HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 17 Nomor 2, Desember 2026, p430-440
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI: https://doi.org/10.52972/hoaq.vol17no2

PIPELINE HADOOP–SPARK UNTUK ANALISIS DAN PREDIKSI SAHAM LQ45

Dian Gita Meilani

Program Studi Teknologi Informasi, Universitas 'Aisyiyah Yogyakarta
Jl. Siliwangi (Ring Road Barat) No. 63, Nogotirto, Gamping, Sleman – Daerah Istimewa Yogyakarta, Indonesia
Email: dgtmeila.a@gmail.com

ABSTRAK

Data pasar modal tumbuh cepat seiring meningkatnya transaksi saham secara digital. Namun, penelitian prediksi saham LQ45 umumnya berfokus pada model dan belum membahas infrastruktur pengolah datanya. Penelitian ini bertujuan merancang dan mengimplementasikan *pipeline Big Data* berbasis Apache Hadoop dan Apache Spark untuk menganalisis dan memprediksi pergerakan harga saham LQ45. Penelitian menggunakan pendekatan kuantitatif eksperimental. Data harian 45 emiten LQ45 periode Juni 2021–Juni 2026 dikumpulkan melalui API *ingestion*, disimpan di HDFS, lalu diproses dengan Spark. Tahap pemrosesan meliputi *preprocessing*, analisis deskriptif, serta pemodelan regresi dan klasifikasi dengan Spark MLlib. Hasil penelitian menunjukkan empat temuan utama. Pertama, *pipeline* mampu mengolah 53.682 baris data, dan format Parquet menghemat ruang penyimpanan sekitar 87%. Kedua, sektor Teknologi memiliki rentang harga harian tertinggi tetapi *return* terendah. Ketiga, Linear Regression menjadi model regresi terbaik dengan R^2 0,9992. Keempat, akurasi seluruh model klasifikasi arah hanya sekitar 57%. Simpulannya, infrastruktur Hadoop–Spark layak diterapkan untuk analitik pasar modal. Nilai R^2 yang tinggi terutama mencerminkan persistensi harga harian, sedangkan arah pergerakan harga tetap sulit diprediksi.

Kata kunci: *big data*; Apache Hadoop; Apache Spark; saham LQ45; *machine learning*

ABSTRACT

Capital market data grow rapidly as digital stock trading increases. However, previous LQ45 stock prediction studies mostly focus on models and rarely address the underlying data processing infrastructure. This study aims to design and implement a Big Data pipeline based on Apache Hadoop and Apache Spark to analyze and predict LQ45 stock price movements. A quantitative experimental approach was used. Daily data of 45 LQ45 constituents from June 2021 to June 2026 were collected through API ingestion, stored in HDFS, and processed with Spark. The processing stages include preprocessing, descriptive analysis, and regression and classification modeling with Spark MLlib. The results show four main findings. First, the pipeline processed 53,682 records, and the Parquet format reduced storage space by about 87%. Second, the Technology sector had the highest daily price range but the lowest return. Third, Linear Regression was the best regression model with an R^2 of 0.9992. Fourth, all direction classification models achieved only about 57% accuracy. In conclusion, a Hadoop–Spark infrastructure is feasible for capital market analytics. The high R^2 mainly reflects daily price persistence, whereas the direction of price movement remains difficult to predict.

Keywords: big data; Apache Hadoop; Apache Spark; LQ45 stocks; machine learning

1. PENDAHULUAN

Platform perdagangan saham berbasis digital telah memperluas partisipasi investor. Setiap transaksi menghasilkan data harga dan volume yang terus bertambah. Karena itu, sektor keuangan menjadi salah satu pengguna utama analitik *Big Data* [1]. Data pasar modal memiliki ciri *Big Data* karena volumenya besar, pertumbuhannya cepat, dan atributnya saling berkaitan. *Machine learning* telah banyak digunakan untuk memprediksi *return* saham [2]. Model *deep learning* seperti *long short-term memory* (LSTM) juga diterapkan untuk memprediksi pergerakan pasar [3].

Pengolahan data berskala besar memerlukan infrastruktur terdistribusi. Apache Hadoop menyediakan *Hadoop Distributed File System* (HDFS), yaitu sistem berkas yang menyimpan data dalam blok-blok terdistribusi. Apache Spark memproses data secara *in-memory*, yaitu langsung di memori utama, sehingga lebih cepat untuk proses berulang seperti pelatihan model [4]. Pilihan format penyimpanan di Hadoop juga memengaruhi volume data dan waktu pemrosesan [5].

Indeks LQ45 terdiri atas 45 saham dengan likuiditas tinggi dan kapitalisasi pasar besar di Bursa Efek Indonesia (BEI). Indeks ini sering dijadikan objek penelitian prediksi saham [6]. [7] menerapkan Random Forest Classifier dengan indikator teknikal pada saham LQ45. Hansun dan Young menggunakan LSTM pada saham

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

sektor keuangan LQ45 [8]. Budiharto menerapkan LSTM pada saham perbankan selama pandemi [9]. [10] membandingkan beberapa model untuk memprediksi *return* saham LQ45. [11] membandingkan Linear Regression, Random Forest Regression, dan Multilayer Perceptron [11].

