

Analisis Komparatif GridSearchCV pada SVM, Logistic Regression, dan Random Forest untuk Klasifikasi Saham TINS

¹Nilam Anggraini, ²Chandra Kirana, ³Delpiah Wahyuningsih

^{1*,2,3}Teknik Informatika, Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indonesia

*Korespondensi: nilamanggraini.work@gmail.com

Submit : 18 Jun 2026 | Diterima : 07 Jul 2026 | Terbit : 01 Sept 2026

ABSTRACT

This study aims to evaluate and compare the performance of Support Vector Machine (SVM), Logistic Regression, and Random Forest in classifying the stock price movement direction of PT Timah Tbk (TINS) in 2025, with hyperparameter optimization based on GridSearchCV. A quantitative experimental approach with a comparative design was applied to 227 daily stock records from January to December 2025, which resulted in 218 modeling-ready observations after feature engineering. The research stages included data preparation, exploratory data analysis, feature engineering based on return, moving average, volatility, momentum, and volume, standardization, baseline model training, hyperparameter optimization using GridSearchCV, and model evaluation using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and cross-validation accuracy. The results indicate that Logistic Regression achieved the best overall performance, with an accuracy of 0.5714, precision of 0.5714, F1-score of 0.4706, and the highest ROC-AUC of 0.6136. Meanwhile, SVM achieved the highest recall at 0.6500, while Random Forest showed relatively stable performance with a ROC-AUC of 0.5227. The findings also show that volatility-based and price-range-based features contributed most significantly to the classification task. The novelty of this study lies in the comparative analysis of three machine learning algorithms with GridSearchCV optimization on TINS stock data using technical features designed to capture nonlinear price patterns and market volatility more representatively.

Keywords: Hyperparameter, Machine Learning, Feature Engineering, Classification, GridSearchCV

ABSTRAK

Penelitian ini bertujuan untuk mengevaluasi serta membandingkan kinerja algoritma *Support Vector Machine* (SVM), *Logistic Regression*, dan *Random Forest* dalam klasifikasi arah pergerakan saham PT Timah Tbk (TINS) tahun 2025 dengan optimasi *hyperparameter* berbasis *GridSearchCV*. Penelitian menggunakan pendekatan kuantitatif eksperimental dengan rancangan komparatif terhadap 227 data harian saham TINS periode Januari–Desember 2025 yang setelah proses *feature engineering* menghasilkan 218 data siap pemodelan. Tahapan penelitian meliputi data *preparation*, *exploratory data analysis*, *feature engineering* berbasis *return*, *moving average*, volatilitas, momentum, dan volume, normalisasi menggunakan *standardization*, pelatihan model *baseline*, optimasi *hyperparameter* menggunakan *GridSearchCV*, serta menilai model menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, ROC-AUC, *confusion matrix*, dan *cross-validation accuracy*. Hasil penelitian mengindikasikan bahwa *Logistic Regression* merupakan model terbaik dengan nilai *accuracy* sebesar 0,5714, *precision* 0,5714, *F1-score* 0,4706, dan ROC-AUC tertinggi sebesar 0,6136. Sementara itu, SVM menghasilkan recall tertinggi sebesar 0,6500, sedangkan *Random Forest* memperoleh performa yang relatif stabil dengan ROC-AUC sebesar 0,5227. Temuan penelitian menunjukkan bahwa fitur berbasis volatilitas dan range harga memiliki pengaruh paling signifikan terhadap klasifikasi arah pergerakan saham TINS. Novelty penelitian ini terletak pada analisis komparatif tiga algoritma *machine learning* dengan Optimasi *GridSearchCV* pada data saham TINS menggunakan fitur teknikal yang dirancang untuk menangkap pola harga nonlinier dan volatilitas pasar secara lebih representatif.

