ISSN: xxxx-xxxx
Prosiding SAINTEK: Sains dan Teknologi Vol.1 No.1 Tahun 2022
Call for papers dan Seminar Nasional Sains dan Teknologi Ke-1 2022
Fakultas Teknik, Universitas Pelita Bangsa, Juli 2022

Analisis Prediksi Harga Saham pada IDX Composite Menggunakan Model Machine Learning dengan Apache Spark (PySpark)

Stock Price Prediction Analysis on IDX Composite Using Machine Learning Model with Apache Spark (PySpark)

Dahyoung Yenuargo[1]
[1]Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]dahyoung8@gmail.com

Abstract

This abstract discusses the application of machine learning methods, such as Decision Tree, Random Forest, and Logistic Regression, using the Apache Spark (PySpark) platform to predict stock price movements in the JCI. The purpose of this study is to evaluate the performance of the three models and identify the contribution of PySpark in improving prediction accuracy. Historical data of JCI from January 1, 2019 to December 31, 2023 is used in this study. After preprocessing, the data was divided into training data (80%) and testing data (20%). The model analysis results show that Decision Tree, Random Forest, and Logistic Regression provide stock price predictions with low Root Mean Squared Error (RMSE) values, high R-squared, and accuracy above 99%. The visualization graph also shows the superiority of the model in predicting stock price movements. The Logistic Regression model achieved optimal accuracy with an Area Under ROC (AUC) value of 1.0. The use of PySpark as a distributed data processing platform proved to be efficient in managing large volumes of data. In conclusion, this study shows that the implementation of machine learning models using PySpark on IDX Composite data can significantly improve the accuracy of predicting stock price movements. The practical implications of this research can help investors, financial analysts, and capital market stakeholders in making investment decisions.

***Keywords**: machine learning, stock price prediction, PySpark, IDX Composite, stock market analysis.*

Abstrak

Abstrak ini membahas penerapan metode *machine learning*, seperti *Decision Tree*, *Random Forest*, dan *Logistic Regression*, menggunakan *platform Apache Spark (PySpark)* untuk memprediksi pergerakan harga saham di *IDX Composite*. Tujuan penelitian ini adalah untuk mengevaluasi kinerja ketiga model tersebut dan mengidentifikasi kontribusi *PySpark* dalam meningkatkan akurasi prediksi. Data historis *IDX Composite* dari periode 1 Januari 2019 hingga 31 Desember 2023 digunakan dalam penelitian ini. Setelah proses *preprocessing*, data dibagi menjadi data pelatihan (80%) dan data pengujian (20%). Hasil analisis model menunjukkan bahwa *Decision Tree*, *Random Forest*, dan *Logistic Regression* memberikan prediksi harga saham dengan nilai *Root Mean Squared Error (RMSE)* yang rendah, *R-squared* tinggi, dan akurasi di atas 99%. Grafik visualisasi juga mengindikasikan keunggulan model dalam memprediksi pergerakan harga saham. Model *Logistic Regression* mencapai tingkat akurasi optimal dengan nilai *Area Under ROC (AUC)* sebesar 1.0. Penggunaan *PySpark* sebagai *platform* pengolahan data distribusi terbukti efisien dalam mengelola *volume* data besar. Kesimpulannya, penelitian ini menunjukkan bahwa implementasi model *machine learning* menggunakan *PySpark* pada data *IDX Composite* dapat meningkatkan akurasi prediksi pergerakan harga saham secara signifikan. Implikasi praktis dari penelitian ini dapat membantu para investor, analis keuangan, dan pemangku kepentingan pasar saham dalam pengambilan keputusan investasi yang lebih informatif dan efektif.

Kata kunci: machine learning, prediksi harga saham, PySpark, IDX Composite, analisis pasar saham.

Pendahuluan

Pasar saham merupakan lingkungan keuangan yang dinamis dan sangat dipengaruhi oleh berbagai faktor ekonomi, politik, dan sosial [1][2][3]. Para investor dan pelaku pasar sering kali dihadapkan pada tugas yang kompleks dalam membuat keputusan investasi yang cerdas [4][5]. Dalam konteks ini, prediksi harga saham menjadi sangat penting untuk membantu para pemangku kepentingan membuat keputusan investasi yang informasional dan berbasis data [6][7][8]. *IDX Composite (JKSE)* sebagai representasi pasar saham Indonesia memiliki peran strategis dalam memberikan gambaran performa ekonomi nasional [9][10][11].