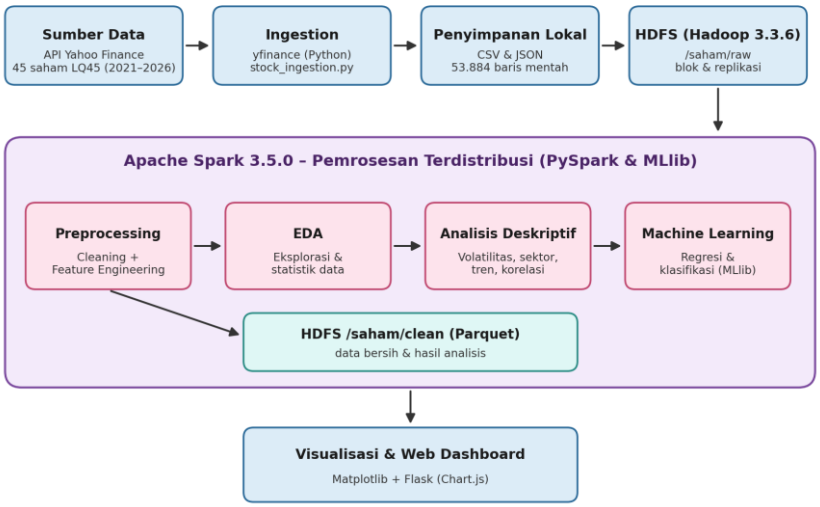
Penelitian terdahulu tersebut menyisakan dua kesenjangan. Pertama, sebagian besar penelitian hanya mencakup sedikit emiten. [8] menggunakan enam saham sektor keuangan, [9] dua saham perbankan, dan [10] enam konstituen LQ45. Penelitian tersebut juga berfokus pada akurasi model tanpa mendokumentasikan infrastruktur pengolah datanya. Kedua, belum banyak penelitian yang membangun *pipeline* terdistribusi utuh untuk seluruh konstituen LQ45. *Pipeline* yang dimaksud adalah rangkaian tahapan pengolahan data dari pengumpulan hingga penyajian hasil. Evaluasi kinerja infrastruktur, seperti waktu eksekusi dan efisiensi penyimpanan, juga jarang dilaporkan.

Berdasarkan kesenjangan tersebut, penelitian ini menjawab tiga rumusan masalah. Pertama, bagaimana mengimplementasikan arsitektur *Big Data* berbasis Hadoop dan Spark untuk mengolah data saham LQ45 dari awal hingga akhir (*end-to-end*). Kedua, bagaimana karakteristik saham LQ45 berdasarkan rentang harga harian, kinerja sektor, tren harga, dan korelasi antarsektor. Ketiga, bagaimana performa model Spark MLlib dalam memprediksi harga penutupan dan arah pergerakan harga harian. Tujuan penelitian ini adalah membangun *pipeline* tersebut, mengevaluasi kinerjanya, menganalisis karakteristik saham LQ45, dan membandingkan model regresi serta klasifikasi. Secara teoretis, penelitian ini menambah kajian penerapan arsitektur *Big Data* pada data pasar modal Indonesia. Secara praktis, hasilnya dapat menjadi rujukan bagi investor dan pengembang sistem analitik saham.

2. METODE PENELITIAN

Desain Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan desain eksperimental terapan. Luarannya berupa *pipeline Big Data* yang diuji secara eksperimental, disertai analisis deskriptif dan eksperimen pemodelan prediktif. Arsitektur sistem terdiri atas empat lapisan, yaitu akuisisi data, penyimpanan (HDFS), pemrosesan (Spark dan MLlib), serta penyajian. Arsitektur tersebut ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur *pipeline Big Data* penelitian

Objek, Populasi, dan Sampel

Objek penelitian adalah saham konstituen indeks LQ45 di BEI. Sampel ditentukan dengan teknik sampling jenuh, sehingga seluruh 45 emiten digunakan. Unit analisisnya adalah data perdagangan harian periode 9 Juni 2021–9 Juni 2026. Karakteristik dataset ditunjukkan pada Tabel 1.

Tabel 1. Karakteristik dataset saham LQ45

Atribut	Keterangan
Jumlah baris	53.884 (mentah); 53.682 (setelah pembersihan)
Jumlah emiten / sektor	45 saham / 11 sektor
Periode	9 Juni 2021 – 9 Juni 2026 (5 tahun)
Atribut data	`date`, `ticker`, `sector`, `open`, `high`, `low`, `close`, `adj_close`, `volume`

Definisi Operasional Variabel

Definisi operasional variabel ditunjukkan pada Tabel 2. Notasi C_t, H_t, dan L_t menyatakan harga penutupan, tertinggi, dan terendah pada hari ke-t. Rentang harga harian (PR_t) digunakan sebagai indikator volatilitas.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 2. Definisi operasional variabel penelitian

Variabel	Definisi Operasional	Satuan
Harga penutupan (C_t)	Harga penutupan harian saham	Rp
Return harian (r_t)	Perubahan C_t terhadap C_{t-1} (Persamaan 1)	%
Rentang harga harian (PR_t)	Selisih harga tertinggi dan terendah terhadap harga terendah (Persamaan 2)	%
Risiko (σ)	Standar deviasi *return* harian (Persamaan 3)	%
MA7, MA30	Rata-rata bergerak harga penutupan 7 dan 30 hari	Rp
Target regresi	Harga penutupan hari berikutnya (C_{t+1})	Rp
Target klasifikasi	NAIK jika $C_{t+1} > C_t$, selain itu TURUN	Kategori

Teknik Pengumpulan Data

Data dikumpulkan melalui API *ingestion*, yaitu pengambilan data terstruktur langsung dari antarmuka pemrograman aplikasi (API). Sumber datanya adalah Yahoo Finance yang diakses dengan pustaka yfinance. Sumber ini lazim digunakan pada penelitian saham Indonesia [8]. Skrip `stock_ingestion.py` mengambil data 45 *ticker* berakhiran `.JK` secara *batch* dan menambahkan atribut sektor. Skrip ini dilengkapi *logging* dan *error handling* agar kegagalan satu *ticker* tidak menghentikan proses. Data mentah disimpan dalam format CSV dan JSON, lalu diunggah ke HDFS.

Infrastruktur dibangun pada lingkungan yang ditunjukkan pada Tabel 3. Hadoop dijalankan dalam mode *pseudo-distributed*, yaitu Hadoop berjalan di satu komputer yang berperan sebagai beberapa komponen *cluster*. Tahapan implementasinya meliputi penyiapan mesin virtual, instalasi Java dan SSH, konfigurasi Hadoop, aktivasi HDFS, integrasi Spark, unggah data, eksekusi *job* Spark, dan penerapan *web dashboard*.