Kata Kunci: Hyperparameter, Machine Learning, Feature Engineering, Klasifikasi, GridSearchCV

PENDAHULUAN

Pasar modal memiliki peranan penting dalam mendukung pertumbuhan ekonomi melalui aktivitas investasi dan penghimpunan dana (Putu & Nilakusmawati, 2024). Saham sebagai instrumen utama di pasar modal memiliki potensi keuntungan yang lebih tinggi, tetapi juga memiliki risiko besar akibat pergerakan harga yang fluktuatif dan sulit diprediksi (Kurniawan et al., 2025). Perubahan harga saham dipengaruhi oleh berbagai faktor seperti kondisi ekonomi, sentimen investor, kebijakan politik, kondisi sosial, dan dinamika global sehingga pasar bersifat dinamis, nonlinier, dan penuh ketidakpastian (Kurniawan et al., 2025), (Labiad et al., 2023). Kondisi tersebut menjadikan prediksi pergerakan saham sebagai permasalahan yang kompleks dan membutuhkan pendekatan analitis yang adaptif terhadap perubahan pola data (S & Nirmala, 2020), (Sidek et al., 2023). Dalam konteks Indonesia, PT Timah Tbk (TINS) merupakan emiten sektor pertambangan strategis yang tercatat di Bursa Efek Indonesia dan memiliki sensitivitas tinggi terhadap perubahan kondisi pasar maupun industri pertambangan global (Asmara & Rudianto, 2024), (Jaladri & Mulyani, 2020). Pergerakan saham perusahaan ini menjadi perhatian investor karena berkaitan dengan pengambilan keputusan investasi dan manajemen risiko (Fitria & Widayanti, 2024).

Perkembangan teknologi komputasi dan data mining mendorong pemanfaatan *machine learning* dalam analisis pasar saham. Pendekatan *machine learning* dinilai mampu mengidentifikasi pola tersembunyi pada data historis dan menghasilkan prediksi yang lebih adaptif dibandingkan metode konvensional (Kurniawan et al., 2025), (Saud & Shakya, 2022), (Islam et al., 2023). Beberapa algoritma klasifikasi yang banyak digunakan dalam analisis data finansial antara lain *Support Vector Machine* (SVM), *Logistic Regression*, dan *Random Forest* (Prasetyo et al., 2024), (Saepudin et al., 2024), (Cahyani & Prasetyaningrum, 2026). *Logistic Regression* efektif digunakan pada klasifikasi biner karena mampu memodelkan probabilitas hubungan antara variabel *input* dan *output* secara interpretatif serta memiliki proses pelatihan yang relatif cepat (Prasetyo et al., 2024), (Mazya et al., 2024). Sementara itu, SVM mampu membangun *hyperplane* optimal sehingga efektif menangani data nonlinier dan berdimensi tinggi (Saepudin et al., 2024), (Latif et al., 2025). *Random Forest* sebagai metode *ensemble learning* juga dinilai mampu meningkatkan akurasi klasifikasi serta mengurangi bias melalui teknik *bootstrap aggregating* dan *random feature selection* (Sandag et al., 2020), (Raghavendra & J, 2020), (Elsayed et al., 2025).

Selain pemilihan algoritma, performa model *machine learning* dipengaruhi oleh konfigurasi *hyperparameter*. Pengaturan *hyperparameter* yang tidak optimal dapat menyebabkan *overfitting* maupun *underfitting* sehingga menurunkan kualitas prediksi (Misriati & Aryanti, 2024), (Hakim et al., 2024). Salah satu metode optimasi *hyperparameter* yang banyak digunakan adalah *GridSearchCV* karena mampu mengevaluasi kombinasi parameter secara sistematis melalui *cross validation* (Putu & Nilakusmawati, 2024), (Misriati & Aryanti, 2024). Penelitian sebelumnya menunjukkan bahwa *GridSearchCV* dapat meningkatkan performa model SVM dan *Random Forest* secara signifikan (Latif et al., 2025), (Misriati & Aryanti, 2024), (Hakim et al., 2024). Namun, sebagian besar penelitian masih berfokus pada analisis sentimen, klasifikasi teks, atau objek saham yang berbeda (Putu & Nilakusmawati, 2024), (Sidek et al., 2023), (Prasetyo et al., 2024). Beberapa penelitian juga menunjukkan bahwa performa algoritma dipengaruhi oleh karakteristik data yang digunakan. *Logistic Regression* cenderung optimal pada pola linear, SVM unggul dalam menangani *boundary* yang kompleks, sedangkan *Random Forest* lebih stabil pada data dengan variasi fitur yang tinggi (Prasetyo et al., 2024), (Saepudin et al., 2024), (Cahyani & Prasetyaningrum, 2026). Meskipun demikian, penelitian yang membandingkan ketiga algoritma tersebut secara bersamaan menggunakan optimasi *hyperparameter* berbasis *GridSearchCV* pada klasifikasi pergerakan saham PT Timah Tbk masih terbatas. Selain itu, sebagian penelitian terdahulu lebih berfokus pada prediksi harga saham dan analisis sentimen dibandingkan klasifikasi arah pergerakan saham secara komprehensif (Islam et al., 2023), (Mazya et al., 2024).

