

PERBANDINGAN PREDIKSI INDEKS HARGA SAHAM GABUNGAN (IHSG) MENGUNAKAN METODE REGRESI *RANDOM FOREST* DAN *EXTREME GRADIENT BOOSTING*

Agnes Monika Simalango¹, Muklas Rivai^{2*}, Fuji Lestari³

^{1,2,3} Program Studi Sains Aktuaria, Fakultas Sains, Institut Teknologi Sumatera

*e-mail: muklas.rivai@at.itera.ac.id

DOI: 10.14710/j.gauss.15.2.356-367

Article Info:

Received: 2026-03-24

Accepted: 2026-09-18

Available Online: 2026-09-30

Keywords:

IHSG; Random Forest; Extreme Gradient Boosting; Prediction; Variable Importance.

Abstract: Indeks Harga Saham Gabungan (IHSG) serves as the benchmark of the Indonesian stock market and is characterized by nonlinear and volatile movements driven by macroeconomic conditions, trading activities, and market sentiment. This study aimed to compare the forecasting performance of Random Forest and Extreme Gradient Boosting (XGBoost) for the IHSG using historical data from January 2019 to July 2025 with an 80:20 chronological train-test split. Model performance was evaluated using RMSE and MAPE, followed by variable importance analysis and Value at Risk (VaR) estimation to assess potential investment losses. The results showed that Random Forest outperformed XGBoost, achieving an RMSE of 117.54 and a MAPE of 0.90%, whereas XGBoost produced an RMSE of 136.45 and a MAPE of 1.03%. Variable importance analysis identified the highest index value as the most influential predictor, followed by the lowest index value, the opening index value, and trading volume. Furthermore, the estimated expected return was 0.00014, with VaR values of IDR 470,000.00, IDR 610,000.00, and IDR 960,000.00 at the 90%, 95%, and 99% confidence levels, respectively. Based on the evaluation results, Random Forest achieved better forecasting accuracy than XGBoost for the IHSG.

1. PENDAHULUAN

Pesatnya perkembangan pasar modal Indonesia ditunjukkan oleh peningkatan jumlah investor per Juli 2025 mencapai 17.465.848 *Single Investor Identification* (SID), meningkat sekitar 17,44% dibandingkan Desember 2024 sebanyak 14.871.639 SID ((KSEI), 2025). Saham adalah instrumen populer karena potensi *capital gain* dan dividen, meskipun memiliki risiko fluktuasi yang tinggi. Indikator yang digunakan dalam melihat pergerakan saham tercatat adalah indeks saham, yang pada pasar saham Indonesia dikenal sebagai Indeks Harga Saham Gabungan (IHSG). Nilai IHSG dipengaruhi oleh berbagai kondisi ekonomi makro, aktivitas perdagangan serta sentimen pasar, sehingga prediksi IHSG menjadi langkah krusial untuk menghindari kerugian investasi dan membantu perumusan kebijakan ekonomi. Namun, karakteristik data IHSG yang nonlinear, fluktuatif dan memiliki ketergantungan terhadap waktu menjadi tantangan dalam menghasilkan prediksi akurat (Indonesia Stock Exchange, 2008).

Berdasarkan karakteristik IHSG tersebut, *machine learning* berbasis *decision tree ensemble* menjadi salah satu pendekatan yang sesuai. Di antara berbagai algoritma *machine learning*, *Random Forest* dan *Extreme Gradient Boosting* (XGBoost) merupakan metode berbasis *decision tree ensemble* yang efektif dalam memodelkan hubungan nonlinear. *Random Forest* menggunakan teknik *bagging* dengan merata-ratakan hasil prediksi dari sekumpulan pohon. Sementara itu, *Extreme Gradient Boosting* menggunakan teknik *boosting* yang akan

membangun pohon secara bertahap untuk meminimalkan kesalahan dari model sebelumnya (Ernawati, 2023).

Kinerja dari kedua metode dipengaruhi oleh pemilihan *hyperparameter* yang tepat dan akan membantu pembentukan model optimal melalui pembelajaran pola data IHSG. Penelitian ini menggunakan teknik *Grid Search* untuk memperoleh kombinasi *hyperparameter* optimal. IHSG merupakan jenis data yang bergantung pada waktu, sehingga proses optimasi akan dilakukan berdasarkan urutan waktu dengan *Time Series Cross Validation* (*TimeSeries CV*) agar urutan waktu tetap terjaga dan kebocoran data (*data leakage*) tidak terjadi. Selain akurasi prediksi, pemahaman mengenai kontribusi variabel prediktor melalui *variable importance* serta estimasi risiko kerugian pada tingkat kepercayaan tertentu dari hasil prediksi melalui *Value at Risk* (VaR) juga menjadi komponen penting untuk memberikan wawasan bagi pelaku kepentingan investasi di pasar saham Indonesia (Solihatun, et al., 2023).

Penelitian (Baihaqi, et al., 2023; Astutiningsih, et al., 2023) menunjukkan bahwa metode *Random Forest* dan *XGBoost* mampu melakukan prediksi akurat berdasarkan akurasi kesalahan prediksi kecil. Sementara itu, (Ernawati, 2023) menyatakan bahwa *Random Forest* lebih unggul dibandingkan *XGBoost* dalam prediksi indeks ISSI tanpa optimasi *hyperparameter*. Berbeda dengan penelitian sebelumnya, penelitian ini akan membandingkan *Random Forest* dan *XGBoost* dalam prediksi IHSG sebagai indeks saham utama Indonesia melalui optimasi *hyperparameter Grid Search* dan *Time Series CV*, serta menganalisis *variable importance* dan *Value at Risk* (VaR).