Tabel 3. Spesifikasi lingkungan implementasi

Komponen	Spesifikasi / Versi
Sistem operasi	Ubuntu (Linux) pada VirtualBox
Java / Python	OpenJDK 11 / Python 3
Apache Hadoop	3.3.6 (mode *pseudo-distributed*)
Apache Spark	3.5.0 (mode lokal)
Pustaka utama	yfinance, pandas, PySpark (MLlib), matplotlib
Memori / prosesor	5,3 GB RAM (mesin virtual) / AMD Ryzen 7 6800H

Teknik Analisis Data

Analisis data dilakukan dengan PySpark dalam empat tahap. Tahap pertama adalah *preprocessing*. Baris tanpa harga penutupan, baris dengan volume nol, dan data duplikat dihapus. Selanjutnya dilakukan *feature engineering*, yaitu pembentukan atribut turunan `prev_close`, `daily_return`, `price_range_pct`, `ma7`, `ma30`, `year`, dan `month`. Hasilnya disimpan ke HDFS dalam format Parquet, yaitu format penyimpanan berbasis kolom yang terkompresi.

Tahap kedua adalah analisis deskriptif dengan Persamaan (1)–(4). Persamaan tersebut menghitung *return* harian, rentang harga harian, standar deviasi *return*, dan koefisien korelasi Pearson antarsektor. Notasi n menyatakan jumlah observasi, sedangkan X_t dan Y_t adalah *return* rata-rata dua sektor pada hari ke-t.

$$r_t = \frac{C_t - C_{t-1}}{C_{t-1}} \times 100\% \tag{1}$$

$$PR_t = \frac{H_t - L_t}{L_t} \times 100\% \tag{2}$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(r_t - \overline{r})^2} \tag{3}$$

$$\rho_{XY} = \frac{\sum_{t=1}^{n}(X_t - \overline{X})(Y_t - \overline{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \overline{X})^2 \sum_{t=1}^{n}(Y_t - \overline{Y})^2}} \tag{4}$$

Tahap ketiga adalah pemodelan dengan Spark MLlib. Model regresi (Linear Regression, Decision Tree, Random Forest, dan Gradient Boosted Tree/GBT) memprediksi C_{t+1}. Model klasifikasi (Logistic Regression, Decision Tree, Random Forest, dan GBT) memprediksi arah pergerakan harga. Data dibagi secara acak menjadi 80% data latih dan 20% data uji. Pembagian menggunakan *seed* tetap, yaitu nilai awal pembangkit bilangan acak, agar hasilnya dapat diulang.

Model regresi dievaluasi dengan RMSE, MAE, dan R^2 (Persamaan 5–7). Notasi y_i adalah nilai aktual, $\hat{y}_i$ nilai prediksi, dan $\bar{y}$ rata-rata nilai aktual. Model klasifikasi dievaluasi dengan *accuracy*, *precision*, *recall*, dan F1 (Persamaan 8–10). Notasi TP, TN, FP, dan FN adalah komponen *confusion matrix*. Nilai *precision*, *recall*, dan F1 dihitung secara tertimbang terhadap proporsi kelas.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (8)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (9)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (10)$$

Model terbaik dipilih menggunakan skor akumulasi karena setiap metrik memiliki skala dan arah yang berbeda. Metrik yang semakin besar semakin baik dinormalisasi dengan Persamaan (11). Metrik galat (RMSE dan MAE) dinormalisasi dengan Persamaan (12). Hasil normalisasi kemudian dirata-ratakan dengan Persamaan (13), dengan *k* adalah jumlah metrik.

$$s = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (11)$$

$$s = \frac{x_{\max} - x}{x_{\max} - x_{\min}} \quad (12)$$

$$\text{Skor} = \frac{1}{k}\sum_{j=1}^{k} s_j \quad (13)$$

Tahap keempat adalah evaluasi performa sistem. Evaluasi ini mengukur waktu eksekusi setiap tahap dengan perintah `time`, membandingkan jumlah *core* Spark, serta meninjau statistik penyimpanan HDFS.

Uji Keabsahan Data

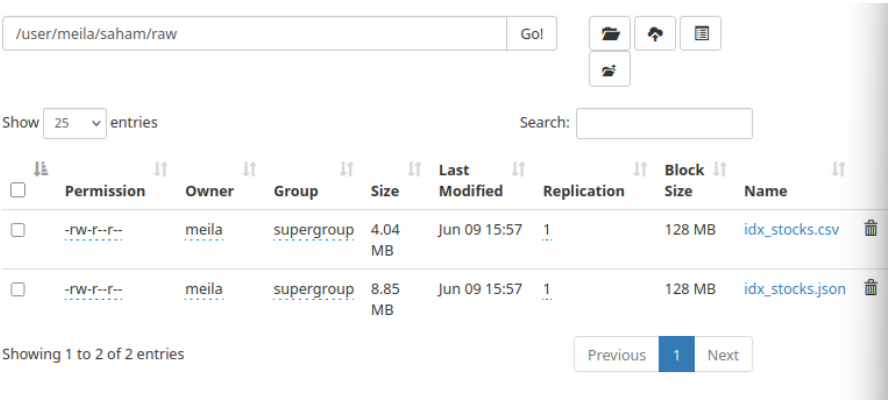
Keabsahan data dijaga melalui empat langkah. Pertama, kelengkapan dan konsistensi data diperiksa melalui nilai kosong, duplikasi, volume nol, dan jumlah baris per emiten. Kedua, *return* harian di atas 25% ditandai sebagai *outlier* untuk diperiksa kewajarannya. Ketiga, integritas penyimpanan diverifikasi dengan perintah `hdfs dfsadmin -report` untuk memastikan tidak ada blok yang rusak atau hilang. Keempat, reliabilitas eksperimen dijaga dengan *seed* tetap, pembagian data yang sama untuk semua model, dan penggunaan beberapa metrik evaluasi sekaligus.

3. HASIL DAN PEMBAHASAN

Hasil

Gambaran Umum Data

Proses *ingestion* menghasilkan 53.884 baris data mentah. Data tersebut diunggah ke direktori HDFS `/user/meila/saham/raw/`. Gambar 2 menunjukkan berkas `idx_stocks.csv` (4,0 MB) dan `idx_stocks.json` (8,9 MB) yang tersimpan dengan ukuran blok 128 MB. Setelah pembersihan, data berkurang menjadi 53.682 baris. Hasil *exploratory data analysis* (EDA) ditunjukkan pada Tabel 4.