Berdasarkan permasalahan tersebut, penelitian ini bertujuan melakukan analisis komparatif terhadap algoritma SVM, *Logistic Regression*, dan *Random Forest* dengan optimasi *hyperparameter* berbasis *GridSearchCV* untuk memperoleh model terbaik dalam klasifikasi pergerakan saham PT Timah Tbk. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan model klasifikasi saham berbasis *machine learning* serta menjadi

referensi bagi investor dan peneliti dalam menentukan pendekatan klasifikasi yang optimal pada data pasar saham.

METODE PENELITIAN

Jenis dan Desain Penelitian

Penelitian ini menerapkan pendekatan kuantitatif eksperimental dengan desain komparatif untuk menilai kinerja algoritma SVM, *Logistic Regression*, dan *Random Forest* dalam klasifikasi pergerakan saham PT Timah Tbk (TINS) tahun 2025. Pendekatan kuantitatif digunakan karena evaluasi model dilakukan secara numerik menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, *ROC-AUC*, dan *confusion matrix*, sedangkan pendekatan eksperimental diterapkan melalui pengujian beberapa algoritma dengan dataset dan prosedur evaluasi yang sama agar hasil dapat dibandingkan secara objektif. Penelitian ini juga menganalisis pengaruh optimasi *hyperparameter* menggunakan *GridSearchCV* untuk memperoleh model yang paling stabil dan sesuai dengan karakteristik data saham yang volatil dan cenderung non-linear.



Gambar 1. Alur Metode Penelitian

Studi Literatur

Studi literatur dilakukan untuk membangun landasan teoritis mengenai klasifikasi saham, *machine learning*, *feature engineering* data runtun waktu, dan optimasi *hyperparameter* menggunakan *GridSearchCV* melalui jurnal ilmiah dan sumber akademik terkait. Hasil kajian menunjukkan bahwa performa klasifikasi saham sangat dipengaruhi oleh kualitas data, pemilihan fitur, dan konfigurasi *hyperparameter*, sehingga penelitian ini menggunakan fitur teknikal dan validasi model yang sistematis agar hasil klasifikasi lebih representatif terhadap pola pergerakan saham TINS.

Data Preparation

Tahap data preparation dilakukan untuk menyiapkan data historis saham PT Timah Tbk (TINS) tahun 2025 yang terdiri atas 227 data harian dengan variabel *Open*, *High*, *Low*, *Close*, *Adj Close*, dan *Volume*. Data diurutkan berdasarkan waktu dan diperiksa terhadap *missing value*, duplikasi, serta konsistensi format, dengan hasil tidak ditemukan data hilang maupun duplikat. Selanjutnya, label kelas dibentuk berdasarkan arah perubahan harga penutupan menjadi kelas Naik dan Turun. Setelah proses *feature engineering*, jumlah data yang siap digunakan dalam pemodelan menjadi 218 observasi karena beberapa data awal digunakan untuk pembentukan fitur historis.

Rumus pembentukan label target. Label kelas ditentukan dari perubahan harga penutupan pada periode berikutnya:

$$y_t = \begin{cases} 1. & \text{Jika } Close_{t+1} > Close_t \\ 0. & \text{Jika } Close_{t+1} \leq Close_t \end{cases} \quad (1)$$

Keterangan:

$y_t = 1$ menyatakan kelas Naik, sedangkan $y_t = 0$ menyatakan kelas Turun.

Feature Engineering

Tahap *feature engineering* dilakukan untuk membentuk fitur turunan berbasis return, volatilitas, tren, momentum, dan volume guna menangkap pola temporal saham yang tidak

terlihat pada data mentah. Fitur yang digunakan meliputi *Return_1d*, *Return_3d*, *Return_5d*, *MA5_ratio*, *MA10_ratio*, *MA20_ratio*, *High_Low_Range*, *Open_Close_Gap*, *Volatility_5d*, *Volatility_10d*, *RSI_14*, *MACD*, *MACD_Signal*, *MACD_Hist*, *BB_Width*, *BB_Pos*, dan *Volume_Ratio*. Hasil analisis menunjukkan bahwa fitur berbasis volatilitas dan *range* harga memberikan kontribusi paling informatif terhadap klasifikasi pergerakan saham, yang juga diperkuat oleh hasil *feature importance* pada model *Random Forest* setelah *tuning*.

Return Harian

Return harian digunakan untuk mengukur perubahan relatif harga penutupan dari satu periode ke periode berikutnya.

$$Return_t = \frac{Close_t - Close_{t-1}}{Close_{t-1}} \times 100\% \quad (2)$$

Untuk *return* 3 hari dan 5 hari, perhitungan dilakukan dengan pendekatan rata-rata perubahan harga pada jangka waktu yang lebih panjang.

Moving Average Ratio

Moving average digunakan untuk menangkap kecenderungan tren harga dalam periode tertentu. *Moving average* sederhana dirumuskan sebagai:

$$MA_n = \frac{1}{n} \sum_{i=0}^{n-1} Close_{t-i} \quad (3)$$

Rasio harga terhadap *moving average* digunakan untuk melihat posisi harga saat ini dibandingkan tren historisnya:

$$MA_{ratio_n} = \frac{Close_t}{MA_n} \quad (4)$$

Pada penelitian ini digunakan *MA5_ratio*, *MA10_ratio*, dan *MA20_ratio*.

High-Low Range dan Open-Close Gap

Fitur *High_Low_Range* digunakan untuk menangkap lebar fluktuasi harga harian:

$$High_{LowRange_t} = High_t - Low_t \quad (5)$$

Sedangkan *Open_Close_Gap* digunakan untuk melihat selisih antara harga pembukaan dan penutupan:

$$Open_Close_Gap_t = Close_t - Open_t \quad (6)$$

Kedua fitur ini penting karena mewakili volatilitas *intraday* yang pada hasil penelitian terbukti lebih relevan terhadap arah pergerakan saham.

Volatilitas

Volatilitas digunakan untuk mengukur seberapa besar penyebaran *return* dalam jangka waktu tertentu. Rumus yang digunakan adalah simpangan baku *return* dalam periode *n*:

$$Volatility_n = \sqrt{\frac{1}{n-1} \sum_{i=0}^n (r_i - \bar{r})^2} \quad (7)$$

dengan:

r_i = *return* harian,

$\bar{r}$ = rata-rata *return* selama periode *n*.

Pada penelitian ini digunakan *Volatility_5d* dan *Volatility_10d*.

RSI (Relative Strength Index)

RSI digunakan untuk mengukur kekuatan momentum harga. Rumus RSI adalah:

$$RSI = 100 - \frac{100}{1+RS} \quad (8)$$

dengan:

$$RS = \frac{\text{Average Gain}}{\text{Average Loss}} \quad (9)$$

Fitur ini digunakan karena hasil analisis *feature importance* menunjukkan bahwa RSI termasuk variabel yang berpengaruh terhadap klasifikasi arah pergerakan saham.

MACD (*Moving Average Convergence Divergence*)

MACD dihitung dari selisih dua EMA (*Exponential Moving Average*):

$$MACD_t = EMA_{12} - EMA_{26} \quad (10)$$

Garis sinyal dihitung sebagai EMA dari MACD:

$$Signal_t = EMA_9(MACD_t) \quad (11)$$

Histogram MACD dihitung sebagai:

$$MACD_Hist_t = MACD_t - Signal_t \quad (12)$$

Fitur ini digunakan untuk menangkap momentum dan perubahan tren jangka pendek.

Bollinger Bands

Bollinger Bands dibentuk menggunakan *moving average* dan simpangan baku:

$$Upper_t = MA_{20} + 2\sigma \quad (13)$$

$$Lower_t = MA_{20} - 2\sigma \quad (14)$$

Lebar *Bollinger Bands* dihitung sebagai:

$$BB_Width_t = \frac{Upper_t - Lower_t}{MA_{20}} \quad (15)$$

Posisi harga terhadap *Bollinger Bands* dihitung sebagai:

$$BB_Pos_t = \frac{Close_t - Lower_t}{Upper_t - Lower_t} \quad (16)$$

Fitur ini digunakan untuk menggambarkan kondisi volatilitas dan posisi harga relatif terhadap rentang pergerakannya.

Volume Ratio

Volume ratio digunakan untuk melihat perubahan aktivitas perdagangan:

$$Volume_Ratio_t = \frac{Volume_t}{MA(Volume)_n} \quad (17)$$

Fitur ini berguna untuk menangkap apakah lonjakan harga didukung oleh peningkatan aktivitas transaksi.

Normalisasi Dataset

Setelah *feature engineering*, data dinormalisasi menggunakan *Standardization (Z-score normalization)* untuk menyamakan skala antarfitur dan mencegah dominasi fitur tertentu pada proses pembelajaran model, terutama karena SVM dan *Logistic Regression* sensitif terhadap perbedaan skala data.

Rumus standardisasi adalah:

$$X^1 = \frac{X - \mu}{\sigma} \quad (18)$$

dengan:

X = nilai asli fitur,

μ = rata-rata fitur,
 σ = simpangan baku fitur.

Data Mining

Tahap *data mining* dilakukan untuk membangun model klasifikasi arah pergerakan saham TINS menggunakan SVM, *Logistic Regression*, dan *Random Forest*. Setiap model terlebih dahulu diuji menggunakan parameter default sebagai *baseline*, kemudian dioptimasi menggunakan *hyperparameter tuning* untuk meningkatkan performa klasifikasi.

Support Vector Machine (SVM)

SVM dipilih karena memiliki kemampuan dalam menangani pola data non-linear serta fitur berdimensi tinggi melalui pencarian *hyperplane* optimal untuk memisahkan dua kelas secara maksimal. Berdasarkan hasil penelitian, model ini menunjukkan peningkatan performa setelah *tuning*, terutama pada metrik *recall* dan *F1-score*.

$$\min_{\{\omega, b, \xi\}} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n \xi_i \quad (19)$$

dengan kendala:

$$y_i(\omega^t x_i + b) \geq 1 - \xi_i, \quad \xi \geq 0 \quad (20)$$

Keterangan:

ω = vektor bobot,

b = bias,

ξ_i = *slack variable*,

C = parameter regularisasi.

Logistic Regression

Logistic Regression digunakan karena merupakan algoritma klasifikasi biner yang sederhana, efisien, dan mudah diinterpretasikan melalui fungsi sigmoid untuk menghitung probabilitas kelas. Model ini digunakan sebagai pembandingan terhadap metode berbasis margin dan *ensemble*, serta menunjukkan performa paling stabil dan terbaik secara keseluruhan berdasarkan hasil penelitian.

$$P(y = 1|x) = \frac{1}{1 + e^{-z}} \quad (21)$$

dengan:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (22)$$

Random Forest

Random Forest digunakan sebagai model *ensemble* berbasis *decision tree* yang mampu menangani pola non-linear dan relatif tahan terhadap *overfitting* melalui mekanisme voting mayoritas. Model ini digunakan untuk menguji efektivitas pendekatan *ensemble* pada data saham TINS, dengan hasil menunjukkan *recall* yang cukup tinggi, meskipun *accuracy* dan kemampuan generalisasinya masih berada di bawah *Logistic Regression*.

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_m(x)\} \quad (23)$$

dengan:

$T_m(x)$ = prediksi dari pohon ke- m ,

m = jumlah pohon.

Optimasi Hyperparameter Menggunakan GridSearchCV

Random Forest digunakan sebagai model *ensemble* berbasis *decision tree* yang mampu menangani pola non-linear dan relatif tahan terhadap *overfitting* melalui mekanisme *voting* mayoritas. Model ini digunakan untuk menguji efektivitas pendekatan *ensemble* pada data saham TINS, dan hasil penelitian menunjukkan bahwa *Random Forest* memiliki *recall* yang

cukup baik, tetapi *accuracy* dan generalisasi masih berada di bawah *Logistic Regression* setelah proses *tuning* dan evaluasi data uji.

$$\theta^* = \arg \max_{\{\theta \in \Theta\}} CV - Score(\theta) \quad (24)$$

dengan:

θ = kombinasi *hyperparameter*,

Θ = ruang parameter,

CV-Score = nilai validasi silang.

Evaluasi Model dan Prediksi

Evaluasi model dilakukan dengan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, ROC-AUC, dan *confusion matrix* untuk menilai serta mengukur kemampuan klasifikasi masing-masing algoritma setelah optimasi, karena pada data keuangan kualitas model tidak dapat dinilai hanya berdasarkan *accuracy* saja.

Confusion Matrix

Confusion matrix digunakan untuk menunjukkan hasil prediksi model secara rinci, yaitu:

True Positive (TP): kelas naik diprediksi naik,

True Negative (TN): kelas turun diprediksi turun,

False Positive (FP): kelas turun diprediksi naik,

False Negative (FN): kelas naik diprediksi turun.

Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (25)$$

Precision

$$Precision = \frac{TP}{TP+FN} \quad (26)$$

Recall

$$Recall = \frac{TP}{TP+FN} \quad (27)$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (28)$$

Cross-Validation Accuracy

$$CV - Accuracy = \frac{1}{k} \sum_{i=1}^k Acc_i \quad (29)$$

dengan k adalah jumlah *fold cross-validation*.

ROC-AUC

ROC-AUC digunakan untuk mengukur kemampuan model dalam membedakan kelas Naik dan Turun secara probabilistik melalui kurva *Receiver Operating Characteristic*, di mana semakin tinggi nilai AUC maka semakin baik kemampuan generalisasi model dalam melakukan klasifikasi.

Komparasi Model

Komparasi model dilakukan untuk membandingkan kinerja SVM, *Logistic Regression*, dan *Random Forest* berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC pada data uji. Penentuan model terbaik tidak hanya didasarkan pada *accuracy* tertinggi, tetapi juga pada keseimbangan performa antar kelas serta kemampuan generalisasi model, sehingga nilai ROC-AUC dan kestabilan metrik menjadi pertimbangan utama.

Penentuan Model Terbaik

Tahap akhir penelitian dilakukan dengan membandingkan performa SVM, *Logistic Regression*, dan *Random Forest* sebelum dan sesudah optimasi *hyperparameter* berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan ROC-AUC. Model terbaik dipilih berdasarkan keseimbangan performa dan kemampuan generalisasi, sehingga *Logistic Regression* ditetapkan sebagai model terbaik karena memiliki ROC-AUC tertinggi dan performa paling stabil pada kedua kelas.

HASIL DAN PEMBAHASAN

Hasil Penelitian

Penelitian ini membandingkan kinerja SVM, *Logistic Regression*, dan *Random Forest* dalam klasifikasi pergerakan saham PT Timah Tbk (TINS) tahun 2025 melalui tahapan data *preparation*, *exploratory data analysis*, *feature engineering*, normalisasi, pemodelan *baseline*, optimasi *GridSearchCV*, dan evaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, ROC-AUC, serta confusion *matrix*. Dataset terdiri atas 227 data harian periode 2 Januari–30 Desember 2025 dengan variabel *Open*, *High*, *Low*, *Close*, *Adj Close*, dan Volume. Data tidak memiliki *missing value* maupun duplikasi, dan setelah *feature engineering* jumlah observasi menjadi 218 data. Hasil penelitian menunjukkan bahwa saham TINS tahun 2025 memiliki pergerakan yang sangat dinamis, dengan pola *sideways* pada awal tahun dan lonjakan tajam hingga sekitar 3.500 pada kuartal terakhir, yang mengindikasikan volatilitas tinggi dan hubungan fitur teknikal yang cenderung non-linear.

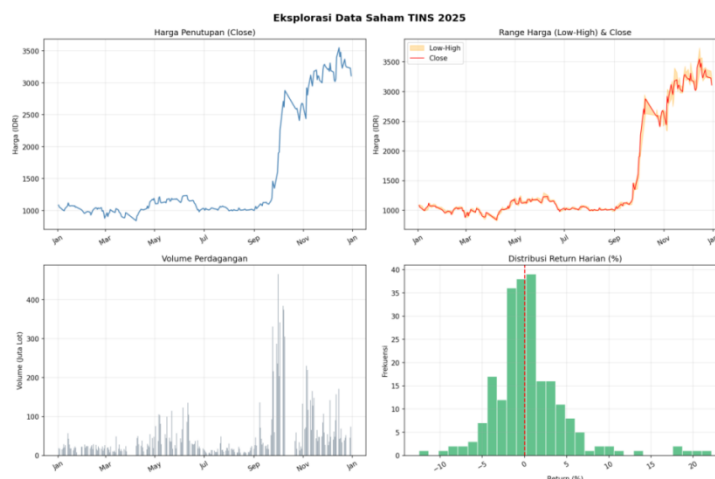
Hasil Exploratory Data Analysis dan Data Preparation

Tahap *exploratory data analysis* menunjukkan bahwa harga penutupan saham TINS meningkat tajam pada akhir tahun dan diikuti kenaikan volume perdagangan, yang menandakan meningkatnya aktivitas pasar. Distribusi target relatif seimbang, yaitu 51,4% kelas turun dan 48,6% kelas naik, sehingga evaluasi model dapat dilakukan secara lebih objektif.

Tabel 1. Ringkasan Dataset Penelitian

Komponen	Nilai
Jumlah data awal	227
Data duplikat	0
Data bilang	0
Data setelah pembersihan	227
Jumlah data siap modeling	218
Kelas Turun	51,4%
Kelas Naik	48,6%

Histogram return harian menunjukkan distribusi yang berpusat di sekitar nol dengan ekor distribusi yang cukup lebar, menandakan adanya perubahan harga ekstrem pada beberapa periode. Selain itu, korelasi fitur terhadap target cenderung lemah, sehingga pergerakan saham TINS tidak mudah dipisahkan secara linear dan memerlukan fitur yang lebih representatif.



Gambar 2. Eksplorasi Data Saham TINS 2025

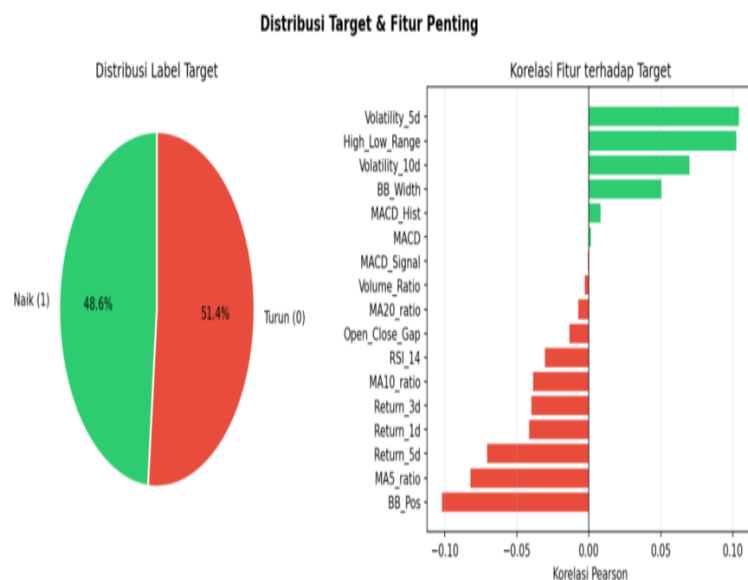
Hasil Feature Engineering

Feature engineering dilakukan untuk menangkap pola temporal dan karakteristik teknik saham yang tidak terlihat langsung pada data mentah. Dalam penelitian ini, fitur yang dibentuk mencakup indikator berbasis return, volatilitas, tren, momentum, dan volume. Fitur-fitur tersebut meliputi:

Tabel 2. Fitur Hasil Feature Engineering

No Fitur	Keterangan
Return_1d	Return harian satu hari sebelumnya
Return_3d	Return rata-rata tiga hari
Return_5d	Return rata-rata lima hari
MA5_ratio	Rasio harga terhadap MA 5 hari
MA10_ratio	Rasio harga terhadap MA 10 hari
MA20_ratio	Rasio harga terhadap MA 20 hari
High_Low_Range	Selisih harga tertinggi dan terendah
Open_Close_Gap	Selisih harga pembukaan dan penutupan
Volatility_5d	Volatilitas 5 hari
Volatility_10d	Volatilitas 10 hari
RSI_14	Relative Strength Index 14 hari
MACD	Moving Average Convergence Divergence
MACD_Signal	Garis sinyal MACD
MACD_Hist	Histogram MACD
BB_Width	Lebar Bollinger Bands
BB_Pos	Posisi harga terhadap Bollinger Bands
Volume_Ratio	Rasio perubahan volume perdagangan