2. TINJAUAN PUSTAKA

Indeks Harga Saham Gabungan (IHSG) merupakan indeks saham utama Indonesia yang mencakup nilai indeks pembukaan, penutupan, tertinggi, terendah dan volume perdagangan (Patriya, 2020). Data historis IHSG termasuk data runtun waktu sehingga mengharuskan penggunaan metode analisis yang mampu mempertahankan urutan waktu selama proses pemodelan dan validasi.

Pembagian data menjadi data latih dan uji menjadi tahapan penting untuk memastikan kinerja model pada data baru setelah pelatihan. Data latih adalah data yang digunakan untuk melatih model sedangkan data uji adalah pengecualian dari data latih untuk validasi model pelatihan (Priyatno, et al., 2023). Uji silang (*Cross Validation* / CV) merupakan teknik validasi untuk mengevaluasi performa model. Penelitian ini menggunakan *Time Series Cross Validation* (*TimeSeries CV*) dengan metode *expanding window* melalui *TimeSeriesSplit* dan untuk jumlah *fold* yang digunakan adalah sejumlah 5 *fold*. Metode ini mempertahankan urutan waktu data, yakni data latih akan bertambah seiring waktu ke depan pada setiap *fold*, sedangkan jumlah data validasi adalah tetap dan berada setelah urutan waktu data latih sehingga urutan waktu tetap terjaga dan *data leakage* dapat dihindari demi periode validasi seimbang (Priyatno, et al., 2023).

Min Max Normalization digunakan untuk menyamaratakan rentang nilai variabel berbeda menjadi kisaran 0 sampai 1, yang bertujuan untuk memudahkan proses pemodelan dengan nilai setiap variabel yang konsisten dan dapat dilihat pada Persamaan (1) (Agus Ambarwari, 2020).

$$v_{norm} = \frac{v_o - v_{min}}{v_{max} - v_{min}} \quad (1)$$

Adapun v_{norm} adalah nilai hasil normalisasi data, v_o adalah nilai asli dari data ke- o , v_{min} adalah nilai minimum dari data, dan v_{max} adalah nilai maksimum.

Pembelajaran mesin (*machine learning*) adalah studi pengembangan algoritma komputer yang mempelajari pola data demi pengambilan keputusan tertentu. Pembelajaran ini terdiri dari *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Teknik ansambel (*ensemble*) menjadi teknik dalam *machine learning* yang terdiri dari

1. *Bagging*, teknik yang menggunakan *bootstrap sampling* yakni membentuk sampel *bootstrap* dari data latih, membangun pohon keputusan secara paralel dan kemudian menggabungkan hasil prediksinya.
2. *Boosting*, teknik yang memusatkan perhatian pada pengurangan residual pada model selanjutnya sehingga akan mengurangi bias pada model.
3. *Stacking*, teknik yang menggabungkan beberapa model dasar menjadi model meta sebagai model prediksi akhir.

Regresi adalah salah satu kategori *supervised learning* yang menggambarkan pengaruh variabel target dengan satu atau lebih variabel penjelas (Hartono, 2024). Regresi *Random Forest* mengacu pada teknik *bagging* dalam memberikan hasil prediksi. Nilai rata-rata dari setiap pohon keputusan akan diambil sebagai prediksi akhir menggunakan Persamaan (2).

$$\hat{y}_t = \frac{1}{n_{tree}} \sum_{j=1}^{n_{tree}} \hat{y}_j \quad (2)$$

Adapun $\hat{y}_t$ adalah hasil prediksi akhir dari semua pohon keputusan akhir ke $-t$, n_{tree} adalah jumlah pohon yang dibangun dan $\hat{y}_j$ adalah hasil prediksi pada pohon ke $-j$. Nilai prediksi pada setiap pohon ke $-j$ dihitung menggunakan Persamaan (3).

$$\hat{y}_j = \frac{1}{|R_{jm}|} \sum_{i \in R_{jm}} y_i \quad (3)$$

Adapun $\hat{y}_j$ adalah hasil prediksi akhir pada pohon keputusan ke $-j$, R_{jm} adalah simpul daun ke $-m$ pada pohon ke $-j$, $|R_{jm}|$ merupakan jumlah data latih yang berada pada simpul daun tersebut, dan y_i adalah nilai target data latih ke $-i$ pada simpul daun R_{jm} . *Hyperparameter* merupakan parameter yang berguna untuk mengoptimalkan performa model prediksi, yang pada *Random Forest* antara lain adalah jumlah pohon ($n_estimators$), jumlah fitur ($max_features$), kedalaman pohon (max_depth), jumlah minimum sampel ($min_samples_leaf$), dan jumlah minimum *split* ($min_samples_split$) (Zhu, et al., 2022).

