JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

Evaluasi Algoritma Metode Machine Learning untuk Memprediksi Harga Saham

Arif Yoga Pratama[1] , Deka Agustiar[1] , Dimas Bayu Samudra Ajie[1] , Intan Tri Yulianti[1] , Martuah Pardamean Wijaya Siahaan[1] , Shalihah Bulan Cinta[1]

[1)]Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Jl. Raya Puspiptek No. 46, Kel. Buaran, Kec. Serpong, Kota Tangerang Selatan. Banten 15310, Indonesia
Email: arifyogapratama69@gmail.com , dekaagustiar@gmail.com , dimas.samudra51@gmail.com , intantriylti@gmail.com , jayawijaya44633@gmail.com , bulancinta98@gmail.com

Abstrak- Penelitian ini mengevaluasi kinerja beberapa algoritma machine learning, yaitu Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), dan Linear Regression, dalam menganalisis data historis saham Bank Negara Indonesia. Penilaian dilakukan melalui analisis klasifikasi dan regresi dengan menggunakan metrik seperti akurasi, Mean Squared Error (MSE), dan R- squared. Logistic Regression menunjukkan performa terbaik dalam tugas klasifikasi dengan akurasi 60%, meskipun memiliki kelemahan dalam memprediksi kelas "Increase". Untuk regresi, algoritma Linear Regression tidak efektif, dengan nilai R-squared -0.02 dan MSE 0.24. Visualisasi data menunjukkan korelasi tinggi antar fitur harga, serta korelasi moderat negatif antara volume perdagangan dan harga saham. Studi ini menyarankan penelitian lebih lanjut untuk mengeksplorasi dampak variabel volume perdagangan terhadap prediksi harga saham.

Kata kunci: Machine Learning, Prediksi Harga Saham, Logistic Regression, Random Forest, Linear Regression, Data Visualisasi, Volume Perdagangan

Abstract- *This study evaluates the performance of several machine learning algorithms, namely Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Linear Regression, in analyzing the historical stock data of Bank Negara Indonesia. The evaluation was conducted through classification and regression analysis using metrics such as accuracy, Mean Squared Error (MSE), and R-squared. Logistic Regression demonstrated the best performance for classification tasks with an accuracy of 60%, although it had weaknesses in predicting the "Increase" class. For regression, the Linear Regression algorithm was ineffective, with an R-squared value of -0.02 and an MSE of 0.24. Data visualization revealed high correlations among price features and a moderate negative correlation between trading volume and stock prices. This study suggests further research to explore the impact of trading volume variables on stock price predictions.*

Keywords: Machine Learning, Stock Price Prediction, Logistic Regression, Random Forest, Linear Regression, Data Visualization, Trading Volume

1. PENDAHULUAN

Pasar saham merupakan salah satu pilar penting dalam ekonomi global dan lokal, memberikan peluang investasi sekaligus tantangan yang kompleks dalam pengelolaan dan analisis data. Prediksi harga saham telah menjadi topik utama dalam dunia keuangan karena memiliki dampak signifikan terhadap pengambilan keputusan investasi. Namun, volatilitas pasar saham yang tinggi dan ketergantungan pada berbagai faktor eksternal, seperti kondisi ekonomi global, kebijakan moneter, dan sentimen pasar, menjadikan prediksi harga saham tugas yang sangat menantang.

Dalam era digital saat ini, kemajuan teknologi telah memungkinkan penggunaan algoritma machine learning untuk menganalisis dan memprediksi pergerakan harga saham. Algoritma ini menawarkan pendekatan berbasis data untuk mengidentifikasi pola dan hubungan yang kompleks dalam dataset yang besar. Namun, tidak semua algoritma machine learning memiliki performa yang sama dalam menghadapi berbagai jenis data dan tugas prediksi.

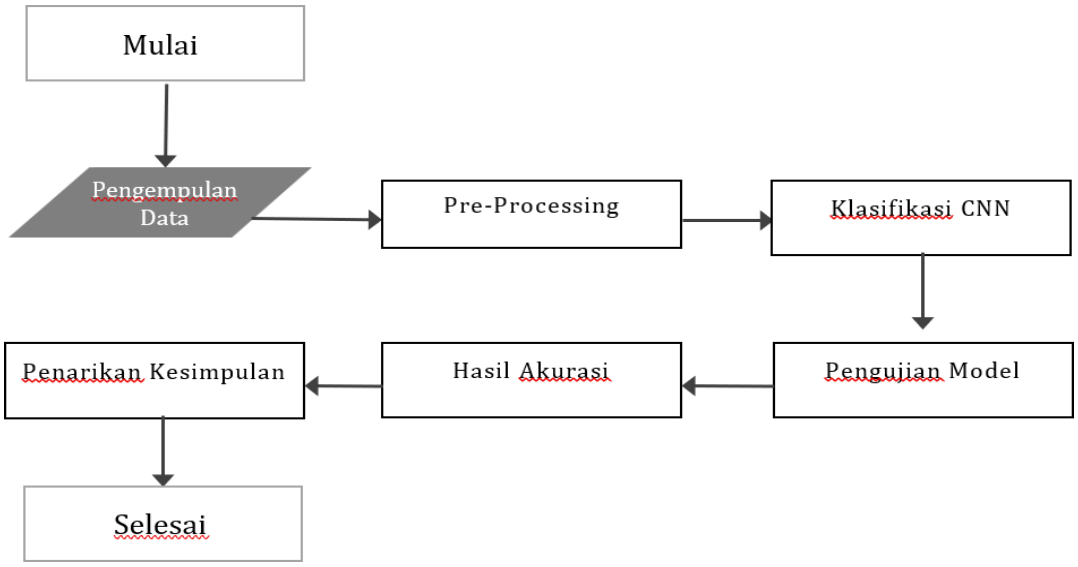
Penelitian ini berfokus pada penggunaan data historis saham Bank Negara Indonesia sebagai kasus studi untuk mengevaluasi kinerja berbagai algoritma machine learning, termasuk Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors (KNN), dan Linear Regression.

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

Dengan menganalisis data yang mencakup harga saham harian dan volume perdagangan, penelitian ini bertujuan untuk memahami kelebihan dan kelemahan masing-masing algoritma dalam tugas klasifikasi dan regresi. Selain itu, kami juga mengeksplorasi hubungan antar variabel dalam dataset untuk memberikan wawasan yang lebih mendalam tentang faktor-faktor yang memengaruhi prediksi harga saham.

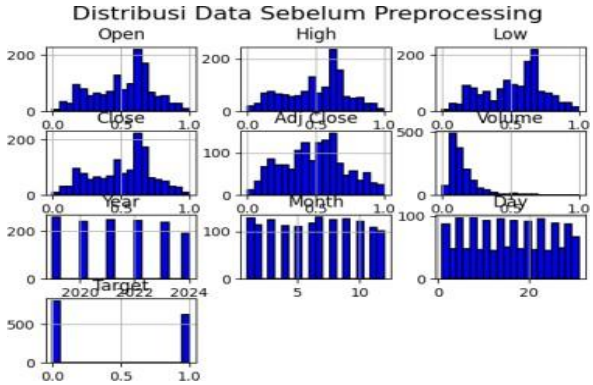
2. METODOLOGI

Penelitian ini menggunakan data historis saham Bank Negara Indonesia yang mencakup fitur numerik seperti harga pembukaan (Open), harga tertinggi (High), harga terendah (Low), harga penutupan (Close), harga penutupan yang disesuaikan (Adj Close), volume perdagangan (Volume), serta fitur waktu (Date). Data yang diperoleh diproses dalam beberapa tahap berikut:



a. *Pengumpulan Data*

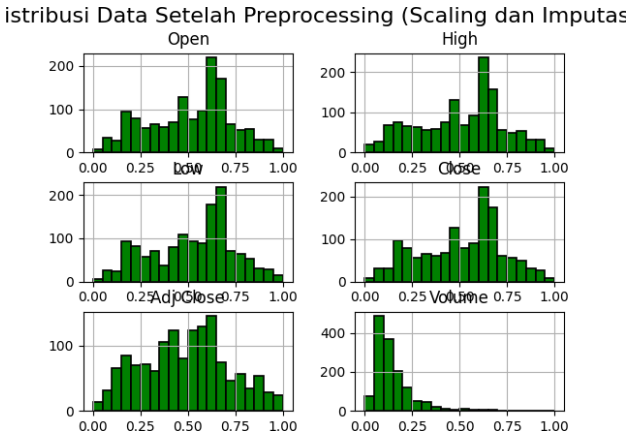
Data diambil dari sumber terpercaya dan meliputi periode tertentu untuk memastikan kualitas serta relevansi dalam analisis.



JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

b. Preprocessing Data

Tahap preprocessing mencakup beberapa langkah penting. Normalisasi dilakukan menggunakan MinMaxScaler untuk menyelaraskan skala nilai fitur numerik ke dalam rentang 0 hingga 1. Penanganan data hilang juga diterapkan dengan mengisi atau menghapus entri yang tidak lengkap guna menjaga integritas data. Selain itu, jika diperlukan, fitur kategori dikonversi menjadi nilai numerik menggunakan metode seperti one-hot encoding untuk memastikan data dapat digunakan oleh model



machine learning.

c. Pemilihan Model

Model yang diuhi meliputi Logistic Regression, Random Forest, Decision Tree, K- Nearst Neighbors (KNN), dan Linear Regression. Pemilihan algoritma-algoritma ini didasarkan pada kemampuan mereka dalam menangani tugas klasifikasi dan regresi, sehingga dapat memberikan wawasan yang lebih mendalam tentang performa masing-masing model dalam analisis data yang dilakukan.

d. Evaluasi Model

Evaluasi model dilakukan menggunakan metrik yang relevan untuk masing-masing tugas. Untuk tugas klasifikasi, metrik seperti akurasi, precision, recall, dan confusion matrix digunakan untuk mengevaluasi performa model dalam mengklasifikasikan data dengan tepat. Sementara itu, untuk tugas regresi, indikator utama yang digunakan adalah Mean Squared Error (MSE) dan R-squared, yang bertujuan untuk mengukur seberapa baik model dapat memprediksi nilai numerik berdasarkan data yang ada.

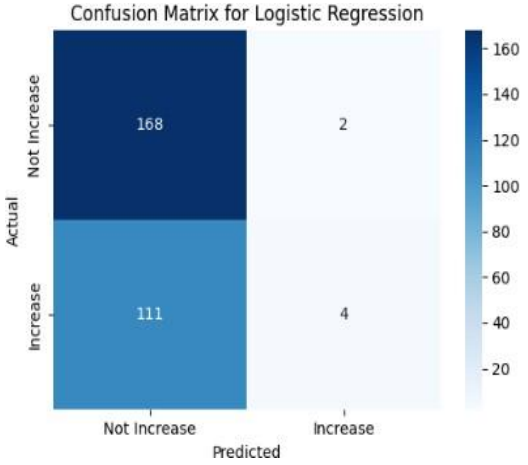
e. *Visualisasi Data dan Hasil*

Visualisasi dilakukan menggunakan heatmap untuk menunjukkan korelasi antar fitur, serta scatter plot untuk membandingkan nilai aktual dengan nilai prediksi dari model regresi. Diagram tambahan, seperti histogram distribusi data, juga digunakan untuk memberikan wawasan lebih mendalam terkait pola data sebelum dan sesudah preprocessing.

3. HASIL DAN DISKUSI

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

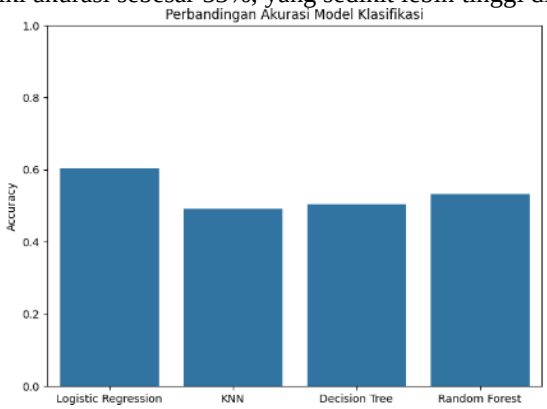
a. Logistic Regression



Logistic Regression mencapai akurasi 60%, yang berarti model ini berhasil membuat prediksi yang benar untuk 60% data yang diuji. Model ini menunjukkan keunggulan dalam memprediksi kelas "Not Increase" (harga saham tidak meningkat), yang mencerminkan kemampuan Logistic Regression dalam mengenali pola data yang merepresentasikan stabilitas atau penurunan harga saham. Namun, model ini memiliki kelemahan signifikan dalam memprediksi kelas "Increase". Sebanyak 111 kasus kelas "Increase" secara keliru diklasifikasikan sebagai "Not Increase". Kelemahan ini kemungkinan disebabkan oleh ketidakseimbangan data, di mana data kelas "Increase" lebih sedikit, atau ketidakmampuan model dalam menangkap pola kompleks yang menunjukkan kenaikan harga saham.

b. Random Forest

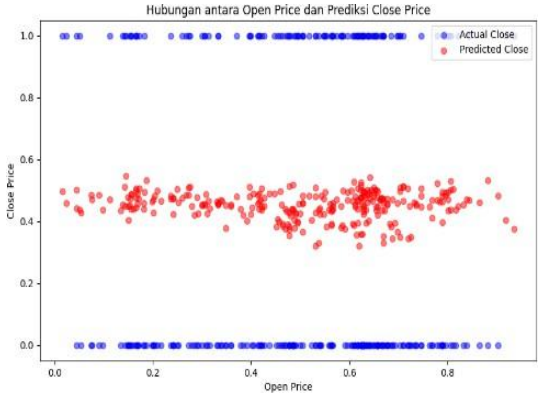
Model Random Forest memiliki akurasi sebesar 53%, yang sedikit lebih tinggi dibandingkan Decision



Tree (51%) namun masih lebih rendah dibandingkan Logistic Regression (60%). Sebagai model ensambel, Random Forest memiliki keunggulan dalam menggabungkan beberapa Decision Tree untuk meningkatkan stabilitas dan akurasi prediksi. Walaupun akurasinya tidak terlalu tinggi, model ini lebih tahan terhadap overfitting dibandingkan Decision Tree. Namun, kelemahan Random Forest terletak pada akurasinya yang hanya sedikit lebih baik dari Decision Tree, yang dapat mengindikasikan bahwa pengaturan hyperparameter seperti jumlah pohon atau kedalaman pohon belum dioptimalkan. Selain itu, Random Forest kurang efisien untuk data yang memiliki variabel kategori yang kompleks, sehingga perlu perhatian lebih dalam penggunaannya.

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

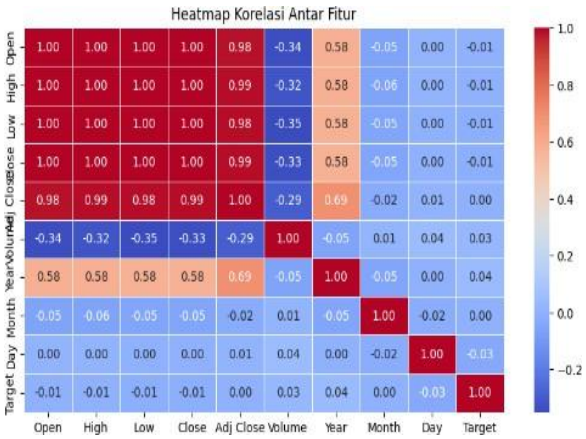
c. Linear Regression



Evaluasi regresi menunjukkan hasil yang kurang memuaskan. Nilai R-squared sebesar

-0.02 mengindikasikan bahwa model Linear Regression tidak mampu menjelaskan variabilitas data, yang berarti model ini hampir tidak memiliki kemampuan untuk memprediksi harga saham yang benar berdasarkan data input. Selain itu, nilai MSE (Mean Squared Error) sebesar 0.24 menunjukkan bahwa kesalahan rata-rata kuadrat dari prediksi cukup besar, mencerminkan performa buruk model dalam memprediksi nilai numerik.

d. Visualisasi Data



Analisis korelasi antar fitur harga menunjukkan bahwa fitur seperti Open, High, Low, Close, dan Adj Close memiliki korelasi yang sangat tinggi satu sama lain. Ini berarti nilai-nilai tersebut saling terkait erat dan cenderung memiliki pola yang serupa, yang mengindikasikan bahwa beberapa variabel ini mungkin redundant karena memberikan informasi yang mirip. Untuk mengatasi hal ini, teknik seperti Principal Component Analysis (PCA) dapat digunakan untuk mengurangi dimensi data. Sementara itu, volume perdagangan menunjukkan korelasi negatif moderat dengan harga saham, yang berarti bahwa ketika volume perdagangan meningkat, harga saham cenderung turun meskipun tidak sepenuhnya linear. Korelasi negatif ini menjadikan volume perdagangan sebagai variabel penting yang dapat digunakan untuk memprediksi harga saham, dan perlu dilakukan analisis lebih lanjut untuk mengeksplorasi pola-pola tersembunyi dalam data.