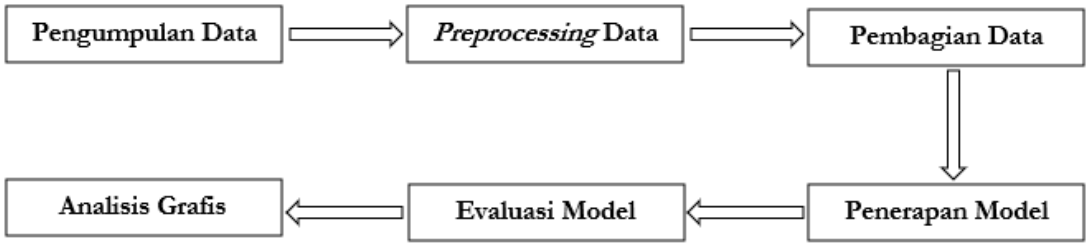
Pendekatan *machine learning*, seperti *Decision Tree*, *Random Forest*, dan *Logistic Regression*, yang diimplementasikan menggunakan *PySpark*, menjadi pilihan dalam mengeksplorasi perilaku pasar saham [12][13][14]. Tinjauan literatur singkat ini menggali penelitian-penelitian terdahulu, menyoroti kelebihan dan kelemahan setiap model, serta mengidentifikasi kesenjangan pengetahuan yang masih perlu diisi. Penelitian ini membawa kontribusi melalui penerapan inovatif terhadap model-model tersebut pada *IDX Composite*.

Pentingnya prediksi harga saham tidak hanya berkaitan dengan keputusan investasi, tetapi juga menjadi kunci bagi pemahaman dinamika pasar [15]. Penelitian ini dilakukan untuk menjawab kebutuhan tersebut, khususnya dalam konteks pasar saham Indonesia. Dengan menggunakan metode *Decision Tree*, *Random Forest*, dan *Logistic Regression*, penelitian ini bertujuan untuk menyajikan pendekatan analitis yang lebih canggih dalam meramalkan pergerakan harga saham pada *IDX Composite*.

Tujuan utama penelitian ini adalah mendapatkan pemahaman mengenai kinerja model *Decision Tree*, *Random Forest*, dan *Logistic Regression* yang diterapkan menggunakan *PySpark* dalam meramalkan pergerakan harga saham di *IDX Composite*. Peneliti berharap hasil penelitian ini dapat memberikan wawasan yang berharga kepada investor, analis keuangan, dan pihak-pihak yang berkepentingan di pasar saham, sehingga mereka dapat membuat keputusan investasi yang lebih informatif dan efektif. Oleh karena itu, penelitian ini diharapkan mampu memberikan kontribusi signifikan dalam pengembangan metode analisis pasar saham Indonesia yang lebih maju dan dapat diandalkan.

Metode Penelitian

Penelitian ini menerapkan metode *machine learning* untuk menganalisis dan membandingkan kinerja tiga model prediksi pergerakan harga saham *IDX Composite*, yaitu *Decision Tree*, *Random Forest*, dan *Logistic Regression* [16]. Pendekatan ini memfasilitasi analisis yang teliti terhadap ketiga model tersebut dalam pergerakan harga saham. Langkah-langkah penelitian dapat dilihat pada gambar 1.



Gambar 1 Tahapan Penelitian

Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari *website finance.yahoo.com,* dapat diakses dengan link https://finance.yahoo.com/quote/%5EJKSE/history?p=%5EJKSE. Rentang waktu pengumpulan data mencakup periode dari 1 Januari 2019 hingga 31 Desember 2023. Data yang diambil mencakup harga pembukaan *(Open)*, harga tertinggi *(High)*, harga terendah *(Low)*, harga penutupan *(Adj Close)*, dan *volume* perdagangan *(Volume)*, yang menjadi dasar bagi analisis prediksi harga saham menggunakan model *machine learning*.