Gambar 2. Tampilan HDFS Web UI yang menampilkan berkas dataset mentah

Tabel 4. Hasil *exploratory data analysis* dataset saham LQ45

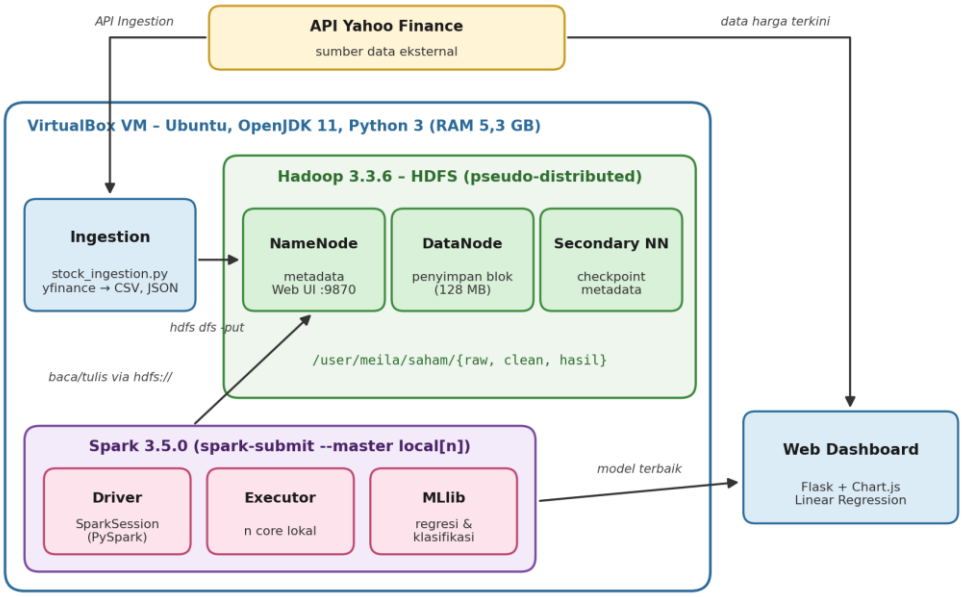
Aspek	Hasil
Nilai kosong (*null*)	Hanya pada `prev_close` dan `daily_return` (45 baris)
Harga penutupan (Rp)	Rata-rata Rp4.669; minimum Rp50; maksimum Rp47.250
Return harian (%)	Rentang −24,8% s.d. +34,9%; rata-rata mendekati 0,0%
Outlier ekstrem (>25,0%)	2 baris dari 53.682 baris
Jumlah baris per saham	Umumnya 1.199 baris; GOTO 986 baris dan BUKA 1.154 baris

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Nilai kosong pada `prev_close` dan `daily_return` berasal dari baris pertama setiap saham. Jumlah baris GOTO dan BUKA lebih sedikit karena kedua emiten melakukan *Initial Public Offering* (IPO) setelah awal periode pengamatan.

Implementasi Infrastruktur

Topologi infrastruktur ditunjukkan pada Gambar 3. Seluruh layanan berjalan dalam satu mesin virtual. HDFS menjalankan tiga *daemon*. NameNode mengelola *namespace*, yaitu struktur direktori dan lokasi blok data. DataNode menyimpan blok data secara fisik. SecondaryNameNode menyimpan salinan metadata secara berkala (*checkpoint*). Spark dijalankan dengan perintah `spark-submit` dalam mode lokal `local[n]`. Pada mode ini, *driver* dan *executor* Spark berjalan dalam satu proses Java dengan *n thread*. Spark tidak dijalankan di atas YARN, yaitu pengelola sumber daya *cluster* Hadoop.



Gambar 3. Topologi implementasi infrastruktur Hadoop dan Spark

Konfigurasi dan perintah utama ditunjukkan pada Tabel 5. Variabel `JAVA_HOME`, `HADOOP_HOME`, dan `SPARK_HOME` ditambahkan ke berkas `~/.bashrc`. Parameter `fs.defaultFS` menentukan alamat NameNode. Parameter `dfs.replication` bernilai 1 karena hanya terdapat satu DataNode. HDFS diinisialisasi dengan `hdfs namenode -format` dan dijalankan dengan `start-dfs.sh`. Statusnya diverifikasi dengan perintah `jps` dan HDFS Web UI pada `localhost:9870`.

Tabel 5. Konfigurasi dan perintah utama infrastruktur

Berkas / Perintah	Parameter / Argumen	Nilai / Fungsi
`~/.bashrc`	`JAVA_HOME`, `HADOOP_HOME`, `SPARK_HOME`, `PATH`	Variabel lingkungan
`hadoop-env.sh`	`JAVA_HOME`	Lokasi Java untuk *daemon* Hadoop
`core-site.xml`	`fs.defaultFS`	`hdfs://localhost:9000` (alamat NameNode)
`hdfs-site.xml`	`dfs.replication`; `dfs.blocksize`	1; 128 MB (bawaan)
`hdfs dfs`	`-mkdir -p`; `-put`	Membuat direktori; mengunggah data mentah
`spark-submit`	`--master local[n]`	Menjalankan *job* Spark dengan *n core*
`hdfs dfsadmin`	`-report`	Memeriksa kapasitas dan kondisi blok

Setiap tahap *pipeline* diimplementasikan sebagai skrip PySpark terpisah. Setiap skrip membaca masukan dari HDFS melalui alamat `hdfs://` dan menuliskan keluarannya ke HDFS atau sistem berkas lokal. Alur tersebut dirangkum pada Tabel 6.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 6. Alur masukan dan keluaran setiap skrip *pipeline*

