengan membandingkan performa model pada data *training* dan *testing*.

7. Melakukan interpretasi model terbaik untuk memahami kontribusi peubah dan perilaku model dalam menghasilkan prediksi. Interpretasi

dilakukan melalui analisis *feature importance*, koefisien *meta learner*, visualisasi *actual vs predicted*, analisis residual, serta pemeriksaan *overfitting* berdasarkan perbandingan performa pada data *training* dan *testing*.

3. Hasil dan Pembahasan

3.1 Preprocessing Data

Pada tahap *preprocessing* ditemukan 20 amatan dengan nilai dimensi fisik x , y , atau z sama dengan nol sehingga dihapus dari dataset karena tidak merepresentasikan kondisi fisik berlian yang valid sehingga amatan yang digunakan dalam analisis berjumlah 53.920. Selanjutnya, peubah kategorik *cut*, *color*, dan *clarity* dikonversi menjadi variabel *dummy* menggunakan *One-Hot Encoding* sehingga total peubah yang digunakan dalam pemodelan menjadi 23 peubah.

Transformasi Box-Cox diterapkan pada peubah respon *price* untuk mengurangi distribusi *right-skewed* dan menstabilkan ragam data. Nilai parameter λ yang diperoleh pada model Regresi Linear menggunakan R adalah $-0,0202$, sedangkan pada model *machine learning* menggunakan Python adalah $-0,0671$. Kedua nilai tersebut mendekati nol sehingga transformasi optimal cenderung menyerupai transformasi logaritmik dan mampu memperbaiki distribusi peubah *price*.

3.2 Eksplorasi Data

Eksplorasi data dilakukan terhadap 53.920 amatan yang terdiri atas tujuh peubah numerik dan tiga peubah kategorik. Peubah *price* memiliki distribusi *right-skewed* dengan rata-rata sebesar USD 3.932,80 dan rentang harga USD 326–18.823. Pada peubah kategorik, kategori yang dominan adalah *cut* = Ideal, *color* = G, dan *clarity* = SI1. Hasil identifikasi menunjukkan tidak terdapat *missing value* pada dataset.

Visualisasi histogram menunjukkan distribusi *price* yang sangat menjulur ke kanan sedangkan transformasi Box-Cox menghasilkan distribusi yang lebih simetris. *Heatmap* korelasi menunjukkan bahwa peubah *carat*, x , y , dan z memiliki korelasi positif yang kuat terhadap *price*. Deteksi *outlier* menggunakan metode IQR menunjukkan bahwa peubah *price* memiliki jumlah *outlier* terbesar, namun *outlier* tidak dihapus karena masih merepresentasikan variasi alami harga berlian premium.

3.3 Pembagian Data dan Validasi Model

Dataset dibagi menjadi 80% data *training* sebanyak 43.136 amatan dan 20% data *testing* sebanyak 10.784 amatan menggunakan *stratified sampling* berdasarkan kuantil *price* dengan *random seed* sebesar 42.

Proses stratifikasi menghasilkan distribusi harga yang relatif seimbang pada kedua subset sehingga dapat mengurangi potensi bias dalam evaluasi model. Konfigurasi terbaik hasil tuning hiperparameter menggunakan *Grid Search Cross Validation* dengan *5-Fold Cross Validation* disajikan pada Tabel 3.

Tabel 3: Konfigurasi tuning hiperparameter *base learner*.

Model	Hiperparameter Terbaik	CV RMSE
<i>Random Forest</i>	n_estimators=300, max_depth=None, min_samples_split=2	0,0612
<i>XGBoost</i>	n_estimators=300, learning_rate=0,1, max_depth=7	0,0527
<i>LightGBM</i>	n_estimators=300, learning_rate=0,1, num_leaves=70	0,0517

Berdasarkan hasil validasi silang, LightGBM menghasilkan nilai CV RMSE terkecil sebesar 0,0517 diikuti XGBoost sebesar 0,0527 dan Random Forest sebesar 0,0612. Hasil ini menunjukkan bahwa LightGBM memiliki kemampuan generalisasi terbaik selama proses validasi model.

3.4 Hasil Model Regresi Linear

Model Regresi Linear dibangun menggunakan 23 peubah hasil *encoding* dan menghasilkan nilai R^2 sebesar 0,9812 pada skala Box-Cox yang menunjukkan bahwa model mampu menjelaskan sekitar 98,12% variasi data pada skala transformasi. Hasil ini mengindikasikan adanya hubungan yang kuat antara karakteristik fisik dan kualitas berlian terhadap harga.