Regresi *XGBoost* adalah jenis regresi *machine learning* yang dikembangkan dari *Gradient Boosting Regression Tree* (GBRT), serta teknik *boosting* dalam membangun pohon keputusan dengan melakukan perbaikan dari model sebelumnya secara berurutan melalui Persamaan (4)

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) \quad (4)$$

Nilai $f_k(x_i)$ adalah model pohon pada prediksi data ke- i periode ke $-t$ dan diumpakan sebagai $\hat{y}_i^{(t)}$. Fungsi pembelajaran metode ini adalah fungsi objektif yang terdiri dari fungsi kerugian dan regularisasi yang bertujuan untuk menemukan algoritma optimum pada Persamaan (5).

$$Obj = \sum_{i=1}^t \ell(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{n_{tree}} \Omega(f_k) \quad (5)$$

n_{tree} adalah jumlah pohon yang akan dibangun, y_i adalah nilai aktual dari data ke- i , $\hat{y}_i^{(t)}$ adalah nilai prediksi dari data ke- i untuk periode ke- t , f_k adalah model prediksi dari pohon ke- k untuk periode ke- t , Ω merupakan fungsi regularisasi yang memberikan penalti pada

kompleksitas untuk mencegah *overfitting*, dan $\ell(y_i, \hat{y}_i^{(t)})$ adalah fungsi kerugian yang digunakan untuk mengukur kesalahan prediksi. Sedangkan fungsi regularisasi digunakan dalam mengurangi kompleksitas data sehingga mencegah *overfitting* yang dihitung menggunakan Persamaan (6)

$$\Omega(f) = \gamma T_i + \frac{1}{2} \lambda \|\omega\|^2 \quad (6)$$

λ adalah parameter regularisasi yang memiliki nilai default sebesar 1, T_i merupakan jumlah ‘leaf’ yang terdapat pada *tree* ke- i , Ω merupakan fungsi untuk menentukan kompleksitas model berdasarkan nilai γ adalah bobot “leaf” yang bertindak sebagai nilai keluaran (*output value*), γ adalah parameter regularisasi yang memiliki nilai default sebesar 0. *Hyperparameter* pada metode ini meliputi jumlah pohon keputusan (*n_estimators*), ukuran langkah dalam memperbaiki model (*learning_rate*), kedalaman maksimum dari pohon keputusan (*max_depth*), bobot minimum yang diperlukan oleh daun (*min_child_weight*), serta proporsi data untuk membangun setiap pohon (*sub_sample*) (Chen & Guestrin, 2016). Masing-masing metode membutuhkan proses penyesuaian hiperparameter (*hyperparameter tuning*) untuk mencari *hyperparameter* optimal dalam membangun model akurat. *Grid Search* menjadi metode yang dapat melakukan pencarian tersebut dengan uji kombinasi satu per satu dan validasi untuk setiap kombinasi serta penggunaan *cross validation* untuk menciptakan indeks evaluasi setiap model (Patriya, 2020).

Model regresi yang akurat dapat diukur menggunakan metrik evaluasi yakni *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE).

- a) *Root Mean Squared Error* (RMSE), metode yang mengukur seberapa jauh nilai dugaan dan aktual berdasarkan akar dari rata-rata kesalahan dalam kuadrat yang dapat dihitung menggunakan Persamaan (7)

$$MSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (7)$$

- b) *Mean Absolute Percentage Error* (MAPE), metode yang digunakan untuk mengukur kesalahan relatif dari model dalam bentuk persentase, dan dihitung menggunakan Persamaan (8).

$$MAPE = \frac{1}{n} \left(\sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \right) \times 100\% \quad (8)$$

Kategori MAPE terdiri dari 4 jenis yang dapat dilihat pada Tabel (1) (Rafirhan, et al., 2023).

Tabel 1. Tingkat Akurasi MAPE	
Tingkat Akurasi Model	Nilai MAPE
Sangat Akurat	<10%
Akurat	10% - 20%
Cukup Akurat	20% - 50%
Tidak Akurat	>50%

Adapun y_t merupakan nilai aktual pada periode ke - t , $\hat{y}_t$ adalah nilai dugaan pada periode ke- t , dan n adalah banyaknya data.

Variable importance digunakan untuk melihat pengaruh variabel prediktor terhadap hasil prediksi dari variabel target. Semakin tinggi nilai suatu *variable importance* suatu variabel prediktor maka semakin besar pengaruh variabel tersebut terhadap hasil prediksi (Lardy, 2019).

Random Forest dan *XGBoost* menyediakan mekanisme bawaan untuk memahami pentingnya variabel prediktor terhadap hasil prediksi.

a) Regresi *Random Forest*

Perhitungan *variable importance* pada metode ini dihitung dari *Mean Decrease in Impurity* (MDI) berdasarkan rata-rata penurunan ketidakmurnian atau ketidakaturan suatu *node* pada setiap variabel yang diukur menggunakan persamaan (9)

$$MDI = \frac{1}{Q} \sum_{q=1}^Q \sum_{c \in T_q} p(c) \Delta V(c) \quad (9)$$

Q adalah jumlah pohon total pada model, T_q adalah himpunan simpul pada pohon ke- q , $p(c)$ adalah proporsi sampel pada simpul ke- c , N_c adalah jumlah total observasi dalam simpul ke- c , N_{total} adalah jumlah data dalam dataset, $\Delta V(c)$ yakni penurunan impuritas pada simpul pohon.

b) Regresi *XGBoost*

XGBoost menyediakan *variable importance* berdasarkan rata-rata peningkatan fungsi kerugian ketika fitur digunakan dalam pemodelan. *Gain* dihitung berdasarkan penurunan nilai *loss* setelah *split* menggunakan suatu fitur, untuk melihat seberapa besar kontribusi yang diberikan variabel tersebut yang dapat dilihat pada Persamaan (10)

$$Gain = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{(H_L + H_R + \lambda)} \right) - \gamma \quad (10)$$