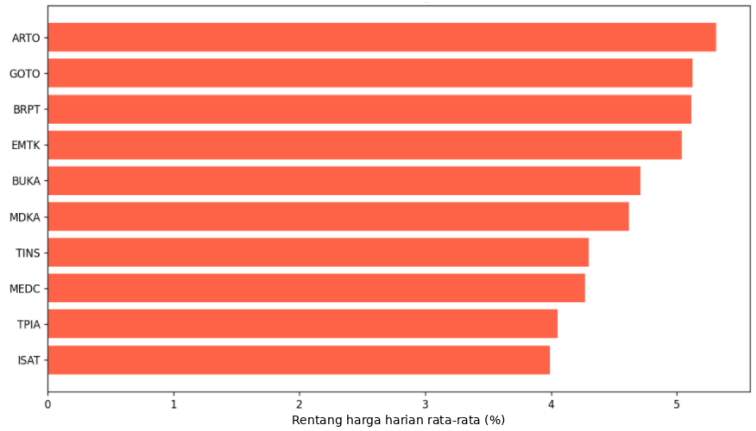
Skrip	Masukan	Keluaran
`stock_ingestion.py`	API Yahoo Finance	CSV dan JSON → `/saham/raw`
`01_preprocessing.py`	`/saham/raw`	`/saham/clean` (Parquet)
`02_eda.py`	`/saham/clean`	Statistik dan kualitas data
`03_analisis.py`	`/saham/clean`	`/saham/hasil` (agregasi analisis)
`04_machine_learning.py`	`/saham/clean`	Metrik evaluasi delapan model
`05_visualisasi.py`	Hasil analisis dan pemodelan	Grafik PNG
`06_prediksi.py`	`/saham/clean`	CSV hasil prediksi
`app.py`	API Yahoo Finance (data terkini)	*Web dashboard*

Hasil Analisis Deskriptif

Saham dengan rentang harga harian rata-rata tertinggi ditunjukkan pada Tabel 7 dan Gambar 4. ARTO memiliki nilai tertinggi (5,315%), diikuti GOTO, BRPT, dan EMTK. ISAT dan TPIA berada di posisi terbawah dalam kelompok sepuluh saham tersebut.

Tabel 7. Saham dengan rentang harga harian rata-rata tertinggi (%)

Ticker	Sektor	Rentang Harga Harian (%)	Standar Deviasi *Return* (%)
ARTO	Perbankan	5,315	3,908
GOTO	Teknologi	5,123	3,915
BRPT	Kimia	5,113	4,028
EMTK	Teknologi	5,038	3,615
BUKA	Teknologi	4,708	3,190
MDKA	Pertambangan	4,618	3,480



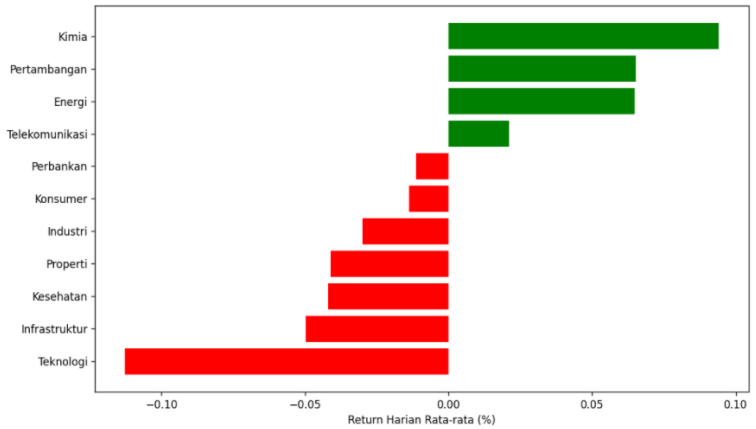
Gambar 4. Sepuluh saham LQ45 dengan rentang harga harian rata-rata tertinggi (%)

Kinerja enam sektor terpilih ditunjukkan pada Tabel 8. Rata-rata *return* harian seluruh 11 sektor ditunjukkan pada Gambar 5. Sektor Kimia, Pertambangan, dan Energi mencatat *return* positif. Sektor Teknologi mencatat *return* terendah (−0,1127%) dengan rentang harga harian tertinggi (4,949%).

Tabel 8. Rata-rata *return* harian dan rentang harga harian per sektor (%)

Sektor	*Return* Harian (%)	Rentang Harga Harian (%)	Jumlah Saham
Kimia	0,0941	4,582	2
Pertambangan	0,0653	4,038	4
Energi	0,0648	3,089	5
Perbankan	−0,0113	2,962	6
Konsumer	−0,0138	2,990	9
Teknologi	−0,1127	4,949	3

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



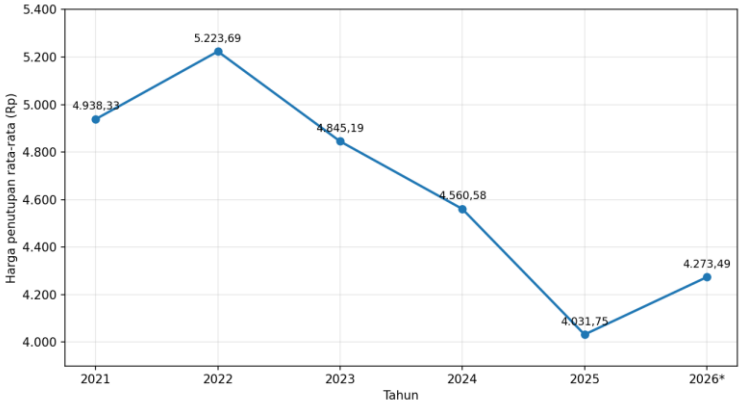
Gambar 5. Rata-rata *return* harian per sektor (%)

Tabel 9 dan Gambar 6 menunjukkan bahwa rata-rata harga penutupan naik hingga 2022. Harga kemudian turun pada 2023–2025 dan naik kembali pada 2026. Rata-rata *return* harian berfluktuasi antartahun.

Tabel 9. Rata-rata harga penutupan (Rp) dan *return* harian (%) saham LQ45 per tahun

Tahun	Harga Penutupan Rata-rata (Rp)	*Return* Harian Rata-rata (%)
2021	4.938,33	0,0225
2022	5.223,69	0,0172
2023	4.845,19	0,0063
2024	4.560,58	−0,0113
2025	4.031,75	0,0704
2026*	4.273,49	−0,2773

*Data 2026 hanya mencakup periode Januari–Juni 2026.



Gambar 6. Rata-rata harga penutupan saham LQ45 per tahun (Rp)

Lima pasangan sektor dengan korelasi *return* tertinggi ditunjukkan pada Tabel 10. Korelasi tertinggi terdapat pada pasangan Perbankan–Properti (0,523). Seluruh nilai berada pada rentang 0,444–0,523.