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1828-1833

4. KESIMPULAN

Secara keseluruhan, analisis yang dilakukan menunjukkan bahwa Logistic Regression memberikan hasil terbaik dalam tugas klasifikasi dibandingkan dengan model lainnya. Model ini menunjukkan kemampuan yang lebih baik dalam mengklasifikasikan data dan menghasilkan prediksi yang lebih akurat. Hal ini menegaskan bahwa Logistic Regression lebih efektif dalam menangani masalah klasifikasi dalam dataset ini. Namun, untuk prediksi nilai numerik, terutama dalam hal prediksi harga saham, Linear Regression tidak memberikan hasil yang memadai. Evaluasi model menunjukkan bahwa Linear Regression gagal dalam menghasilkan prediksi yang akurat, dengan nilai R- squared yang sangat rendah dan Mean Squared Error (MSE) yang tinggi, yang mencerminkan kesalahan yang besar dalam prediksi harga saham berdasarkan data input yang ada.

Selain itu, analisis korelasi antar fitur harga seperti Open, High, Low, Close, dan Adj Close menunjukkan adanya korelasi yang sangat tinggi di antara fitur-fitur tersebut. Hal ini mengindikasikan bahwa variabel-variabel tersebut cenderung memberikan informasi yang mirip atau redundant, yang dapat memengaruhi kualitas dan kinerja model. Untuk menangani masalah ini, pendekatan seperti Principal Component Analysis (PCA) dapat diterapkan untuk mengurangi dimensi data, sehingga informasi yang berlebihan bisa dihilangkan tanpa mengorbankan kualitas prediksi model. Oleh karena itu, mengatasi multikolinearitas antar fitur harga menjadi hal yang penting untuk meningkatkan kinerja model prediktif.

Selanjutnya, volume perdagangan menunjukkan korelasi negatif moderat dengan harga saham, yang berarti bahwa ketika volume perdagangan meningkat, harga saham cenderung turun, meskipun hubungan ini tidak sepenuhnya linear. Korelasi negatif ini menunjukkan bahwa volume perdagangan bisa menjadi faktor prediktif yang signifikan dalam memodelkan harga saham. Meskipun demikian, hubungan antara volume perdagangan dan harga saham memerlukan analisis lebih lanjut untuk memahami pola- pola yang lebih mendalam dan potensi pengaruhnya terhadap fluktuasi harga saham. Penelitian lebih lanjut diperlukan untuk mengeksplorasi faktor-faktor lain yang mungkin memengaruhi hubungan ini dan untuk mengoptimalkan penggunaan volume perdagangan sebagai variabel dalam model prediksi harga saham.

Secara keseluruhan, meskipun Logistic Regression terbukti efektif dalam klasifikasi, masih ada tantangan besar dalam memprediksi harga saham secara akurat dengan model regresi. Untuk meningkatkan performa model prediksi harga saham, pendekatan lebih lanjut seperti pengurangan dimensi melalui PCA dan eksplorasi lebih dalam terhadap volume perdagangan serta variabel lainnya sangat diperlukan. Hal ini akan membantu dalam mengidentifikasi faktor-faktor penting yang dapat meningkatkan akurasi prediksi harga saham di masa depan.

DAFTAR PUSTAKA

Brownlee, J. (2020). *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End*. Machine Learning Mastery.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kelleher, J. D., Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.

Khan, R., & Shah, N. A. (2021). Stock Price Prediction Using Machine Learning Techniques: A Comparative Study. *Journal of Financial Analytics and Machine Learning*, 4(1), 34–45.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Yeo, I. K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87(4), 954–959.

Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998). Forecasting with Artificial Neural Networks: The State of the Art. *International Journal of Forecasting*, 14(1), 35– 62.