G_L adalah total gradien (turunan pertama fungsi kerugian) untuk *node* kiri, G_R adalah total gradien (turunan pertama fungsi kerugian) untuk *node* kanan, H_L adalah total hessian (turunan kedua fungsi kerugian) untuk *node* kiri, H_R adalah total hessian (turunan kedua fungsi kerugian) untuk *node* kanan, λ adalah parameter regularisasi L2 untuk menghindari *overfitting*, γ adalah parameter penalti *split*, sedangkan nilai *gain* adalah nilai penurunan fungsi *loss* karena *split*, yaitu seberapa bermanfaat *split* tersebut (Chen & Guestrin, 2016).

Imbal hasil (*return*) merupakan yang ingin dicapai oleh investor melalui kegiatan investasi yang dipengaruhi oleh kondisi emiten dan pasar yang dihitung menggunakan Persamaan (11)

$$r_t = \frac{(S_t - S_{t-1})}{S_{t-1}} \quad (11)$$

r_t adalah *return* indeks pada periode ke- t , S_t adalah nilai indeks penutupan pada periode - t , S_{t-1} adalah nilai indeks penutupan pada periode ke- $t - 1$. Sedangkan ekspektasi imbal hasil dihitung menggunakan Persamaan (12)

$$E(r_t) = \frac{\sum_{t=1}^m r_t}{m} \quad (12)$$

$E(r_t)$ adalah *expected return* indeks pada periode ke- t , r_t adalah *return* pada periode - t , m adalah jumlah total data *return*. Risiko merupakan tingkat kemungkinan perbedaan antara *return* yang terjadi dan tidak terjadi dan dapat dihitung dengan *Value at Risk* (VaR). VaR merupakan teknik menghitung kerugian maksimum pada periode dan tingkat kepercayaan tertentu menggunakan Persamaan (13)

$$VaR(1 - \alpha) = V_0 \times r_\alpha \quad (13)$$

$VaR(1 - \alpha)$ adalah besarnya potensi kerugian maksimal pada tingkat kepercayaan tertentu, r_α adalah *return* pada persentil ke- α , V_0 merupakan dana investasi awal. Persentil pada perhitungan VaR simulasi historis dinyatakan menggunakan Persamaan (14)

$$r_\alpha = \alpha \times z \quad (14)$$

α adalah tingkat signifikansi, dan z merupakan jumlah observasi dari data historis (Solihatun, et al., 2023).

3. METODE PENELITIAN

Data IHSG pada penelitian ini bersumber dari (www.investing.com) periode 01 Januari 2019 - 31 Juli 2025 dengan jumlah data 1.588 baris. Variabel yang digunakan mencakup indeks penutupan (Y) sebagai variabel respon, indeks pembukaan (X_1), indeks tertinggi (X_2), indeks terendah (X_3) serta volume perdagangan (X_4) IHSG sebagai variabel prediktor. Adapun tahapan yang perlu diikuti dalam penelitian ini, meliputi:

1. Mengumpulkan data historis harian IHSG serta melakukan analisis deskriptif data.
2. Membagi data IHSG data latih dan uji berdasarkan urutan waktu dengan rasio 80:20, yakni data latih mencakup periode 1 Januari 2019 - 20 Maret 2024 dan data uji periode 21 Maret 2024 - 31 Juli 2025.
3. Membagi data latih menjadi 5 *fold* menggunakan *cross validation* untuk *time series* demi pelatihan model secara bertahap sesuai urutan waktu.
4. Normalisasi nilai variabel prediktor pada data latih menggunakan *Min Max Normalization* agar tidak terjadi *data leakage*.
5. Mengimplementasikan metode regresi *Random Forest* dan *Extreme Gradient Boosting*
 - a. Metode regresi *Random Forest*
 1. Menentukan ruang *hyperparameter* awal.
 2. Membangun pohon keputusan menggunakan data latih dengan teknik *bagging* dan *grid search CV time series* demi model prediksi dengan kombinasi optimal berdasarkan rata-rata MSE terkecil.
 3. Model dengan kombinasi terbaik akan divalidasi pada data uji untuk mendapatkan nilai dugaan IHSG.
 - b. Metode regresi *Extreme Gradient Boosting*
 1. Menentukan ruang *hyperparameter* awal.
 2. Melatih model pada data latih menggunakan teknik *boosting* dan *grid search CV time series* untuk mendapatkan model terbaik dengan kombinasi optimal berdasarkan nilai rata-rata MSE terkecil.
 3. Model dengan kombinasi terbaik akan divalidasi pada data uji untuk mendapatkan nilai dugaan IHSG.
6. Mengevaluasi model menggunakan RMSE dan MAPE untuk mendapatkan model terbaik.
7. Menganalisis *variable importance* untuk melihat pengaruh variabel prediktor menggunakan metode terbaik.
8. Meramalkan nilai penutupan IHSG 30 hari ke depan menggunakan metode terbaik.
9. Menghitung VaR simulasi historis berdasarkan *return* hasil peramalan IHSG sebagai estimasi kerugian dengan taraf signifikansi 90%, 95% dan 99%.