Tabel 10. Lima pasangan sektor dengan koefisien korelasi *return* tertinggi

Sektor 1	Sektor 2	Koefisien Korelasi
Perbankan	Properti	0,523
Industri	Konsumer	0,491
Industri	Properti	0,489
Industri	Perbankan	0,461
Konsumer	Perbankan	0,444

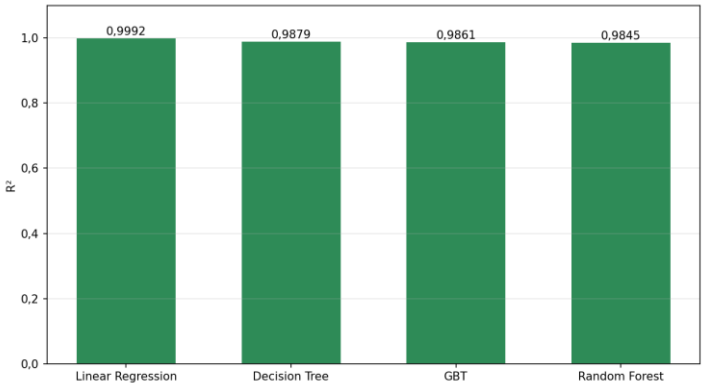
Hasil Pemodelan

Hasil evaluasi model regresi ditunjukkan pada Tabel 11 dan Gambar 7. Linear Regression memperoleh skor akumulasi 1,0000. Model ini memiliki RMSE dan MAE terendah serta R^2 tertinggi. Model berbasis pohon menghasilkan RMSE Rp676,413–Rp765,954.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

Tabel 11. Hasil evaluasi model regresi (RMSE dan MAE dalam Rp)

Model	RMSE (Rp)	MAE (Rp)	R^2	Skor Akumulasi
Linear Regression	177,316	76,099	0,9992	1,0000
GBT	724,374	238,705	0,9861	0,1565
Decision Tree	676,413	305,103	0,9879	0,1278
Random Forest	765,954	298,755	0,9845	0,0092



Gambar 7. Perbandingan nilai R^2 model regresi

Contoh hasil prediksi Linear Regression pada data uji ditunjukkan pada Tabel 12. Galat sampel berkisar 0,25%–2,26%, dan rata-rata galat prediksi tercatat sebesar 1,83%.

Tabel 12. Contoh hasil prediksi harga penutupan dengan Linear Regression

Ticker	Harga Aktual (Rp)	Harga Prediksi (Rp)	Selisih (Rp)	Galat (%)
ANTM	2.480,00	2.423,92	−56,08	2,26
BBCA	6.470,00	6.604,05	134,05	2,07
ICBP	8.150,00	8.173,43	23,43	0,29
INCO	4.850,00	4.837,80	−12,20	0,25
MYOR	2.540,00	2.554,42	14,42	0,57
ASII	5.150,00	5.180,23	30,23	0,59

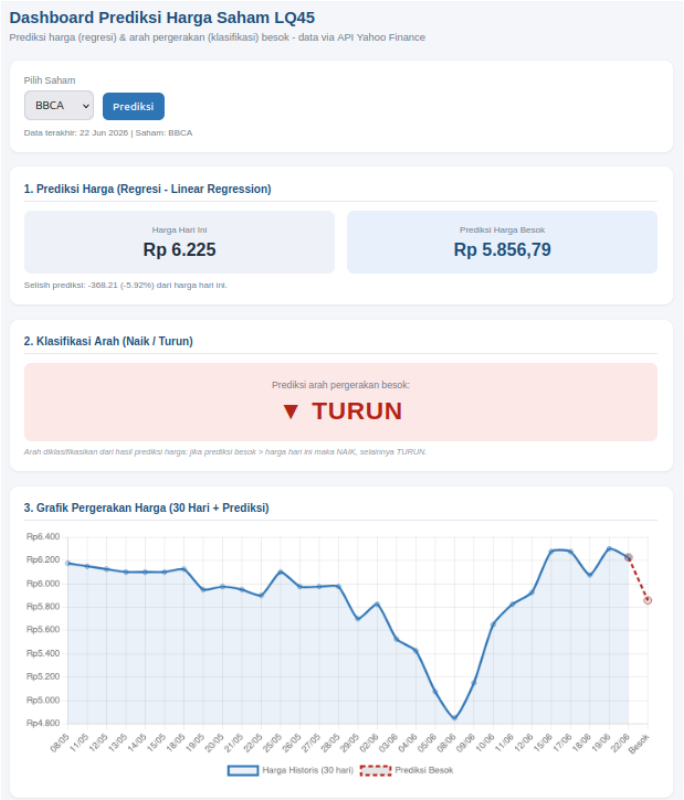
Hasil model klasifikasi ditunjukkan pada Tabel 13. Decision Tree memperoleh skor akumulasi tertinggi (0,8729). Namun, akurasi keempat model berada pada rentang sempit 57,07%–57,39%. Pada 10.552 data uji, model terbaik menghasilkan 6.022 prediksi benar dan 4.530 prediksi salah.

Tabel 13. Hasil evaluasi model klasifikasi arah pergerakan harga

Model	*Accuracy*	*Precision*	*Recall*	F1	Skor Akumulasi
Decision Tree	0,5739	0,5451	0,5739	0,4490	0,8729
Logistic Regression	0,5723	0,5369	0,5723	0,4517	0,4878
Random Forest	0,5736	0,5311	0,5736	0,4198	0,4531
GBT	0,5707	0,5383	0,5707	0,4792	0,3786

Model terbaik diterapkan pada *web dashboard* berbasis Flask (Gambar 8). Server mengambil data terkini dari API Yahoo Finance dan melatih ulang Linear Regression. Model kemudian memprediksi harga penutupan hari berikutnya. Arah ditetapkan NAIK jika harga prediksi lebih tinggi dari harga terakhir, dan TURUN jika sebaliknya.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi



Gambar 8. Tampilan *web dashboard* prediksi harga saham

Hasil Evaluasi Performa Sistem

Waktu eksekusi setiap tahap ditunjukkan pada Tabel 14. Tahap *machine learning* paling lama, yaitu 323,7 detik, karena melatih delapan model. Tahap prediksi regresi paling cepat, yaitu 38,8 detik. Pada uji jumlah *core*, tahap *preprocessing* memerlukan 24,8 detik dengan `local[1]` dan 23,2 detik dengan `local[2]`.