Preprocessing **Data**

Data yang diunduh kemudian melalui proses *preprocessing* untuk membersihkan nilai-nilai yang hilang dan memilih fitur-fitur yang relevan. Pengorganisasian data dilakukan agar sesuai dengan format yang dibutuhkan untuk analisis lebih lanjut.

Pembagian Data

Dataset diterapkan pembagian menjadi dua bagian, data pelatihan (80%) dan data pengujian (20%). Hal ini dilakukan untuk melatih model pada dataset historis dan menguji kinerjanya pada data yang belum pernah diakses sebelumnya.

Penerapan Model

Tiga model prediksi harga saham, yaitu *Decision Tree*, *Random Forest*, dan *Logistic Regression*, diimplementasikan menggunakan *PySpark*. *Decision Tree* dan *Random Forest* menggunakan label berdasarkan perbandingan harga penutupan dan pembukaan, sedangkan *Logistic Regression* digunakan untuk klasifikasi biner.

Evaluasi Model

Performa model *Decision Tree* dan *Random Forest* diukur menggunakan metrik *Root Mean Squared Error (RMSE)* untuk mengevaluasi sejauh mana prediksi mendekati nilai sebenarnya. *Model Logistic Regression* dievaluasi dengan metrik *Area Under ROC (AUC)* untuk mengukur kemampuan klasifikasi biner.

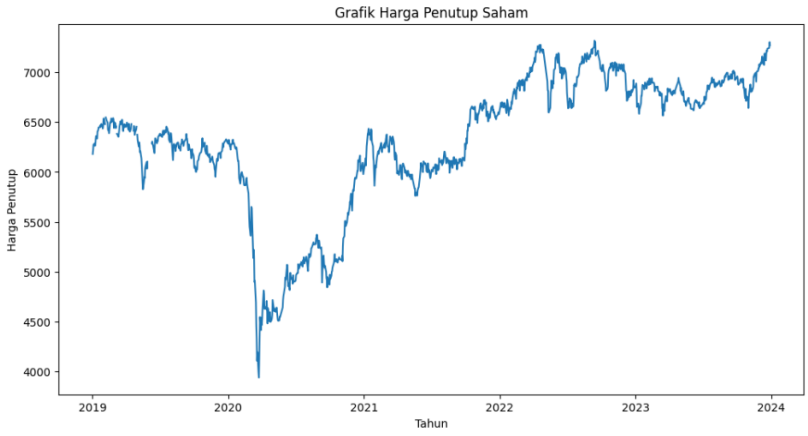
Analisis Grafis

Grafik digunakan untuk menyajikan visualisasi terhadap performa model. Visualisasi ini memberikan gambaran yang jelas mengenai sejauh mana model-model tersebut memprediksi pergerakan harga saham dengan baik.

Hasil dan Pembahasan

Analisis Grafis

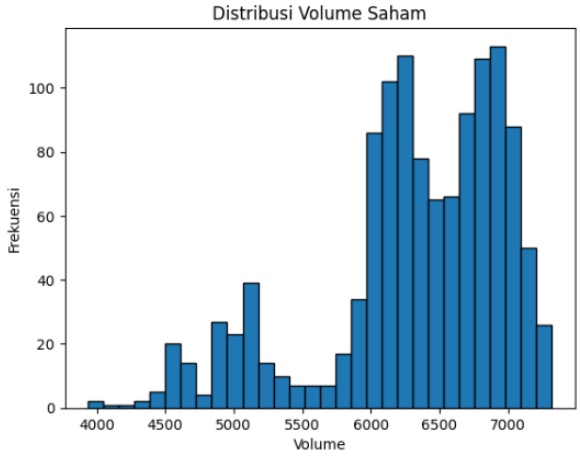
Grafik Harga Saham



Gambar 2 Grafik Harga Penutup Saham

Grafik pada gambar 2 yang menampilkan harga penutup saham memberikan representasi visual mengenai tren pergerakan harga saham *IDX Composite* selama periode penelitian. Dinamika pergerakan harga tersebut memberikan konteks yang penting terhadap hasil prediksi model.