4. HASIL DAN PEMBAHASAN

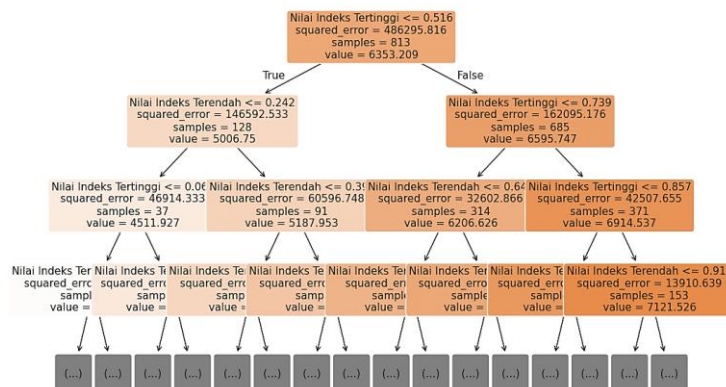
Analisis deskriptif dari data historis IHSG memberikan gambaran umum terkait data berdasarkan statistika deskriptif untuk masing-masing nilai indeks yang disajikan pada **Tabel 2**.

Tabel 2. Statistika deskriptif data historis IHSG

Sari Numerik	Nilai Indeks Pembukaan (X_1)	Nilai Indeks Tertinggi (X_2)	Nilai Indeks Terendah (X_3)	Nilai Volume Perdagangan (X_4)	Nilai Indeks Penutupan (Y)
Rata- rata	6.516,80	6.551,17	6.475,83	154.319.300	6.513,22
Standar deviasi	709,99	707,56	712,65	64.065.280	709,50
Maksimum	7.904,40	7.910,56	7.853,35	615.071.900	7.905,39
Minimum	3.937,63	4.123,56	3.911,72	20.620.000	3.937,63
Kuartil-1	6.152,00	6.178,68	6.115,30	113.648.000	6.148,16
Median	6.660,54	6.691,66	6.621,84	151.916.400	6.652,75
Kuartil-3	7.032,04	7.070,27	6.987,45	192.393.400	7.031,48

Berdasarkan Tabel 2, nilai indeks penutupan IHSG selama periode Januari 2019 – Juli 2025 berkisar antara 3.937,63 dan 7.905,39. Periode tersebut mencakup kondisi pandemi COVID-19, pemulihan hingga penguatan pasar pada akhir periode pengamatan yang menunjukkan terdapat karakteristik yang fluktuatif. Rasio pembagian data latih dan uji menggunakan rasio 80:20, yakni periode data latih dimulai dari 01 Januari 2019 – 20 Maret 2024 dengan jumlah data 1.271 sedangkan data uji berjumlah 317 baris dengan periode 21 Maret 2024 – 31 Juli 2025. *Cross validation* untuk data *time series* akan digunakan pada data latih untuk pelatihan model secara bertahap. Seluruh data latih akan dibagi menjadi 5 ukuran seimbang (*5 fold*) dengan jumlah validasi internal yang sama yakni 211 data validasi dan akan bertambah seiring dengan urutan waktu.

Min Max Normalization digunakan baik pada variabel prediktor data latih maupun data uji. Hasil normalisasi data pada data latih akan digunakan sebagai skala untuk data uji, sehingga urutan waktu akan terjaga sehingga tidak terjadinya kebocoran data (*leakage*). *Hyperparameter* pada metode *Random Forest* terdiri dari *n_estimators* [100,200,300], *max_depth* [3,5,7], *max_features* [1, None], *min_sample_leaf* [1,2,4], *min_sample_split* [2,5,10]. Pelatihan model akan dilakukan dengan membangun pohon keputusan oleh teknik *bagging* serta *grid search* untuk mendapatkan model dengan kombinasi terbaik dan dapat dilihat pada **Gambar 1**.



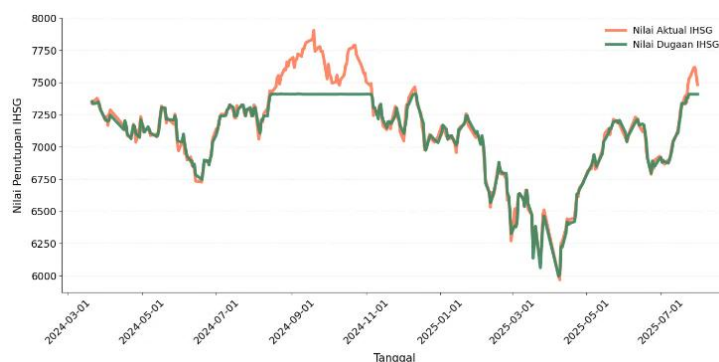
Gambar 1. Visualisasi pohon keputusan *Random Forest*

Pembangunan pohon dilakukan sampai dengan batas dari ruang *hyperparameter* tercapai dengan rata-rata MSE terkecil. Model terbaik pada metode ini mencakup *n_estimators* 100, *max_depth* sebanyak 7, *max_features* adalah *None* yakni semua variabel prediktor digunakan dalam prediksi, *min_samples_leaf* sebanyak 1, serta *min_samples_split* adalah 2. Model terbaik tersebut kemudian akan digunakan untuk memberikan nilai dugaan untuk IHSG, yang dapat dilihat pada Tabel 3

Tabel 3. Perbandingan nilai aktual dan dugaan IHSG pada *Random Forest*

Tanggal	Nilai Aktual IHSG	Nilai Dugaan IHSG
21/03/2024	7.338,35	7.351,80
22/03/2024	7.350,15	7.335,84
25/03/2024	7.377,76	7.340,25
26/03/2024	7.365,66	7.350,55
27/03/2024	7.310,09	7.332,34
⋮	⋮	⋮
25/07/2025	7.543,50	7.409,05
28/07/2025	7.614,77	7.409,05
29/07/2025	7.617,91	7.409,05
30/07/2025	7.549,89	7.409,05
31/07/2025	7.484,34	7.409,05

Berdasarkan Tabel (3) prediksi pada awal periode tidak jauh berbeda dengan nilai aktual, namun pada akhir pengamatan cenderung konstan. Kondisi ini menunjukkan keterbatasan *Random Forest* dalam merespons perubahan tajam dan melakukan ekstrapolasi terhadap pola di luar data latih. Oleh karena itu, penambahan indikator volatilitas dapat dipertimbangkan agar model lebih mampu menangkap dinamika runtun waktu finansial. Grafik perbandingan antar nilai aktual dan dugaan IHSG metode *Random Forest* terdapat pada **Gambar 2**.

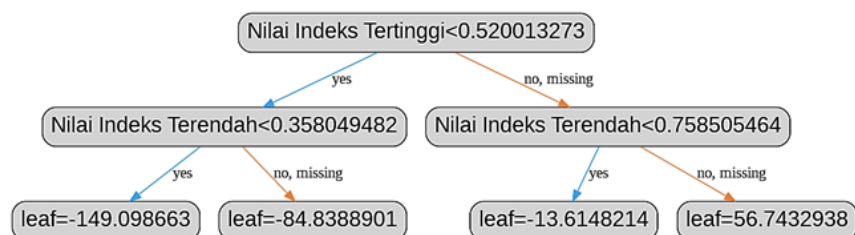


Gambar 2. Grafik perbandingan nilai aktual dan dugaan pada data uji metode *Random Forest*

Tahapan selanjutnya adalah mengevaluasi kinerja model berdasarkan nilai RMSE dan MAPE. Nilai RMSE dan MAPE dari model terbaik *Random Forest* adalah 117,54 dan 0,90% yang mengartikan bahwa nilai dugaan IHSG menggunakan metode ini cukup mendekati nilai aktual.

Sementara itu, untuk metode regresi *XGBoost* *hyperparameter* yang digunakan adalah *n_estimators* [100,200,300], *max_depth* [3;5;7], *learning_rate* [0,01;0,1;0,2], *subs_sample* [0,7;0,8;0,9], *min_child_weight* [1;2;3]. Pembangunan pohon keputusan dengan kombinasi tersebut akan dilakukan hingga akhir dan contoh dari pembangunan pohon keputusan dengan metode ini dapat dilihat pada Gambar (3). Pembangunan pohon keputusan dilakukan dengan

teknik *boosting* dan akan terus diulangi hingga semua kombinasi dijalankan. Model terbaik adalah kombinasi dari $n_estimators$ sebanyak 100, max_depth sebanyak 7, $learning_rate$ yang digunakan sebesar 0,2, $subs_sample$ sebesar 0,9, serta min_child_weight sebanyak 3. Model ini akan digunakan untuk mendapatkan nilai dugaan IHSG pada data uji yang dapat dilihat pada Tabel 4.

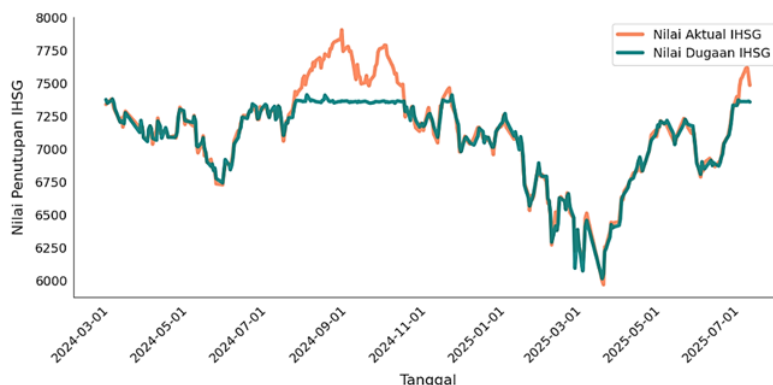


Gambar 3. Visualisasi pohon keputusan *XGBoost*

Tabel 4. Perbandingan nilai aktual dan dugaan IHSG pada *XGBoost*

Tanggal	Nilai Aktual IHSG	Nilai Dugaan IHSG
21/03/2024	7.338,35	7.370,85
22/03/2024	7.350,15	7.342,99
25/03/2024	7.377,76	7.365,54
26/03/2024	7.365,66	7.374,36
27/03/2024	7.310,09	7.355,71
⋮	⋮	⋮
25/07/2025	7.543,50	7.363,18
28/07/2025	7.614,77	7.354,31
29/07/2025	7.617,91	7.354,31
30/07/2025	7.549,89	7.357,81
31/07/2025	7.484,34	7.354,79

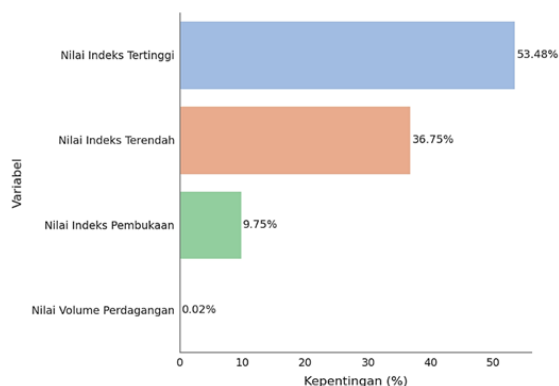
Berdasarkan hasil perbandingan di atas, dapat dilihat bahwa selisih antara data aktual dan dugaan IHSG pada *XGBoost* tidak jauh berbeda, dan untuk grafik perbandingan antar keduanya pada **Gambar 4**.