Tabel 14. Waktu eksekusi setiap tahap *pipeline* (detik)

Tahap	Waktu Eksekusi (detik)
Preprocessing	103,7
Exploratory data analysis	77,6
Analisis deskriptif	89,7
Machine learning (8 model)	323,7
Prediksi regresi	38,8
Prediksi klasifikasi	68,2

Statistik penyimpanan HDFS ditunjukkan pada Tabel 15. Laporan *cluster* menunjukkan kapasitas terkonfigurasi 39,1 GB dan pemakaian DFS 15,1 MB (0,1%). Tidak ada blok yang rusak maupun hilang.

Tabel 15. Ukuran data dan ukuran tereplikasi setiap direktori HDFS

Direktori HDFS	Ukuran Data	Ukuran Tereplikasi
`/saham/raw` (data mentah)	12,9 MB	12,9 MB
`/saham/clean` (Parquet)	1,7 MB	5,0 MB
`/saham/hasil` (hasil analisis)	3,4 KB	10,1 KB

Pembahasan

Kinerja Infrastruktur Big Data

Hasil implementasi menjawab rumusan masalah pertama. Arsitektur Hadoop–Spark mampu menjalankan seluruh tahap *pipeline* dengan waktu yang wajar dan integritas blok yang terjaga. Penambahan *core* hanya sedikit mempercepat proses. Hal ini terjadi karena ukuran data masih kecil. Pada data kecil, beban tambahan (*overhead*) untuk menyiapkan dan menjadwalkan tugas Spark lebih besar daripada manfaat paralelisasi. Penghematan ruang dari format Parquet sejalan dengan temuan Belov dkk. bahwa pilihan format penyimpanan di Hadoop memengaruhi volume data dan waktu pemrosesan [5].

Ukuran tereplikasi direktori hasil tulisan Spark sekitar tiga kali ukuran datanya. Kondisi ini berbeda dengan data mentah yang diunggah melalui klien Hadoop. Hal ini mengindikasikan bahwa Spark menggunakan faktor

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

replikasi bawaan (3), kemungkinan karena berkas `hdfs-site.xml` tidak terbaca oleh Spark. Pada satu DataNode, replikasi tersebut hanya bersifat nominal. Oleh karena itu, konfigurasi Hadoop dan Spark perlu diselaraskan, misalnya melalui variabel `HADOOP_CONF_DIR`.

Karakteristik Saham LQ45

Untuk rumusan masalah kedua, rentang harga harian yang tinggi pada saham teknologi dan perbankan digital mencerminkan karakter saham pertumbuhan (*growth stock*). Harga saham jenis ini sensitif terhadap perubahan ekspektasi pasar. Pada sektor Teknologi, rentang harga yang tinggi tidak diimbangi *return* yang memadai. Sebaliknya, sektor berbasis komoditas menunjukkan kinerja yang lebih baik selama periode pengamatan. Rentang harga yang tinggi juga menyulitkan prediksi. [12] menemukan bahwa model prediksi bekerja lebih baik pada indeks yang lebih stabil. Korelasi tertinggi pada pasangan Perbankan–Properti dapat dikaitkan dengan ketergantungan sektor properti pada pembiayaan bank. Korelasi antarsektor yang lemah hingga sedang menunjukkan bahwa diversifikasi lintas sektor dapat membantu menurunkan risiko portofolio.

Kinerja Model Prediksi

Untuk rumusan masalah ketiga, perbedaan besar antara kinerja regresi dan klasifikasi perlu ditafsirkan dengan hati-hati. Harga penutupan harian memiliki autokorelasi yang kuat, yaitu harga hari ini sangat dipengaruhi harga hari sebelumnya. Akibatnya, harga esok hampir selalu dekat dengan harga hari ini. Dengan fitur `prev_close` dan rata-rata bergerak, model linear secara alami unggul. Nilai R^2 yang tinggi terutama mencerminkan persistensi harga, bukan kemampuan menebak arah perubahan. Hal ini juga menjelaskan mengapa galat dalam penelitian ini jauh lebih kecil daripada galat LSTM pada saham sektor keuangan LQ45 [8]. Perbedaan horizon dan skema evaluasi membuat kedua hasil tersebut tidak dapat dibandingkan secara langsung.

Akurasi klasifikasi yang hanya sedikit di atas 50% sejalan dengan *Efficient Market Hypothesis*. Hipotesis ini menyatakan bahwa harga telah mencerminkan informasi yang tersedia, sehingga perubahan jangka pendek sulit diprediksi [13]. Temuan ini memperkuat hasil Indra dkk. yang melaporkan akurasi arah 49%–54% pada saham LQ45 [10]. Dalam penelitian yang sama, model *random walk* sebagai pembanding mencapai akurasi arah sekitar 56%. Valiant dkk. juga hanya memperoleh akurasi di atas 60% pada tiga kode saham untuk horizon satu hari [7]. Penelitian lain melaporkan akurasi yang jauh lebih tinggi, misalnya pada LSTM dengan indikator teknikal di pasar Vietnam [14] dan pada saham perbankan Indonesia [9]. Namun, penelitian tersebut menggunakan fitur, horizon, dan definisi akurasi yang berbeda. Nilai F1 yang lebih rendah daripada akurasi mengindikasikan bahwa model cenderung lebih sering memprediksi kelas mayoritas. Tantangan serupa juga dibahas dalam evaluasi model *ensemble* untuk prediksi tren indeks saham [15].

Implikasi dan Keterbatasan

Secara praktis, infrastruktur Hadoop–Spark dapat dibangun dengan sumber daya terbatas dan dapat diperluas ke data yang lebih besar. Penelitian ini memiliki empat keterbatasan. Pertama, data dibagi secara acak sehingga sebagian data uji berada di antara data latih secara kronologis. Untuk data runtun waktu yang tidak stasioner, metode evaluasi yang mempertahankan urutan waktu memberikan estimasi yang lebih akurat [16]. Kedua, model belum dibandingkan dengan *baseline* naif, misalnya harga hari ini sebagai prediksi harga esok. Ketiga, infrastruktur masih berupa satu node dengan ukuran data yang kecil. Akibatnya, manfaat skalabilitas dan toleransi kegagalan belum teruji sepenuhnya. Keempat, data 2026 hanya mencakup setengah tahun dan kualitas data bergantung pada Yahoo Finance.