Distribusi *Volume*



Gambar 3 Distribusi *Volume* Saham

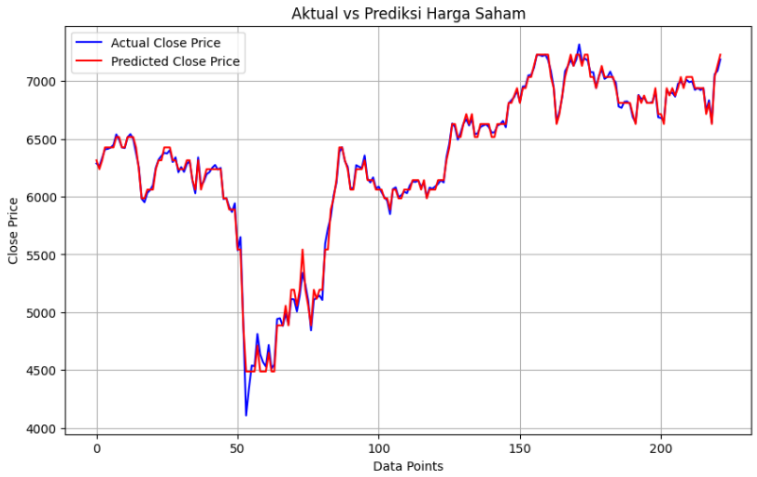
Grafik pada gambar 3, yang menggambarkan distribusi volume saham, memberikan pemahaman tentang sebaran *volume* perdagangan selama periode penelitian. Analisis distribusi volume ini dapat menjadi elemen kunci dalam mengartikan hasil prediksi model.

Perfroma Model

Decision Tree

Tabel 1 Performa Model *Decision Tree*

No	Metrik	Nilai
1	*Root Mean Squared Error (RMSE)*	47.46
2	*R-squared*	0.9942
3	Akurasi Relatif	99.97%



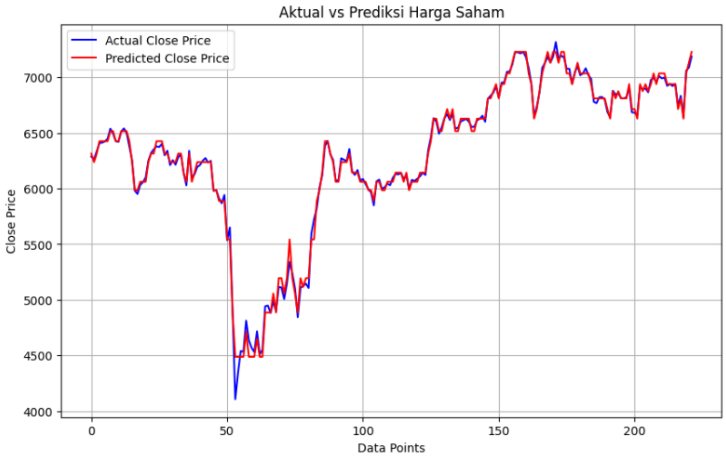
Gambar Aktual vs Prediksi Harga Saham

Model *Decision Tree* menunjukkan nilai *Root Mean Squared Error (RMSE)* sebesar 47.46 dan *R-squared* sebesar 0.9942. Akurasi relatif model mencapai 99.97%. Grafik pada gambar 4 memberikan representasi visual mengenai seberapa dekat hasil prediksi dengan nilai sebenarnya.

Random Forest

Tabel 2 Performa Model *Random Forest*

No	Metrik	Nilai
1	*Root Mean Squared Error (RMSE)*	43.90
2	*R-squared*	0.9957
3	Akurasi Relatif	99.95%



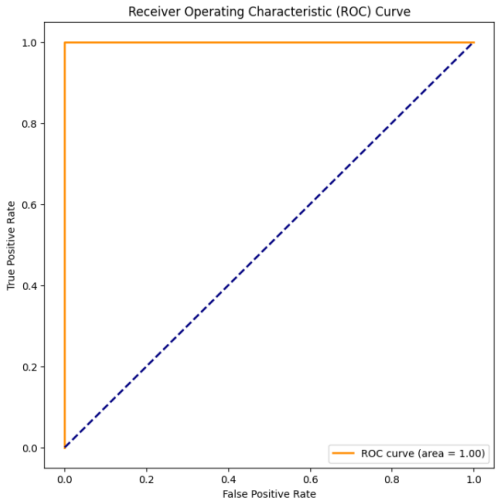
Gambar 5 Aktual vs Prediksi Harga Saham

Model *Random Forest* menunjukkan peningkatan kinerja dengan *Root Mean Squared Error (RMSE)* sebesar 43.90 dan *R-squared* sebesar 0.9957. Akurasi relatif model *Random Forest* mencapai 99.95%. Grafik pada gambar 5 memberikan representasi visual mengenai seberapa dekat hasil prediksi dengan nilai sebenarnya.

Logistic Regression

Tabel 3 Performa Model *Logistic Regression*

No	Metrik	Nilai
1	*Area Under ROC (AUC)*	1.0



Gambar 5 Grafik *Receiver Operating Characteristic (ROC)*

Model *Logistic Regression* mencapai tingkat akurasi optimal dengan nilai *Area Under ROC (AUC)* sebesar 1.0. Grafik *Receiver Operating Characteristic (ROC)* pada gambar 6 memberikan representasi visual terhadap kemampuan model dalam mengklasifikasikan pergerakan harga saham.

Pembahasan

Hasil analisis dan visualisasi menunjukkan bahwa ketiga model (*Decision Tree*, *Random Forest*, dan *Logistic Regression*) secara signifikan unggul dalam memprediksi pergerakan harga saham pada *IDX Composite*. Grafik Aktual vs Prediksi Harga Saham memberikan representasi visual tentang seberapa dekat prediksi dengan nilai sebenarnya, sedangkan grafik harga penutup saham dan distribusi *volume* memberikan konteks pasar yang luas.

Efektivitas model dalam memberikan prediksi yang akurat, tercermin dalam nilai-nilai *Root Mean Squared Error (RMSE)* yang rendah, *R-squared* yang tinggi, dan tingkat akurasi di atas 99%, menjadikannya sebagai alat yang sangat berguna dalam mendukung pengambilan keputusan di pasar saham. Dengan analisis grafis yang kuat, penelitian ini membentuk dasar yang kuat untuk strategi prediksi harga saham yang efektif dan relevan di pasar saham *IDX Composite*.

Kesimpulan

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa penerapan model *Decision Tree*, *Random Forest*, dan *Logistic Regression* menggunakan *Apache Spark (PySpark)* pada data historis *IDX Composite* secara signifikan meningkatkan akurasi prediksi pergerakan harga saham. Ketiga model ini menunjukkan kinerja yang sangat baik, terbukti dari nilai *Root Mean Squared Error (RMSE)* yang rendah, *R-squared* yang tinggi, dan akurasi relatif di atas 99%. Penggunaan *PySpark* sebagai *platform* pengolahan data distribusi memberikan keunggulan dalam mengelola *volume* data yang besar secara efisien. Grafik Aktual vs. Prediksi Harga Saham memberikan visualisasi yang kuat terhadap seberapa dekat hasil prediksi dengan nilai sebenarnya, dan *PySpark* efisien dalam memfasilitasi proses ini. Keberhasilan *PySpark* dalam mengelola dan memproses big data dalam konteks prediksi saham memberikan kontribusi yang signifikan terhadap ketepatan hasil analisis.

Disarankan untuk mempertimbangkan penambahan faktor-faktor eksternal dan informasi tambahan, seperti berita pasar dan kondisi ekonomi global, dalam model prediktif untuk meningkatkan keakuratan dan responsivitas terhadap perubahan pasar yang dinamis. Sebagai langkah penelitian berikutnya, eksplorasi lebih lanjut terhadap teknik *ensemble* dan integrasi fitur-fitur baru dapat diterapkan dengan memanfaatkan kemampuan distribusi *PySpark*. Oleh karena itu, penelitian ini tidak hanya menekankan pencapaian model prediktif dalam konteks *IDX Composite*, tetapi juga mengungkap potensi besar *PySpark* sebagai alat pengolahan data untuk analisis prediktif saham yang efisien dan dapat diukur secara luas.

Daftar Rujukan

[1] T. Hidayati, D. Wulandari, and W. G. Aedi, "Scientia Sacra : Jurnal Sains , Teknologi dan Masyarakat Implementasi Algoritma C4 . 5 Dalam Memprediksi Harga," vol. 3, no. 4, pp. 1–7, 2023.

[2] D. Tambunan, "Investasi Saham di Masa Pandemi COVID-19," vol. 4, no. 2, pp. 117–123, 2020.

[3] P. C.- Terhadap, P. Saham, D. I. Indonesia, E. Purnaningrum, and V. Ariyanti, "PEMANFAATAN GOOGLE TRENDS UNTUK MENGETAHUI INTERVENSI PANDEMI COVID-19 TERHADAP PASAR SAHAM DI INDONESIA Evita Purnaningrum 1 , Viki Ariyanti 2 2," vol. 25, no. 1411, pp. 93–101, 2020.

[4] N. Hindayani, "ANALISIS REAKSI PASAR SAHAM ATAS PERISTIWA COVID-19," vol. 4, no. 3, pp. 1645–1661, 2020.

[5] B. Hartono, A. Setyo, D. Purnomo, and M. M. Andhini, "PERILAKU INVESTOR SAHAM INDIVIDU DALAM PERPEKTIF TEORI MENTAL ACCOUNTS," pp. 173–183.

[6] D. Rapidminer, "PREDIKSI HARGA SAHAM DENGAN ALGORITMA REGRESI LINIER DENGAN

RAPIDMINER," vol. 10, no. 3, 2022.

[7] E. Patriya, "IMPLEMENTASI SUPPORT VECTOR MACHINE PADA PREDIKSI HARGA SAHAM GABUNGAN (IHSG)," vol. 25, no. 100, pp. 24–38.

[8] O. P. Barus, C. Wijaya, U. P. Harapan, N. N. Backpropagation, and D. Mining, "IMPLEMENTASI METODE NEURAL NETWORK BACKPROPAGATION DALAMPREDIKSI INDEKS HARGA SAHAM GABUNGAN (IHSG)," 2021, doi: 10.47002/seminastika.v3i1.252.

[9] R. A. E. V. T. Sapanji, S. Lestari, and R. Samihardjo, "Prediksi Indeks Bursa Efek Indonesia 2023 Pendekatan ARIMA , Machine Learning dengan R Programming Indonesia Stock Exchange Index Prediction 2023 with the ARIMA Approach , Machine Learning with R Programming," vol. 13, pp. 163–177, 2023.

[10] H. Safitri, "Pengaruh Korea Composite Stock price Index , Hang Seng Index , Staits Times Index dan Dow Jones Industrial Average Terhadap Indeks Harga Saham Gabungan di Bursa Efek Indonesia," vol. 8, pp. 350–359, 2021.

[11] K. M. W. Seso, "PENGARUH KURS DAN HARGA EMAS TERHADAP INDEKS HARGA SAHAM GABUNGAN (IHSG) DI BURSA EFEK INDONESIA (BEI) PERIODE 2020-2021," no. Idx, 2021.

[12] H. S. Telkom and D. A. N. Xl, "PERBANDINGAN MODEL MULTIPLE LINEAR REGRESSION DAN DECISION TREE REGRESSION (STUDI KASUS: PREDIKSI HARGA SAHAM TELKOM, INDOSAT, DAN XL)," vol. 28, no. 1, pp. 78–92.

[13] E. Fitri and D. Riana, "ANALISA PERBANDINGAN MODEL PREDICTION DALAM PREDIKSI HARGA SAHAM MENGGUNAKAN METODE LINEAR REGRESSION , RANDOM FOREST REGRESSION DAN MULTILAYER PERCEPTRON," vol. 6, no. 1, pp. 69–78, 2022.

[14] A. Wulandari and A. A. Rohmawati, "Prediksi Pergerakan Harga Saham PT . Astra Internasional tbk Menggunakan Vector Auto Regressive (VAR) Stasioner dan Logistic Regression," vol. 7, no. 1, pp. 2614–2626, 2020.

[15] P. Harga and S. Menggunakan, "Prediksi Harga Saham Menggunakan BiLSTM dengan Faktor Sentimen Publik," vol. 5, pp. 1–2, 2022.

[16] K. Valiant, Y. Lukito, and R. G. Santosa, "Sistem Prediksi Harga Saham LQ45 Dengan Random Forest Classifier," no. 2, pp. 127–136, 2019, doi: 10.21460/jutei.2019.32.187.