Hasil evaluasi model menunjukkan nilai CV-RMSE sebesar 0,1399 pada skala Box-Cox. Pada data *testing*, model menghasilkan MAE sebesar USD 604,51, RMSE sebesar USD 865,23, MAPE sebesar 10,73%, dan R^2 sebesar 0,9538. Meskipun performa model tergolong baik, hasil tersebut menunjukkan bahwa pendekatan *machine learning* masih berpotensi memberikan peningkatan akurasi yang mengindikasikan adanya pola nonlinier pada hubungan antara karakteristik berlian dan harga yang belum sepenuhnya ditangkap oleh model linear.

3.5 Hasil Model *Machine Learning*

3.5.1 Level 0 (*Base Learner*)

Pada Level 0 digunakan tiga algoritma *machine learning* yaitu Random Forest, XGBoost, dan LightGBM yang dilatih secara terpisah menggunakan hiperparameter terbaik hasil *Grid Search Cross*

Validation. Random Forest menghasilkan MAE sebesar USD 295,76, RMSE sebesar USD 636,13, MAPE sebesar 6,81%, dan R^2 sebesar 0,9747. XGBoost menghasilkan MAE sebesar USD 273,97, RMSE sebesar USD 570,25, MAPE sebesar 6,33%, dan R^2 sebesar 0,9797. Sementara itu, LightGBM mencapai performa terbaik di antara seluruh *base learner* dengan MAE sebesar USD 270,21, RMSE sebesar USD 558,66, MAPE sebesar 6,21%, dan R^2 sebesar 0,9805. Hasil tersebut menunjukkan bahwa pendekatan *gradient boosting* khususnya LightGBM memiliki kemampuan yang lebih baik dalam menangkap pola hubungan data dibandingkan metode berbasis *bagging*.

3.5.2 Pembentukan *Meta-Feature Dataset*

Prediksi dari ketiga *base learner* pada Level 0 dikombinasikan menjadi *meta-feature dataset* yang terdiri atas tiga kolom prediksi yaitu hasil prediksi dari Random Forest, XGBoost, dan LightGBM. Dataset ini digunakan sebagai masukan pada Level 1 untuk membentuk prediksi akhir melalui proses *stacking*. Pendekatan ini memungkinkan model *meta learner* mempelajari kombinasi prediksi optimal dari setiap *base learner*.

3.5.3 Level 1 (*Meta Learner*)

Pada Level 1 digunakan Regresi Linear sebagai *meta learner* untuk mengombinasikan prediksi dari seluruh *base learner*. Persamaan *stacking* pada skala Box-Cox yang diperoleh adalah:

$$\hat{Y}_{BC} = -0,0066 + 0,0427\hat{Y}_{RF} + 0,3306\hat{Y}_{XGB} + 0,6278\hat{Y}_{LGBM} \quad (9)$$

LightGBM memiliki kontribusi terbesar dengan koefisien sebesar 0,6278, diikuti XGBoost sebesar 0,3306 dan Random Forest sebesar 0,0427. Hasil tersebut sejalan dengan konsep *stacking ensemble* yang dikemukakan oleh Breiman (1996), di mana *meta learner* secara adaptif memberikan bobot yang lebih besar kepada model dengan performa prediksi yang lebih baik. Dominannya bobot LightGBM juga konsisten dengan hasil validasi silang yang menunjukkan bahwa model tersebut memiliki nilai CV-RMSE terkecil di antara seluruh *base learner*. Sebaliknya, rendahnya bobot Random Forest mengindikasikan bahwa kontribusinya terhadap prediksi akhir relatif lebih kecil dibandingkan model lainnya.

3.6 Evaluasi dan Perbandingan Model

3.6.1 Perbandingan Metrik Evaluasi

Evaluasi performa dilakukan pada data *testing* menggunakan metrik MAE, RMSE, MAPE, dan R^2 dalam satuan USD untuk membandingkan kemampuan prediksi masing-masing model secara objektif. Hasil evaluasi seluruh model disajikan pada Tabel 4.

Tabel 4: Hasil *tuning* hiperparameter *base learner*.

Model	MAE Test	RMSE Test	MAPE Test (%)	R^2 Test
Regresi Linear + Box-Cox	604,51	865,23	10,73	0,9538
<i>Random Forest</i>	295,76	636,13	6,81	0,9747
<i>XGBoost</i>	273,97	570,25	6,33	0,9797
<i>LightGBM</i>	270,21	558,66	6,21	0,9805
<i>Stacked Regression</i>	266,70	552,67	6,13	0,9809

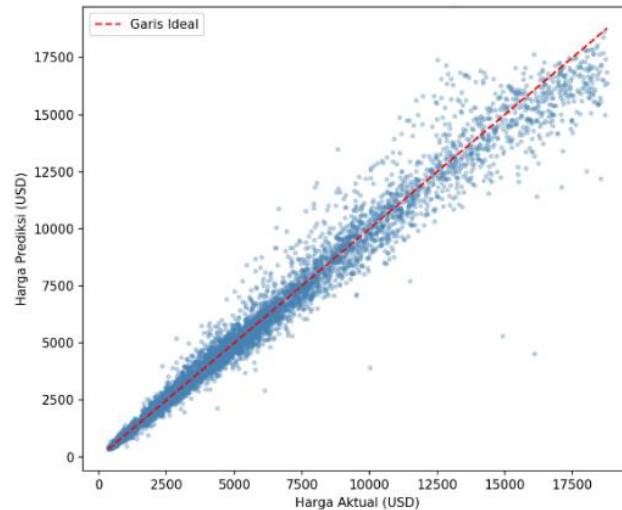
Berdasarkan Tabel 4, seluruh model *machine learning* menghasilkan performa yang lebih baik dibandingkan Regresi Linear dengan transformasi Box-Cox pada seluruh metrik evaluasi. *LightGBM* menjadi *base learner* terbaik dengan RMSE sebesar USD 558,66 dan MAPE sebesar 6,21%, sedangkan *Stacked Regression* menghasilkan performa terbaik secara keseluruhan dengan MAE sebesar USD 266,70, RMSE sebesar USD 552,67, MAPE sebesar 6,13%, dan R^2 sebesar 0,9809. Hasil ini menunjukkan bahwa pendekatan *ensemble learning* mampu meningkatkan akurasi prediksi melalui kombinasi beberapa *base learner*.

Pemeriksaan *overfitting* dilakukan melalui perbandingan performa model pada data *training* dan *testing* dengan selisih nilai R^2 . *Random Forest* menunjukkan selisih terbesar sehingga mengalami *overfitting* ringan, sedangkan *XGBoost*, *LightGBM*, dan *Stacked Regression* memiliki selisih di bawah 0,02 sehingga menunjukkan kemampuan generalisasi yang baik terhadap data baru. Secara keseluruhan, *Stacked Regression* menjadi model terbaik karena mampu menggabungkan keunggulan masing-masing *base learner* sehingga menghasilkan prediksi yang paling akurat dan stabil.

3.7 Interpretasi Model Terbaik

Analisis *feature importance* dari model *LightGBM* menunjukkan bahwa peubah y (lebar), *depth*, x (panjang), z (kedalaman), *carat*, dan *table* merupakan peubah dengan kontribusi terbesar terhadap prediksi harga berlian. Dominannya peubah numerik menunjukkan bahwa ukuran fisik berlian lebih berpengaruh terhadap pembentukan harga dibandingkan

atribut kualitas kategorik. Temuan ini sejalan dengan Kigo et al. (2023) serta Basha dan Oveis (2024) yang menunjukkan bahwa dimensi fisik dan *carat* merupakan faktor utama dalam prediksi harga berlian.



Gambar 1: Visualisasi *actual vs predicted* pada model Stacked Regression.

Visualisasi *actual vs predicted* pada Gambar 1 menunjukkan bahwa sebagian besar titik berada di sekitar garis diagonal $y = x$ yang mengindikasikan kesesuaian antara nilai aktual dan nilai prediksi. Selain itu, residual tersebar secara acak di sekitar nol tanpa pola tertentu yang menunjukkan bahwa model mampu menangkap pola hubungan data dengan baik (James et al., 2021). Nilai R^2 testing sebesar 0,9809 dan selisih R^2 antara data training dan testing sebesar 0,0087 menunjukkan bahwa Stacked Regression memiliki kemampuan generalisasi yang baik terhadap data baru serta tidak mengalami *overfitting* yang signifikan. Temuan ini sejalan dengan konsep *stacking ensemble* yang dikemukakan Breiman (1996) yaitu bahwa kombinasi beberapa model dasar dapat meningkatkan stabilitas dan akurasi prediksi.

4. Simpulan dan Saran

Penelitian ini membandingkan Regresi Linear dengan transformasi Box-Cox dan metode *ensemble learning* untuk prediksi harga berlian menggunakan *Diamonds Dataset*. Hasil evaluasi menunjukkan bahwa seluruh model *machine learning* menghasilkan performa yang lebih baik dibandingkan Regresi Linear dengan transformasi Box-Cox. Model terbaik diperoleh pada Stacked Regression dengan MAE sebesar USD 266,70, RMSE sebesar USD 552,67, MAPE sebesar 6,13%, dan R^2 sebesar 0,9809. Hasil tersebut menunjukkan bahwa pendekatan *ensemble learning* mampu meningkatkan akurasi prediksi dan memiliki kemampuan generalisasi yang baik.

Penelitian selanjutnya dapat mengembangkan metode *ensemble learning* lain seperti *Blending Ensemble* dan *Voting Regression* serta menggunakan optimasi hiperparameter berbasis *Bayesian Optimization* atau *Optuna*. Selain itu, penerapan *Explainable Artificial Intelligence* (XAI) seperti SHAP dan LIME dapat digunakan untuk meningkatkan interpretabilitas model dan memahami faktor-faktor yang memengaruhi harga berlian.

Daftar Pustaka

- Basha, M.S.A. and Oveis, P.M. (2024) 'Predictive modeling and benchmarking for diamond price estimation: integrating classification, regression, hyperparameter tuning and execution time analysis', *International Journal of System Assurance Engineering and Management*, 15, pp. 5279–5313. <https://doi.org/10.1007/s13198-024-02535-0>
- Blum, L., Elgendi, M. and Menon, C. (2022) 'Impact of Box-Cox transformation on machine-learning algorithms', *Frontiers in Artificial Intelligence*, 5, p. 877569. <https://doi.org/10.3389/frai.2022.877569>
- Box, G.E.P. and Cox, D.R. (1964) 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), pp. 211–243. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Breiman, L. (1996) 'Stacked regressions', *Machine Learning*, 24(1), pp. 49–64. <https://doi.org/10.1023/A:1018046112532>
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) *An Introduction to Statistical Learning: with Applications in R*. 2nd edn. New York: Springer. <https://doi.org/10.1007/978-1-0716-1418-1>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) 'LightGBM: a highly efficient gradient boosting decision tree', in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, 4–9 December 2017. Red Hook: Curran Associates, pp. 3146–3154.
- Kigo, S.N., Omondi, E.O. and Omolo, B.O. (2023) 'Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model', *Scientific Reports*, 13(1), p. 17315. <https://doi.org/10.1038/s41598-023-44326-w>
- Mankawade, A.A., Kokate, C., Soman, K., Mohite, A., Vispute, A. and More, O. (2023) 'Diamond price prediction using machine learning

algorithms', *International Journal for Research in Applied Science and Engineering Technology*, 11(5), pp. 4867–4871. <https://doi.org/10.22214/ijraset.2023.52741>

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2021) *Introduction to Linear Regression Analysis*. 6th edn. Hoboken: John Wiley & Sons.

Nti, I.K., Adekoya, A.F. and Weyori, B.A. (2020) 'A comprehensive evaluation of ensemble learning for stock-market prediction', *Journal of Big Data*, 7(1), p. 20. <https://doi.org/10.1186/s40537-020-00299-5>

Randhawa, A.K., Sawarkar, A.D., Mundhe, S.B., Deshmukh, A.W. and Sonarkar, P.R. (2025) 'Predicting diamond prices using regression in machine learning', in Gupta, D., Kamble, V., Satpute, V. and Kothari, A. (eds.) *Paradigm Shifts in Communication, Embedded Systems, Machine Learning, and Signal Processing. PCEMS 2024. Communications in Computer and Information Science*, vol. 2490. Cham: Springer, pp. [halaman]. https://doi.org/10.1007/978-3-031-90580-3_13

Wolpert, D.H. (1992) 'Stacked generalization', *Neural Networks*, 5(2), pp. 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)