4. SIMPULAN

Penelitian ini berhasil membangun *pipeline Big Data* berbasis Apache Hadoop dan Apache Spark untuk seluruh konstituen indeks LQ45. *Pipeline* tersebut mencakup pengumpulan data, penyimpanan di HDFS, pemrosesan dan pemodelan dengan Spark, hingga penyajian melalui *web dashboard*. Infrastruktur berjalan stabil dan hemat ruang penyimpanan berkat format Parquet. Analisis deskriptif menunjukkan bahwa saham teknologi dan perbankan digital memiliki risiko tertinggi tanpa imbal hasil yang memadai. Sektor berbasis komoditas mencatat kinerja yang lebih baik, dan pergerakan antarsektor cukup beragam.

Prediksi harga harian menghasilkan galat kecil karena kuatnya persistensi harga. Sebaliknya, prediksi arah pergerakan hanya sedikit lebih baik daripada tebakan acak. Implikasinya, evaluasi model prediksi saham sebaiknya tidak hanya bertumpu pada R^2. Pasar LQ45 juga menunjukkan perilaku yang mendekati efisiensi bentuk lemah. Keterbatasan penelitian terletak pada pembagian data acak, ketiadaan *baseline* naif, dan lingkungan satu node. Penelitian selanjutnya disarankan menggunakan pembagian data berbasis waktu dan *baseline* naif. Penelitian lanjutan juga dapat menambahkan indikator teknikal, menguji model lain seperti XGBoost atau LSTM, dan mengembangkan *pipeline streaming* dengan Kafka. Infrastruktur juga dapat diperluas ke *cluster* banyak node dengan YARN dan seluruh emiten BEI.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi

DAFTAR PUSTAKA

[1] M. M. Hasan, J. Popp, and J. Oláh, "Current Landscape and Influence of Big Data on Finance," *Journal of Big Data*, vol. 7, no. 1, pp. 1–17, Mar. 2020, doi: 10.1186/s40537-020-00291-z.

[2] S. Gu, B. Kelly, and D. Xiu, "Empirical Asset Pricing Via Machine Learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, May 2020, doi: 10.1093/rfs/hhaa009.

[3] T. Fischer and C. Krauss, "Deep Learning With Long Short-Term Memory Networks for Financial Market Predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.

[4] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, H. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica, "Apache Spark: A Unified Engine for Big Data Processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016, doi: 10.1145/2934664.

[5] V. Belov, A. N. Kosenkov, and E. Nikulchev, "Experimental Characteristics Study of Data Storage Formats for Data Marts Development Within Data Lakes," *Applied Sciences*, vol. 11, no. 18, pp. 1–9, Sep. 2021, doi: 10.3390/app11188651.

[6] G. Achyar and O. Rohaeni, "Penggunaan Hybrid K-Means dan General Regression Neural Network untuk Prediksi Harga Saham Indeks LQ45," *Jurnal Riset Matematika*, vol. 2, no. 2, pp. 111–120, Dec. 2022, doi: 10.29313/jrm.v2i2.1193.

[7] K. Valiant, Y. Lukito, and R. G. Santosa, "Sistem Prediksi Harga Saham LQ45 dengan Random Forest Classifier," *Jurnal Terapan Teknologi Informasi*, vol. 3, no. 2, pp. 127–136, Oct. 2019, doi: 10.21460/jutei.2019.32.187.

[8] S. Hansun and J. C. Young, "Predicting LQ45 Financial Sector Indices Using RNN-LSTM," *Journal of Big Data*, vol. 8, no. 1, pp. 1–13, Aug. 2021, doi: 10.1186/s40537-021-00495-x.

[9] W. Budiharto, "Data Science Approach to Stock Prices Forecasting in Indonesia During Covid-19 Using Long Short-Term Memory (LSTM)," *Journal of Big Data*, vol. 8, no. 1, pp. 1–9, Mar. 2021, doi: 10.1186/s40537-021-00430-0.

[10] Indra, S. Supian, Sukono, Riaman, M. P. A. Saputra, A. S. Azahra, and D. I. Pirdaus, "Stock Return Prediction on the LQ45 Market Index in the Indonesia Stock Exchange Using a Machine Learning Algorithm Based on Technical Indicators," *Journal of Risk and Financial Management*, vol. 18, no. 12, pp. 1–33, Dec. 2025, doi: 10.3390/jrfm18120714.

[11] E. Fitri and D. Riana, "Analisa Perbandingan Model Prediction dalam Prediksi Harga Saham Menggunakan Metode Linear Regression, Random Forest Regression dan Multilayer Perceptron," *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, vol. 6, no. 1, pp. 69–78, Apr. 2022, doi: 10.46880/jmika.Vol6No1.pp69-78.

[12] N. Yeldho, D. Thomas, V. G. Kurian, C. Arathy, and A. V. N. Biju, "Are Machine Learning Models Effective in Predicting Emerging Markets? Investigating the Accuracy of Predictions in Emerging Stock Market Indices," *Quality & Quantity*, vol. 59, no. 1, pp. 839–904, Feb. 2025, doi: 10.1007/s11135-024-01964-0.

[13] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, May 1970, doi: 10.2307/2325486.

[14] T. Phuoc, P. T. K. Anh, P. H. Tam, and C. V. Nguyen, "Applying Machine Learning Algorithms to Predict the Stock Price Trend in the Stock Market – The Case of Vietnam," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–18, Mar. 2024, doi: 10.1057/s41599-024-02807-x.

[15] X. Wei, Y. Tian, N. Li, and H. Peng, "Evaluating Ensemble Learning Techniques for Stock Index Trend Prediction: A case of China," *Portuguese Economic Journal*, vol. 23, no. 3, pp. 505–530, Sep. 2024, doi: 10.1007/s10258-023-00246-1.

[16] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods," *Machine Learning*, vol. 109, no. 11, pp. 1997–2028, Nov. 2020, doi: 10.1007/s10994-020-05910-7.

Copyright (c) HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi