

# Comparative Study of Linear Regression, SVR, and XGBoost for Stock Price Prediction After a Stock Split

Muhammad Yusuf Andrika<sup>1\*</sup>, Majid Rahardi<sup>2</sup>

\* Informatics, Universitas Amikom Yogyakarta, Indonesia  
[andrika00748@gmail.com](mailto:andrika00748@gmail.com)<sup>1</sup>, [majid@amikom.ac.id](mailto:majid@amikom.ac.id)<sup>2</sup>

## Article Info

### Article history:

Received 2025-07-14

Revised 2025-07-26

Accepted 2025-07-30

### Keyword:

BBCA,  
Linear Regression,  
Stock Price Prediction,  
SVR,  
XGBoost.

## ABSTRACT

This study aims to identify the most effective regression method for predicting the closing stock price of Bank Central Asia (BBCA) following the stock split event on October 12, 2021. Accurate post-split price predictions are crucial for helping investors comprehend new market dynamics, yet there is limited research evaluating the performance of regression models on BBCA's stock after such corporate actions. Using data obtained through web scraping from the Indonesia Stock Exchange, this study tested three regression algorithms Linear Regression, Support Vector Regression, and XGBoost Regressor on post-split data. The selected input features were open\_price, first\_trade, high, low, and volume, while the target was close\_price. The dataset was divided using an 80:20 train-test split and evaluated with RMSE, MAPE, and R-squared metrics. Results showed that Linear Regression achieved the best performance RMSE: 50.41, MAPE: 0.0048, R<sup>2</sup>: 0.9971, followed by XGBoost RMSE: 69.12, MAPE: 0.0058, R<sup>2</sup>: 0.9946, and SVR RMSE: 80.98, MAPE: 0.0069, R<sup>2</sup>: 0.9925. These findings indicate that BBCA's post-split stock data exhibits a linear pattern, making Linear Regression the most suitable and efficient method. This suggests that simpler models can outperform more complex algorithms when applied to stable and structured financial datasets.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. PENDAHULUAN

Stock split adalah kebijakan korporasi yang meningkatkan jumlah saham beredar dengan menurunkan nilai nominal per lembar saham, tanpa mengubah nilai perusahaan secara total[1]. Namun, menurut penelitian sebelumnya[2], setelah terjadinya stock split, harga saham cenderung mengalami fluktuasi yang signifikan. Namun secara psikologis, pasar kadang bereaksi berlebihan, menyebabkan peningkatan volatilitas serta volume transaksi yang signifikan. Fluktuasi ini menimbulkan tantangan dalam melakukan prediksi harga saham harian secara akurat, terutama bagi investor maupun analisis pasar modal yang bergantung pada data historis sebagai dasar pengambilan keputusan[3]. Prediksi harga saham pasca-stock split menjadi krusial karena data historis sebelum split seringkali tidak lagi merefleksikan pola harga baru, sehingga pendekatan konvensional yang bergantung pada kestabilan pola historis jadi kurang efektif[4].

Salah satu fenomena penting yang muncul pasca-stock split adalah perubahan signifikan pada harga dan volume transaksi saham. Hal ini sejalan dengan temuan bahwa terdapat pengaruh signifikan terhadap *return* dan *abnormal return* saham sebelum dan sesudah *stock split*[5]. Data dari Bursa Efek Indonesia menunjukkan bahwa dalam seminggu setelah *stock split*, harga saham mengalami penurunan yang cukup signifikan dibandingkan periode sebelum aksi korporasi terjadi. Selain itu, volume transaksi mengalami perubahan dibandingkan dengan pekan sebelumnya. Ketidakstabilan harga ini menjadi indikasi adanya perubahan perilaku pasar yang belum sepenuhnya dapat dijelaskan menggunakan metode konvensional.

Dalam kondisi tersebut, muncul kebutuhan untuk mengevaluasi efektivitas algoritma prediksi harga saham dengan pendekatan yang lebih kompleks dan adaptif. Linear Regression sebagai metode statistik dasar, Support Vector Regression (SVR) sebagai pendekatan berbasis margin, dan XGBoost Regressor sebagai teknik *ensemble*

yang telah terbukti unggul dalam prediksi[6], dipilih untuk dianalisis performanya terhadap data saham pasca *stock split*.

Berbagai studi sebelumnya telah mencoba memprediksi harga saham maupun saham emiten lain dengan pendekatan yang beragam, mulai dari regresi linear sederhana hingga model pembelajaran mesin kompleks seperti XGBoost, LSTM, dan Prophet. Penelitian oleh[7] menggunakan XGBoost untuk memprediksi harga saham BBCA dengan indikator teknikal dan memperoleh MAPE sekitar 4,01%. Penelitian[8] menguji regresi linear, SVR, Random Forest, dan XGBoost pada saham LQ45 dan menemukan bahwa tidak ada model yang mencapai MAPE <50% di data uji. Studi[9] menggunakan regresi linear pada saham BRI dan memperoleh RMSE tinggi, menandakan presisi rendah. Sementara itu, penelitian dengan LSTM[10] dan Prophet[11] menunjukkan bahwa metode tersebut dapat memberikan hasil akurat bila konfigurasi parameter dilakukan secara optimal. Penelitian lain oleh[12] menggunakan jaringan saraf tiruan dan mencatat akurasi sebesar 91,67% untuk saham BBCA. Hingga saat ini, Sedikit ditemukan studi komprehensif yang menerapkan algoritma regresi secara bersamaan untuk kondisi *post-split*, meskipun beberapa penelitian telah menyelidiki efek stock split secara umum atau di saham lainnya[13].

Meskipun banyak pendekatan telah digunakan, belum ada kajian yang secara eksplisit membandingkan performa regresi linear, SVR, dan XGBoost terhadap data saham pasca *stock split*. Di sinilah letak *gap* penelitian ini, Sebagian besar studi tersebut berfokus pada kondisi pasar normal atau keseluruhan periode perdagangan tanpa mempertimbangkan perubahan struktural akibat aksi korporasi stockplit dan kurangnya kajian komparatif antara beberapa algoritma regresi terhadap data saham setelah *stock split*. Padahal, memahami metode prediksi terbaik sangat krusial karena perubahan struktur harga akibat *stock split* dapat menyebabkan perbedaan pola data yang signifikan.

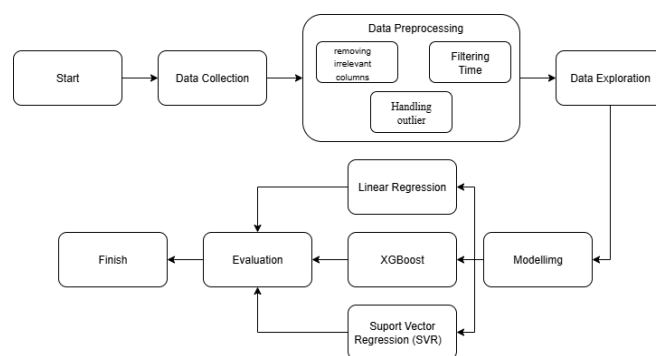
Dalam ranah keilmuan terapan, khususnya di bidang data mining dan machine learning, prediksi harga saham menjadi salah satu tantangan yang kompleks karena sifat data pasar keuangan yang bersifat time-series, non-stasioner, dan sangat dipengaruhi oleh dinamika pasar serta aksi korporasi. Salah satu tantangan utama adalah ketika data mengalami perubahan struktural, seperti akibat aksi stock split, yang menyebabkan pola historis menjadi kurang representatif terhadap kondisi saat ini. Dalam kondisi ini, pemodelan prediktif tidak hanya membutuhkan algoritma yang akurat, tetapi juga adaptif terhadap perubahan pola data. Oleh karena itu, penting dilakukan penelitian yang menguji efektivitas berbagai algoritma regresi terhadap data yang telah mengalami perubahan signifikan, untuk mengevaluasi sejauh mana pendekatan linier maupun non-linier mampu beradaptasi terhadap dinamika pasar. Penelitian ini berkontribusi pada

pengembangan keilmuan dalam hal seleksi model (model selection), pengelolaan data keuangan pasca-peristiwa (event-driven modeling), serta pemahaman terhadap kecocokan algoritma dalam konteks data yang telah terdistorsi oleh kebijakan korporasi, yang menjadi aspek penting dalam perancangan sistem prediksi berbasis pembelajaran mesin.

Penelitian ini untuk mengevaluasi dan membandingkan performa tiga algoritma prediksi saham, yaitu Linear Regression, Support Vector Regression, dan XGBoost, terhadap data saham setelah *stock split*. Data yang digunakan merupakan data historis saham dari situs resmi Bursa Efek Indonesia yang diperoleh melalui *web scraping* menggunakan Selenium, mencakup periode setelah *stock split*. Fitur yang digunakan dalam penelitian meliputi *open\_price*, *first\_trade*, *high*, *low*, dan *volume*, sedangkan variabel target adalah *close\_price*.

## II. METODE

Proses penelitian dirancang sesuai dengan gambar 1, yang menggambarkan tahap-tahapan penelitian. Penjelasan mengenai langkah-langkah yang dilakukan untuk mencapai tujuan penelitian sebagai berikut.



Gambar 1. Gambar Diagram Alur Penelitian

Tahapan utama dalam metode penelitian terdiri dari Data Collection, Data Processing, Data Exploration. Modeling (Linear Regression, XGBoost, Suport Vector Regression (SVR) Beberapa proses melibatkan tahapan-tahapan tersebut.

### A. Data Collection

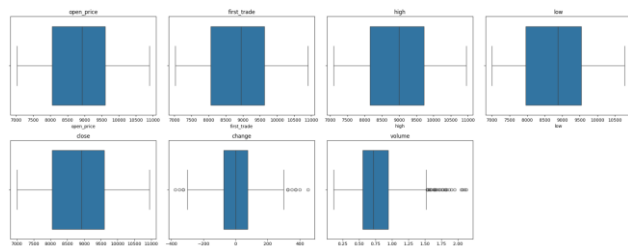
Data yang digunakan merupakan data historis saham BBCA dari situs resmi Bursa Efek Indonesia (*idx.co.id*) yang diperoleh melalui *web scraping* menggunakan Selenium, mencakup periode setelah *stock split*. Data saham BBCA diambil dari tanggal 29 Juli 2019 hingga 21 Februari 2025. Frekuensi data adalah harian, dengan 1 baris per hari perdagangan, mencakup informasi yaitu date, previous, openprice, firsttrade, high, low, close, change, volume, value, frequency, offer, indexindividual, offervolume, bid, bidvolume, tradebleshares, weightforindex, delistingdate, foreignsell, foreignbuy, nonregularfrequency,

nonregularvolume, nonregularvalue. Jumlah data didapatkan yaitu 1355 row data.

### B. Data Preprocessing

Tahap awal yang dilakukan adalah data preprocessing, yang bertujuan untuk mempersiapkan data mentah agar layak digunakan dalam proses pemodelan[14]. Proses ini diawali dengan penghapusan kolom-kolom yang tidak relevan seperti `delisting_date`, `non_regular_value`, dan `offer_volume`, karena tidak memiliki kontribusi signifikan terhadap proses prediksi harga saham harian BBKA. Selanjutnya, dilakukan filtering data berdasarkan waktu, yakni hanya mempertahankan data setelah tanggal 13 Oktober 2021, sebagai batas waktu dilakukannya stock split. Langkah ini penting agar struktur harga yang digunakan sudah sepenuhnya mencerminkan hasil perubahan nominal akibat stock split.

Tahapan berikutnya adalah penanganan outlier. Karena data keuangan seperti harga dan volume transaksi sering mengandung nilai ekstrem, dilakukan visualisasi seperti boxplot untuk mendeteksi anomali dan mengambil keputusan apakah outlier tersebut akan dihapus atau disesuaikan. Keberadaan outlier dapat menyebabkan distorsi dalam model[15] prediksi, seperti regresi linier dan XGBoost, yang sangat sensitif terhadap fluktuasi ekstrem. Oleh karena itu, dilakukan visualisasi boxplot



Gambar 2. Visualisasi Boxplot Setelah Pembersihan Outlier

Visualisasi boxplot fitur numerik saham BBKA setelah dilakukan pembersihan outlier, yang mencakup variabel `open_price`, `first_trade`, `high`, `low`, `close`, `change`, dan `volume`. Setelah proses pembersihan, distribusi data terlihat lebih seragam dan tidak menunjukkan banyak nilai ekstrem seperti sebelumnya, terutama pada fitur `change` dan `volume` yang awalnya memiliki banyak outlier. Hal ini menunjukkan bahwa data sudah lebih bersih dan siap digunakan untuk tahap pemodelan, dengan risiko bias akibat anomali yang diminimalkan. Pembersihan ini penting untuk menjaga akurasi dan generalisasi model prediksi yang dibangun. Setelah itu, dilakukan seleksi fitur.

Pemilihan variabel input seperti `open_price`, `first_trade`, `high`, `low`, dan `volume` didasarkan pada teori market microstructure, yang menjelaskan bagaimana informasi pasar (harga pembukaan, harga tertinggi/rendah) dan aktivitas volume menjadi sinyal penting dalam pembentukan harga penutupan saham[16]. Dalam literatur prediksi harga saham, fitur OHLCV (Open, High, Low, Close, Volume) sering digunakan sebagai input standar dan terbukti efektif[17].

Selanjutnya seluruh fitur dinormalisasi menggunakan teknik Min-Max Scaling dengan bantuan `MinMaxScaler` dari library `scikit-learn`. Proses ini dilakukan agar seluruh fitur memiliki rentang nilai yang seragam, yaitu antara 0 hingga 1. Normalisasi ini penting karena beberapa algoritma machine learning, seperti Support Vector Regression (SVR) dan Extreme Gradient Boosting (XGBoost), bersifat sensitif terhadap perbedaan skala antar fitur[18]. Proses scaling dilakukan setelah data dibagi ke dalam data latih dan data uji, di mana `MinMaxScaler` di-fit menggunakan data latih, dan transformasi dilakukan pada kedua set secara terpisah untuk menghindari data leakage.



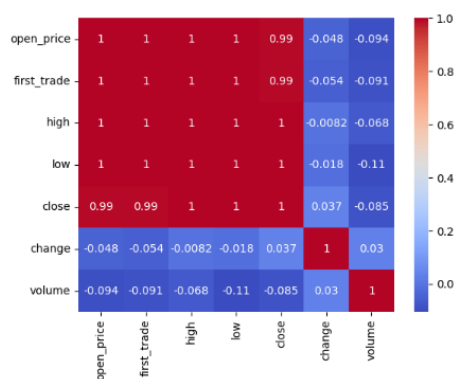
Gambar 3. Hasil Pre-processing Data

Gambar 3 menunjukkan hasil data harga penutupan saham BBKA setelah melalui tahap pre-processing, seperti konversi tanggal, penghapusan nilai null, deteksi dan penyesuaian outlier, serta normalisasi. Terlihat bahwa fluktuasi harga mengalami pola yang relatif stabil dengan kecenderungan tren naik, namun disertai dengan beberapa penurunan tajam pada titik-titik tertentu. Pola ini mencerminkan karakteristik saham pasca stock split yang lebih terjangkau dan cenderung menunjukkan respons pasar yang konsisten terhadap sentimen dan volume perdagangan.

### C. Data Exploration

Tahap eksplorasi data (Exploratory Data Analysis/EDA) dilakukan untuk memahami struktur dan pola dari dataset yang digunakan. Analisis statistik deskriptif diterapkan untuk mengetahui nilai minimum, maksimum, rata-rata, dan standar deviasi dari fitur-fitur seperti `open_price`, `high`, `low`, `close`, dan `volume`. Distribusi data juga dianalisis untuk melihat apakah penyebaran data bersifat normal atau skewed, yang dapat memengaruhi performa algoritma regresi tertentu.

Selanjutnya, korelasi antar fitur dianalisis menggunakan visualisasi heatmap. Analisis korelasi Pearson dilakukan untuk memastikan kekuatan hubungan antar fitur input dan `close_price`. Heatmap korelasi menghasilkan nilai  $r > 0.9$  untuk `open_price`, `first_trade`, `high`, dan `low`, menandakan hubungan linier yang sangat kuat dengan target. Sementara itu, `volume` memiliki korelasi moderat namun tetap dipertahankan karena mewakili tekanan pasar (demand-supply) dan memiliki dampak terhadap prediksi harga saham[19].



Gambar 4. Korelasi Heatmap

Insight dari tahap eksplorasi ini digunakan untuk mendukung proses feature selection pada tahap preprocessing. Fitur input yang dipilih adalah open\_price, first\_trade, high, low, dan volume, karena dianggap sebagai indikator utama pembentukan harga penutupan. Sedangkan target variabel atau output yang akan diprediksi adalah close, yaitu harga penutupan saham.

#### D. Modelling

Dataset dalam penelitian ini dibagi menjadi dua bagian menggunakan fungsi train\_test\_split dari pustaka Scikit-learn, dengan proporsi 80% untuk data latih (training set) dan 20% untuk data uji (testing set). Tujuan dari pembagian ini adalah untuk memastikan bahwa model memiliki data yang cukup untuk proses pelatihan, serta dapat diuji pada data yang tidak pernah dilihat sebelumnya, sehingga dapat menghindari overfitting dan mengevaluasi kemampuan generalisasi model secara objektif.

Tiga algoritma regresi digunakan untuk memodelkan dan memprediksi harga penutupan saham BBKA pasca stock split, yaitu Linear Regression, XGBoost, dan Support Vector Regression (SVR). Berikut penjelasan masing-masing:

##### 1) Linear Regression

Linear Regression adalah metode regresi linier sederhana yang digunakan sebagai baseline model. Model ini bertujuan untuk mencari hubungan linier antara fitur input dan nilai target[20]. Fungsi prediksi dari model ini dituliskan sebagai:

$$y^{\wedge} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n \quad (1)$$

##### 2) XGBoost

XGBoost (*Extreme Gradient Boosting*) merupakan metode *ensemble learning* berbasis *decision tree* yang menggabungkan sejumlah model pohon keputusan secara bertahap dengan tujuan memperbaiki kesalahan model sebelumnya. XGBoost terkenal dengan efisiensinya dan kemampuan menangani data non-linier yang kompleks.

$$Obj = i = 1 \sum nl(y_i, y^{\wedge}i(t)) + k = 1 \sum t \Omega(fk) \quad (2)$$

Untuk menghindari overfitting dan meningkatkan generalisasi model, digunakan GridSearchCV dengan 5-fold cross-validation untuk melakukan hyperparameter tuning.

Parameter yang diuji meliputi jumlah pohon (n\_estimators), kedalaman pohon (max\_depth), dan laju pembelajaran (learning\_rate). Hasil tuning menunjukkan bahwa parameter terbaik untuk dataset ini yaitu learning\_rate = 0.1, max\_depth = 3, dan n\_estimators = 250.

##### 3) Support Vector Regression (SVR)

SVR adalah algoritma yang merupakan perluasan dari Support Vector Machine (SVM) untuk tugas regresi. Tujuan SVR adalah mencari fungsi  $f(x)$  yang memiliki deviasi maksimal  $\epsilon$  dari nilai target  $y$ , dengan kompleksitas model seminimal mungkin:

$$f(x) = \langle w, x \rangle + b \quad (3)$$

Untuk memaksimalkan performa model, dilakukan hyperparameter tuning menggunakan GridSearchCV dengan skema 5-fold cross-validation. Parameter yang ditelusuri meliputi jenis kernel (linear, rbf, poly), nilai C (regularisasi), epsilon (margin toleransi), dan gamma (khusus untuk kernel non-linear). Hasil tuning menunjukkan bahwa konfigurasi optimal untuk dataset ini adalah C = 100, epsilon = 0.5, gamma = 'scale', dan kernel = 'linear'.

#### E. Evaluasi

Setiap model dievaluasi menggunakan empat metrik performa, yaitu Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), dan R-squared ( $R^2$ ). Evaluasi dilakukan pada data uji (*testing set*) untuk menilai sejauh mana model mampu melakukan generalisasi terhadap data yang tidak pernah dilihat sebelumnya. Berikut adalah table evaluasi model.

TABEL I  
HASIL EVALUASI MODEL PADA DATA LATIH

Model	RMSE	MAPE	MSE	$R^2$
Linear Regression	52.66	0.0046	2772.70	0.9969
XGBoost	32.75	0.0030	1072.70	0.9988
SVR	74.76	0.0066	5589.76	0.9938

Berdasarkan Tabel II, model XGBoost menunjukkan performa terbaik pada data latih, ditunjukkan oleh nilai RMSE terendah sebesar 32.75 dan  $R^2$  tertinggi sebesar 0.9988. Hal ini mengindikasikan bahwa model mampu menyesuaikan diri dengan data pelatihan secara sangat akurat. Linear Regression menempati posisi kedua dengan RMSE sebesar 52.66 dan  $R^2$  sebesar 0.9969, yang juga menunjukkan kinerja cukup baik, terutama dalam konteks data yang bersifat linier dan stabil. Sementara itu, SVR mencatatkan nilai RMSE tertinggi sebesar 74.76 dan  $R^2$  sebesar 0.9938, mengindikasikan bahwa model ini memiliki tingkat kesalahan terbesar di antara ketiganya saat melatih data.

### III. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil evaluasi performa dari tiga algoritma regresi dalam memprediksi harga penutupan saham PT Bank Central Asia Tbk (BBCA) setelah pelaksanaan aksi korporasi berupa stock split. Model yang dibandingkan meliputi Linear Regression sebagai baseline model linier sederhana, Support Vector Regression (SVR) yang mengandalkan konsep margin maksimal, serta XGBoost Regressor, sebuah pendekatan berbasis ensemble yang mampu menangani data non-linier dan kompleks secara efisien.

Seluruh model dilatih dan diuji menggunakan dataset yang telah melalui tahapan pre-processing, mencakup penghapusan kolom tidak relevan, deteksi dan penyesuaian outlier, normalisasi menggunakan Min-Max Scaler, serta filtering data berdasarkan periode waktu pasca stock split, yaitu setelah tanggal 13 Oktober 2021. Dataset dibagi menjadi dua bagian menggunakan fungsi `train_test_split` dari pustaka Scikit-learn dengan rasio 80% untuk data latih dan 20% untuk data uji. Pembagian ini bertujuan untuk menjaga kualitas pelatihan dan memastikan bahwa model diuji pada data yang tidak pernah dilihat sebelumnya, guna mengukur kemampuan generalisasi model secara obyektif dan menghindari overfitting.

Evaluasi model dilakukan dengan menggunakan empat metrik evaluasi utama, yaitu Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), dan R-squared ( $R^2$ ). Metrik-metrik ini dipilih karena mampu memberikan gambaran yang menyeluruh mengenai akurasi prediksi model baik dari segi skala kesalahan, persentase deviasi relatif, maupun kemampuan model dalam menjelaskan variansi data target. Selain evaluasi kuantitatif, dilakukan juga visualisasi prediksi versus nilai aktual dalam bentuk line plot dan residual plot, guna memberikan pemahaman kualitatif terhadap kesesuaian prediksi model terhadap tren aktual. Visualisasi ini berguna untuk mendeteksi pola kesalahan sistematis yang mungkin tidak tertangkap oleh metrik numerik.

Dengan pendekatan ini, diharapkan analisis yang dilakukan tidak hanya menunjukkan model mana yang paling akurat secara numerik, tetapi juga memberikan wawasan tentang kesesuaian karakteristik model dengan struktur data harga saham BBCA pasca stock split.

#### A. Hasil Evaluasi Model

Dari tabel berikut, model Linear Regression memberikan hasil terbaik dibandingkan dua algoritma lainnya, dengan nilai RMSE terendah (50.41) dan  $R^2$  tertinggi (0.9971). Ini menunjukkan bahwa model linier mampu menjelaskan 99.71% variasi harga penutupan saham berdasarkan fitur yang digunakan.

TABEL II  
HASIL EVALUASI MODEL PADA DATA UJI

Model	RMSE	MAPE	MSE	$R^2$
Linear Regression	50.41	0.0048	2541.51	0.9971
XGBoost	69.13	0.0059	4778.60	0.9946
SVR	80.98	0.0070	6558.38	0.9926

Sebaliknya, model SVR memiliki nilai RMSE dan MSE yang paling tinggi, serta nilai  $R^2$  terendah (0.9926), yang menandakan bahwa model ini tidak mampu mengikuti pola data dengan baik. Ini bisa jadi disebabkan oleh sensitivitas SVR terhadap skala data serta pemilihan parameter yang tidak optimal pada data BBCA pasca stock split.

Model XGBoost menempati posisi tengah: performanya masih cukup baik dengan  $R^2$  sebesar 0.9946, namun lebih buruk dibanding Linear Regression. Hal ini menunjukkan bahwa kompleksitas model tidak selalu berbanding lurus dengan peningkatan akurasi, terutama pada data yang bersifat linier dan stabil.

Hasil ini memberikan gambaran yang jelas bahwa dalam konteks data saham yang diolah setelah pelaksanaan aksi korporasi stock split, yang cenderung memiliki volatilitas harga yang lebih rendah, model regresi yang sederhana dan mudah diinterpretasikan seperti Linear Regression justru menunjukkan performa yang lebih baik dibandingkan dengan model-model kompleks berbasis machine learning seperti Support Vector Regression (SVR) dan XGBoost.

Hal ini dapat dijelaskan oleh karakteristik data pasca stock split yang relatif stabil, di mana harga saham BBCA cenderung bergerak lebih seragam dan tidak terpengaruh oleh fluktuasi ekstrim. Kondisi ini memungkinkan hubungan antara fitur-fitur yang digunakan untuk memprediksi harga penutupan saham seperti harga pembukaan, harga tertinggi, dan harga terendah bersifat linier dan mudah dipahami, tanpa memerlukan pemrosesan data yang rumit atau model yang sangat kompleks.

Sebaliknya, model-model seperti SVR dan XGBoost, meskipun dirancang untuk menangani data non-linier dan lebih volatile, kurang efektif pada data yang stabil ini karena kompleksitas model yang tidak diperlukan untuk pola yang relatif sederhana. Struktur yang lebih rumit pada kedua model tersebut tidak mampu menangkap pola hubungan yang sederhana, sehingga menghasilkan prediksi yang kurang akurat dibandingkan dengan Linear Regression. Oleh karena itu, dalam konteks ini, Linear Regression, yang lebih sederhana dengan parameter yang terbatas dan mudah diinterpretasikan, terbukti lebih efektif dalam menghasilkan prediksi yang akurat dan dapat diandalkan. Pendekatan ini juga lebih efisien dari segi waktu pelatihan dan komputasi, menjadikannya pilihan yang lebih tepat untuk prediksi harga saham dalam jangka pendek setelah aksi stock split yang cenderung menstabilkan harga saham.



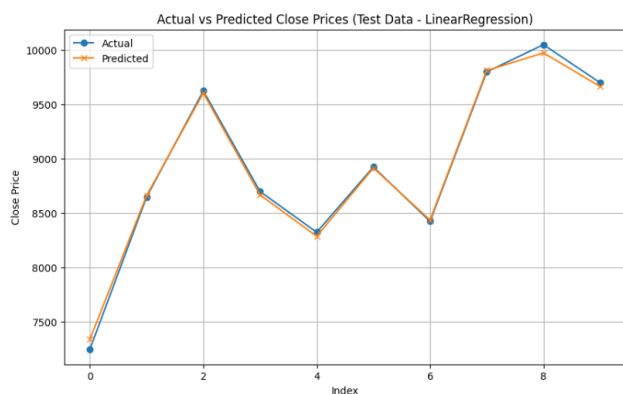
### B. Visualisasi Hasil Prediksi

Visualisasi hasil prediksi merupakan langkah penting dalam mengevaluasi kualitas model secara lebih intuitif. Meskipun evaluasi kuantitatif dengan metrik seperti RMSE, MAPE, MSE, dan  $R^2$  dapat menunjukkan performa secara numerik, namun visualisasi memungkinkan kita untuk menilai apakah prediksi model mengikuti pola fluktuasi harga saham secara keseluruhan atau justru menyimpang pada periode tertentu.

Dalam bagian ini, ditampilkan tiga grafik perbandingan antara nilai aktual dan nilai prediksi untuk masing-masing model regresi yang digunakan, yakni Linear Regression, XGBoost, dan Support Vector Regression (SVR). Grafik-grafik tersebut merepresentasikan 10 data uji terakhir setelah dilakukan `train_test_split` sebesar 80:20. Model yang ideal adalah model yang kurvanya paling mendekati kurva aktual, dengan penyimpangan (residual) yang kecil dan tidak menunjukkan pola error sistematis.

#### 1) Linear Regression

Model Linear Regression menghasilkan kurva prediksi yang sangat mendekati nilai aktual. Hampir semua titik prediksi sejajar atau sangat dekat dengan titik harga aktual, yang mengindikasikan kesalahan prediksi yang rendah. Hasil ini sejalan dengan nilai RMSE sebesar 50.41 dan  $R^2$  sebesar 0.9971, yang merupakan nilai terbaik dibandingkan model lainnya.



Gambar 5. Prediksi vs Aktual Linear Regression

Tingkat akurasi ini menunjukkan bahwa data saham BBKA pasca stock split bersifat linier, dan pendekatan Linear Regression dapat menangkap pola tersebut secara optimal tanpa memerlukan kompleksitas tinggi.

#### 2) XGBoost

Kurva prediksi dari model XGBoost juga mengikuti pola harga saham aktual dengan baik, namun terdapat beberapa penyimpangan pada titik-titik ekstrem. Nilai RMSE sebesar (69.13) serta nilai MSE sebesar 4778.60 menandakan bahwa model ini cenderung lebih sensitif terhadap fluktuasi harga.



Gambar 6. Prediksi vs Aktual XGBoost

Kecenderungan overfitting mungkin terjadi karena model ini terlalu kompleks untuk data yang stabil dan relatif linier. Walau demikian, nilai  $R^2$  masih cukup tinggi (0.9946), yang menunjukkan bahwa model tetap memiliki kemampuan prediksi yang kuat.

#### 3) Support Vector Regression (SVR)

Model SVR menunjukkan performa paling rendah dibandingkan dua model lainnya. Kurva prediksi terlihat memiliki deviasi yang lebih besar terhadap harga aktual, terutama pada titik-titik puncak dan lembah. Hal ini menyebabkan RMSE mencapai 80.98 dan  $R^2$  sebesar 0.9926



Gambar 7. Prediksi vs Aktual Support Vector Regression

Ketidaktepatan model SVR ini kemungkinan besar disebabkan oleh penggunaan kernel RBF yang tidak selaras dengan pola data yang linier dan stabil, serta karena kompleksitas SVR yang berlebihan untuk karakteristik data seperti saham BBKA pasca stock split.

### C. Analisis Per Model

Linear Regression, model ini memberikan hasil paling akurat karena karakteristik data BBKA pasca stock split menunjukkan hubungan linier antar fitur. Dengan struktur harga yang cenderung stabil, pendekatan linier mampu menangkap pola tanpa kompleksitas berlebih. Hal ini juga tercermin dari nilai  $R^2$  yang tinggi dan error prediksi yang rendah.

XGBoost, meskipun XGBoost dikenal unggul dalam berbagai tugas prediksi, hasil menunjukkan bahwa model ini sedikit overfit pada data latih. Hal ini terlihat dari performa testing yang tidak melebihi model linier. Kemungkinan ini disebabkan oleh kompleksitas model yang terlalu tinggi untuk data yang stabil dan linear.

SVR, model SVR menghasilkan performa paling rendah. MAPE lebih dari 1% dan nilai  $R^2$  di bawah 0.96 menunjukkan keterbatasan model ini pada konteks data stabil dan linier.

#### D. Diskusi Komparatif

Hasil evaluasi menunjukkan bahwa Linear Regression menghasilkan performa terbaik dibandingkan XGBoost dan SVR. Hal ini dikarenakan struktur data saham BBKA pasca stock split menunjukkan kecenderungan pola yang lebih linear dan stabil, yang lebih cocok ditangani oleh model linier sederhana. Hal ini diperkuat oleh hasil analisis korelasi Pearson, di mana fitur-fitur seperti *open\_price*, *first\_trade*, *high*, dan *low* memiliki korelasi sangat tinggi ( $r > 0.9$ ) terhadap *close\_price*, mengindikasikan adanya hubungan linier yang kuat antar fitur. Kondisi ini menyebabkan model Linear Regression mampu menangkap pola hubungan antar variabel dengan cukup baik tanpa memerlukan kompleksitas tambahan dari model non-linear.

Namun demikian, hasil feature importance dari model XGBoost menunjukkan bahwa fitur *open\_price*, *high*, dan *low* masih memberikan kontribusi dominan dalam proses prediksi. Ini menandakan bahwa XGBoost tetap mampu mengenali fitur-fitur penting, meskipun performanya secara keseluruhan tidak melebihi Linear Regression dalam konteks data ini.

Kedepan, SVR maupun XGBoost berpotensi menghasilkan performa yang lebih baik dibanding Linear Regression jika data yang digunakan mengandung volatilitas tinggi, non-linearitas, atau pola kompleks lainnya yang tidak dapat ditangkap oleh model linier. Misalnya, jika terjadi perubahan struktural di pasar, gejolak makroekonomi, atau fluktuasi sentimen pasar secara ekstrem, maka model seperti XGBoost yang dapat mempelajari pola-pola non-linier dengan baik akan menjadi lebih unggul. Oleh karena itu, pemilihan model yang tepat sangat bergantung pada karakteristik dan dinamika data yang digunakan.

Perbedaan utama penelitian ini adalah fokus pada data setelah aksi korporasi (stock split) yang menyebabkan perubahan struktur data. Maka dari itu, Linear Regression yang sederhana justru paling mampu menangkap pola harga saham pasca aksi tersebut.

#### IV. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan performa tiga algoritma regresi—Linear Regression, XGBoost, dan Support Vector Regression (SVR)—dalam memprediksi harga penutupan saham BBKA setelah terjadinya aksi korporasi berupa stock split. Evaluasi dilakukan menggunakan metrik Root Mean Squared Error (RMSE),

Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), dan koefisien determinasi ( $R^2$ ), berdasarkan hasil prediksi pada data uji yang telah melalui proses normalisasi dan pembagian data sebesar 80:20.

Linear Regression memberikan hasil terbaik dengan RMSE sebesar 50.41 dan  $R^2$  mencapai 0.9971, menunjukkan bahwa model ini dapat menjelaskan 99.71% variansi harga saham BBKA. Model ini bekerja optimal karena karakteristik data pasca stock split yang bersifat linier dan stabil. Dalam konteks ini, data saham BBKA cenderung memiliki fluktuasi harga yang tidak terlalu besar, dan hubungan antara fitur-fitur seperti harga pembukaan, harga tertinggi, dan harga terendah dengan harga penutupan saham sangat erat.

Pola harga yang stabil ini mendukung penggunaan model yang lebih sederhana dan interpretable seperti Linear Regression. Hal ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa dalam kasus data saham dengan pola yang lebih stabil dan sedikit volatilitas, model linier sering kali memberikan akurasi yang lebih tinggi daripada model yang lebih kompleks.

Sebaliknya, meskipun XGBoost dan SVR merupakan model yang lebih kompleks dan sering kali unggul dalam menangani data dengan non-linearitas tinggi, pada data ini performanya justru sedikit lebih buruk dibanding Linear Regression. Model XGBoost memiliki nilai RMSE sebesar 69.13 dan  $R^2$  sebesar 0.9946, sedangkan SVR menunjukkan performa yang lebih rendah dengan RMSE sebesar 80.98 dan  $R^2$  sebesar 0.9926.

Meskipun XGBoost dikenal memiliki kemampuan untuk menangani data non-linier dan kompleks, pada data BBKA yang relatif linier pasca stock split, model ini cenderung terlalu kompleks untuk pola yang sederhana, sehingga tidak menunjukkan akurasi yang lebih tinggi pada data uji dibandingkan dengan Linear Regression. Begitu pula dengan SVR, yang mengalami kesulitan dalam mengikuti pola data yang stabil dan linier. Kompleksitas model-model ini tidak diperlukan untuk menangani data yang relatif sederhana, dan pendekatan machine learning yang lebih rumit justru tidak memberikan peningkatan akurasi yang signifikan dalam konteks data ini. Hal ini mengindikasikan bahwa penggunaan model yang terlalu kompleks belum tentu memberikan peningkatan akurasi, khususnya ketika struktur data bersifat sederhana dan minim fluktuasi.

Hasil ini dapat dimanfaatkan oleh investor, analis keuangan, dan pengembang sistem prediksi saham yang ingin menggunakan pendekatan ringan namun tetap akurat. Linear Regression memberikan keunggulan dari segi kecepatan pelatihan, interpretabilitas model, serta efisiensi komputasi, menjadikannya pilihan yang layak untuk prediksi jangka pendek maupun analisis kuantitatif berbasis data historis pasca aksi korporasi.

Secara keseluruhan, penelitian ini memberikan kontribusi penting dalam menegaskan bahwa pemilihan model yang tepat sangat bergantung pada karakteristik data, bukan semata-mata pada tingkat kompleksitas algoritma yang digunakan.

## DAFTAR PUSTAKA

- [1] E. Mardiyarningsih and R. Andhityara, "Analisis Perbandingan Sebelum Dan Sesudah Stock Split Dengan Tingkat Likuiditas Saham, Harga Saham, Dan Return Saham Pada Indeks Saham Kompas 100 Tahun 2014 - 2018," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 4, no. 1, pp. 1–13, 2020, [Online]. Available: <http://journal.stmikjayakarta.ac.id/index.php/jisamar/article/view/166>
- [2] A. Rohim, "Analisis Perbedaan Nilai Harga Saham Sebelum Dan Setelah Stock Split," *JMD J. Ris. Manaj. Bisnis Dewantara*, vol. 4, no. 1, pp. 55–66, 2021, doi: 10.26533/jmd.v4i1.771.
- [3] A. Wijaya and I. Fenriana, "Prediksi Harga Saham Top 10 NASDAQ dengan Time Series Prophet," vol. 7, no. 2, 2024, doi: 10.32877/bt.v7i2.1736.
- [4] I. B. P. S. Ni Kadek Wiwik Yuniartini, "Dampak stock split terhadap harga saham dan aktivitas volume perdagangan saham di Bursa Efek Indonesia (Doctoral dissertation, Udayana University)."
- [5] P. Sabdowati and Z. Zulmaita, "Analisis Pengaruh Stock Split Pada Pergerakan Saham PT. Bank Central Asia Tbk," *Semin. Nas. Akunt. dan ...*, 2022, [Online]. Available: <http://prosiding-old.pnj.ac.id/index.php/snampnj/article/download/5890/2993>
- [6] Galih Adhi Putratama, Satya Maulana Fahriza, and Yudhistira Rakha Ramandhani, "Evaluasi Komparatif Metode Machine Learning Untuk Memprediksi Perubahan Harga Saham," *Antivirus J. Ilm. Tek. Inform.*, vol. 17, no. 2, pp. 278–285, 2024, doi: 10.35457/antivirus.v17i2.2871.
- [7] B. Jange, "Prediksi Harga Saham Bank BCA Menggunakan XGBoost," *Arbitr. J. Econ. Account.*, vol. 3, no. 2, pp. 231–237, 2022, doi: 10.47065/arbitrase.v3i2.495.
- [8] Ricky, "Implementasi Model Machine Learning Dalam Memprediksi Return Saham," UIN Syarif Hidayatullah Jakarta, 2022. [Online]. Available: <http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf>
- [9] M. Magdalena, A. P. Safira, and I. Maulida, "Penerapan Algoritma Linear Regression Dalam Memprediksi Harga Saham Bank BRI," vol. 2, no. 3, 2024.
- [10] R. Julian and M. R. Pribadi, "Peramalan Harga Saham Pertambangan Pada Bursa Efek Indonesia (BEI) Menggunakan Long Short Term Memory (LSTM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 3, pp. 1570–1580, 2021, doi: 10.35957/jatisi.v8i3.1159.
- [11] B. Jange, "Prediksi Harga Saham Bank BCA Menggunakan Prophet," *Arbitr. J. Econ. Account.*, vol. 3, no. 2, pp. 231–237, 2022, doi: 10.47065/arbitrase.v3i2.495.
- [12] A. P. W. Ihda Innar Ridho, Cerah Fitri Ramadhani, "Penerapan Artificial Neural Network dengan Algoritma Backpropagation untuk Memprediksi Harga Saham," *J. Ris. Stat.*, vol. 8, pp. 107–118, 2023, doi: 10.29313/jrs.v3i2.2953.
- [13] S. Soewignjo, Sediono, M. F. F. Mardianto, and E. Pusporani, "Prediksi Harga Saham Bank BCA (BBCA) Pasca Stock Split dengan Artificial Neural Network dengan Algoritma Backpropagation," *G-Tech J. Teknol. Terap.*, vol. 7, no. 4, pp. 1683–1693, 2023, doi: 10.33379/gtech.v7i4.3363.
- [14] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.
- [15] A. A. Shehadeh, S. M. Alwadi, and M. I. Almaharmeh, "Detecting and Analysing Possible Outliers in Global Stock Market Returns," *Cogent Econ. Financ.*, vol. 10, no. 1, 2022, doi: 10.1080/23322039.2022.2066762.
- [16] K. Alkhatib, H. Khazaleh, H. A. Alkhazaleh, A. R. Alsoud, and L. Abualigah, "A New Stock Price Forecasting Method Using Active Deep Learning Approach," *J. Open Innov. Technol. Mark. Complex.*, vol. 8, no. 2, 2022, doi: 10.3390/joitmc8020096.
- [17] S. Kang and J.-K. Kim, "Stock Price Prediction Using Triple Barrier Labeling and Raw OHLCV Data: Evidence from Korean Markets," 2025, [Online]. Available: <https://arxiv.org/pdf/2504.02249v1>
- [18] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling In Machine Learning: Effects on Regression and Classification Tasks," vol. XX, no. X, 2025, [Online]. Available: <http://arxiv.org/abs/2506.08274>
- [19] R. Plachý, "Impact of trading volume on prediction of stock market development," *Acta Univ. Agric. Silvic. Mendelianae Brun.*, vol. 62, no. 6, pp. 1373–1380, 2014, doi: 10.11118/actaun201462061373.
- [20] K. Qu, "Research on linear regression algorithm," *MATEC Web Conf.*, vol. 395, p. 01046, 2024, doi: 10.1051/mateconf/202439501046.