Gambar 4. Grafik perbandingan nilai aktual dan dugaan pada data uji metode *XGBoost*

Grafik tersebut menunjukkan bahwa terdapat nilai yang tidak jauh berbeda antara aktual dan dugaan, yang juga didukung oleh metrik evaluasi yakni RMSE dan MAPE berturut-turut yakni

136,45 dan 1,04% yang menyatakan bahwa metode ini sudah baik dalam memberikan nilai dugaan IHSG. Berdasarkan hasil tersebut, metode *Random Forest* menjadi model terbaik dengan nilai RMSE dan MAPE terkecil. Oleh karena itu, metode ini dipilih untuk mencari nilai *variable importance* serta melakukan peramalan IHSG 30 hari ke depan dengan nilai dugaan pada data uji. Diagram dari analisis *variable importance* dari metode *Random Forest* terdapat pada Gambar (5)



Gambar 5. *Variable Importance* pada metode *Random Forest*

Gambar 5 yang menunjukkan bahwa variabel yang memiliki pengaruh terbesar adalah indeks tertinggi (53,48%), dilanjutkan oleh indeks terendah (36,75%), indeks pembukaan (9,75%) dan volume perdagangan (0,02%). Hal ini menandakan bahwa nilai indeks yang terjadi selama perdagangan lebih memengaruhi indeks penutupan pada suatu periode dibandingkan dengan indeks awal maupun volume perdagangan yang terjadi. Langkah selanjutnya adalah mendapat nilai peramalan IHSG untuk 30 hari ke depan menggunakan *Random Forest*, yang ditunjukkan pada Tabel 5.

Tabel 5. Hasil peramalan IHSG 30 hari ke depan menggunakan *Random Forest*

No.	Tanggal	Hasil peramalan	No.	Tanggal	Hasil peramalan
1.	01/08/2025	7.409,05	16.	25/08/2025	7.364,04
2.	04/08/2025	7.375,54	17.	26/08/2025	7.330,50
3.	05/08/2025	7.340,57	18.	27/08/2025	7.350,77
4.	06/08/2025	7.358,17	19.	28/08/2025	7.394,55
5.	07/08/2025	7.394,55	20.	29/08/2025	7.402,24
6.	08/08/2025	7.323,75	21.	01/09/2025	7.406,57
7.	11/08/2025	7.289,26	22.	02/09/2025	7.407,06
8.	12/08/2025	7.342,68	23.	03/09/2025	7.376,17
9.	13/08/2025	7.322,73	24.	04/09/2025	7.371,54
10.	14/08/2025	7.337,07	25.	08/09/2025	7.373,95
11.	15/08/2025	7.305,94	26.	09/09/2025	7.378,80
12.	19/08/2025	7.366,90	27.	10/09/2025	7.334,10
13.	20/08/2025	7.337,71	28.	11/09/2025	7.333,35
14.	21/08/2025	7.363,33	29.	12/09/2025	7.348,23
15.	22/08/2025	7.348,39	30.	15/09/2025	7.376,31

Berdasarkan hasil peramalan tersebut tentunya akan dapat dihitung imbal hasil yang mungkin terjadi selama periode peramalan, Perhitungan nilai *return* dilakukan untuk semua nilai peramalan yang dapat dilihat pada Tabel 6. Nilai ekspektasi pengembalian (*expected*

return) berdasarkan hasil peramalan adalah -0,00014, yang menunjukkan bahwa rata-rata *return* diperkirakan bernilai negatif sebesar 0,00014.

Value at Risk simulasi historis dapat dihitung dengan data *return* yang sudah diurutkan dari terburuk ke terbaik dengan tingkat kepercayaan tertentu. Pada taraf signifikansi 10% data yang digunakan data *return* terburuk ketiga (-0,0047), 5% menggunakan data kedua (-0,0061), sedangkan untuk 1% akan menggunakan nilai *return* terburuk yakni data pertama (-0,0096). Data tersebut akan digunakan untuk menghitung VaR simulasi historis dengan dana investasi awal adalah Rp 100.000.000,00 apabila mengikuti IHSG sebagai tolak ukur dalam investasi. Taraf signifikansi 10%, VaR bernilai Rp 470.000,00, VaR pada taraf signifikansi 5% adalah Rp 610.000,00, serta VaR pada taraf signifikansi 1% adalah Rp 960.000,00. Nilai-nilai tersebut menunjukkan semakin tinggi tingkat kepercayaan, maka semakin besar pula estimasi potensi kerugian maksimum yang mungkin dialami investor. Secara keseluruhan, peramalan IHSG memberikan gambaran dari IHSG untuk periode kedepannya, sedangkan VaR menjadi indikator untuk memantau kerugian pada tingkat kepercayaan tertentu.

Tabel 6. Nilai *return* dari hasil peramalan 30 hari IHSG pada *Random Forest*

No.	Tanggal	Return IHSG	No.	Tanggal	Return IHSG
1.	01/08/2025	-	16.	25/08/2025	0,0021
2.	04/08/2025	-0,0045	17.	26/08/2025	-0,0046
3.	05/08/2025	-0,0047	18.	27/08/2025	0,0028
4.	06/08/2025	0,0024	19.	28/08/2025	0,0060
5.	07/08/2025	0,0049	20.	29/08/2025	0,0010
6.	08/08/2025	-0,0096	21.	01/09/2025	0,0006
7.	11/08/2025	-0,0047	22.	02/09/2025	0,0001
8.	12/08/2025	0,0073	23.	03/09/2025	-0,0042
9.	13/08/2025	-0,0027	24.	04/09/2025	-0,0006
10.	14/08/2025	0,0020	25.	08/09/2025	0,0003
11.	15/08/2025	-0,0042	26.	09/09/2025	0,0007
12.	19/08/2025	0,0083	27.	10/09/2025	-0,0061
13.	20/08/2025	-0,0040	28.	11/09/2025	-0,0001
14.	21/08/2025	0,0035	29.	12/09/2025	0,0020
15.	22/08/2025	-0,0020	30.	15/09/2025	0,0038

5. KESIMPULAN

Penelitian ini menunjukkan bahwa metode *Random Forest* memiliki performa lebih baik dibandingkan *Extreme Gradient Boosting (XGBoost)* berdasarkan evaluasi RMSE dan MAPE yang lebih rendah dalam memprediksi nilai penutupan IHSG. Analisis *variable importance* menunjukkan kontribusi terbesar terhadap hasil prediksi adalah nilai tertinggi dan terendah IHSG yang terjadi selama proses perdagangan, sedangkan volume perdagangan memberikan kontribusi yang relatif kecil. Selain itu, hasil peramalan menggunakan model *Random Forest* dapat dimanfaatkan untuk mengestimasi risiko investasi pada tingkat kepercayaan tertentu melalui metode *Value at Risk* (VaR). Penelitian selanjutnya disarankan menggunakan periode data yang lebih panjang, menambahkan fitur *lag*, *return*, volatilitas, serta variabel makroekonomi seperti nilai tukar rupiah serta indeks saham asing.

DAFTAR PUSTAKA

- (KSEI), K. S. E. I., 2025. *Statistik Pasar Modal Indonesia Periode Juli 2025*, Jakarta: Kustodian Sentral Efek Indonesia.
- Agus Ambarwari, Q. J. A. Y. H., 2020. Analisis Pengaruh Data Scalling Terhadap Performa Algoritma Machine Learning untuk Identifikasi Tanaman. *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4(1), pp. 117-122.
- Astutiningsih, T., Saputro, D. R. S. & S., 2023. Optimasi Algoritma eXtreme Gradient Boosting (XGBoost) pada Harga Saham PT. United Tractors Tbk.. *SPECTA Journal of Technology*, VII(3), pp. 632-641.
- Baihaqi, M. I., D., Syaripudin, A. & Nugroho, F. A., 2023. Implementasi Algoritma Random Forest pada Prediksi Harga Saham Berdasarkan Data Historis. *Jubitek: Jurnal Biga Data dan Teknologi informasi*, I(3), pp. 42-51.
- Breiman, L., 2001. Random Forest. *Machine Learning*, 45(1), pp. 5-32.
- Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *KDD*.
- Ernawati, F. R. P. d. I., 2023. Perbandingan Algoritma Extreme Gradient Boosting Dan Random Forest Untuk Memprediksi Indeks Harga Terendah Saham Dengan Indeks ISSI. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, pp. 438-449.
- Hartono, 2024. *Modul Digital Machine Learning*. Jakarta: LMS-SPADA INDONESIA.
- Indonesia Stock Exchange, 2008. *Buku Panduan Indeks Harga Saham Bursa Efek Indonesia*. Jakarta: Indonesia Stock Exchange.
- Lardy, M. M. d. J., 2019. Credit Spread Approximation and Improvement using Random Forest Regression. *European Journal of Operational Research*, 277(1), pp. 351-365.
- Patriya, E., 2020. Implementasi Support Vector Machine pada Prediksi Indeks Harga Saham Gabungan (IHSG). *Jurnal Ilmiah Teknologi dan Rekayasa Volume*, XXV(1), pp. 24-38.
- Priyatno, A. M. et al., 2023. Comparison Random Forest Regression and Linear Regression For Forecasting BBKA Stock Price. *JUTIN : Jurnal Teknik Industri Terintegrasi*, VI(3), pp. 718-732.
- Rafirhan, M. N., Kaniaingsih, A. & Komarudin, A., 2023. Peramalan Indeks Harga Saham PT XYZ Menggunakan Metode Double Moving Average (DMA). *IJESPG Journal*, I(3), pp. 183-189.
- Solihatun, A. et al., 2023. Perhitungan Value at Risk(Var) Pada Portofolio Saham IDX Sektor Keuangan (IDXFINANCE) Menggunakan Metode Simulasi Historis (Historical Simulation Method) . *Jurnal Matematika, Komputasi dan Statistika*, III(1), pp. 245-254.
- Zhu, N. et al., 2022. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm. *MDPI*, XII(20), pp. 1-16.