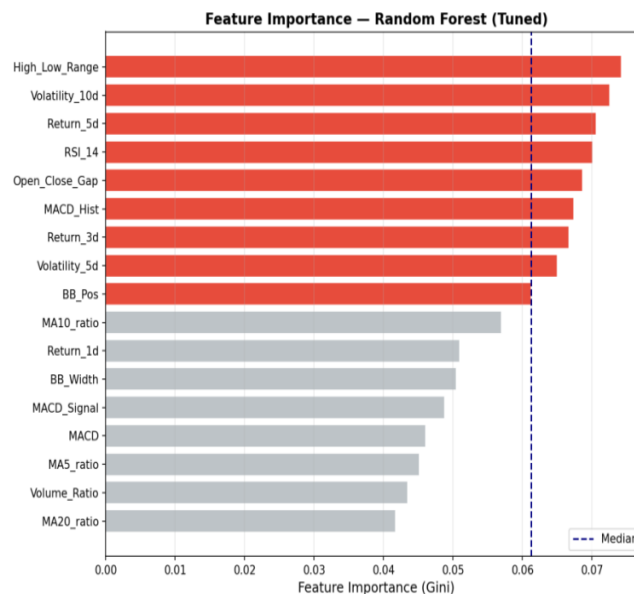
Hasil korelasi menunjukkan bahwa fitur berbasis volatilitas dan range harga, seperti *High_Low_Range*, *Volatility_5d*, dan *Volatility_10d*, memiliki hubungan paling informatif terhadap arah pergerakan saham dibandingkan fitur lainnya, meskipun nilai korelasinya relatif rendah.



Gambar 3. Distribusi Target dan Korelasi Fitur terhadap Target

Temuan ini diperkuat oleh hasil feature importance pada Random Forest setelah *tuning*, di mana fitur dominan meliputi *High_Low_Range*, *Volatility_10d*, *Return_5d*, *RSI_14*, *Open_Close_Gap*, *MACD_Hist*, *Return_3d*, *Volatility_5d*, dan *BB_Pos*. Hasil tersebut

menunjukkan bahwa model lebih dipengaruhi oleh volatilitas dan momentum jangka pendek dibandingkan tren harga jangka panjang.



Gambar 4. Feature Importance Model Random Forest Setelah Tuning

Hasil Normalisasi Dataset

Setelah *feature engineering*, seluruh fitur dinormalisasi menggunakan *Standardization* (*Z-score normalization*) untuk menyamakan skala antarfitur, terutama karena SVM dan *Logistic Regression* sensitif terhadap perbedaan rentang nilai. Proses ini membantu mengurangi dominasi fitur bernilai besar dan meningkatkan kestabilan pembelajaran model, khususnya pada fitur teknikal berbasis volume dan volatilitas yang memiliki rentang nilai sangat berbeda.

Tabel 3. Hasil Normalisasi Dataset

Komponen	Hasil
Metode Normalisasi	Standardization
Rata-rata Fitur	Mendekati 0
Simpangan Baku Fitur	Mendekati 1

Hasil Data Mining: Evaluasi Baseline

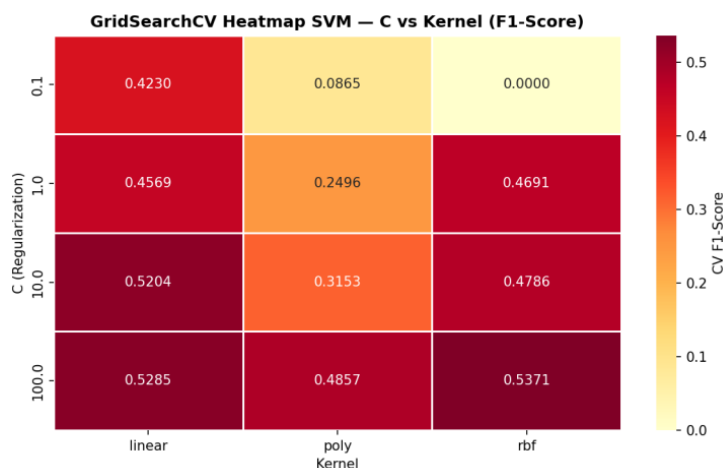
Tahap evaluasi awal menggunakan parameter default menunjukkan bahwa *Logistic Regression* memiliki performa baseline terbaik dibandingkan SVM dan *Random Forest*, dengan *accuracy* sebesar 0,5952 dan *F1-score* 0,4848. SVM masih memiliki *recall* yang rendah sehingga kemampuan mendeteksi kelas positif belum optimal, sedangkan *Random Forest* menunjukkan performa yang relatif moderat. Hasil ini menunjukkan bahwa seluruh model masih memerlukan optimasi lebih lanjut untuk menyesuaikan karakteristik data saham TINS yang kompleks dan volatil.

Tabel 4. Performa Model Baseline

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV-Acc
SVM	0,5238	0,5000	0,2500	0,3333	0,4159	0,5483
Logistic Regression	0,5952	0,6154	0,4000	0,4848	0,6000	0,5431
Random Forest	0,5000	0,4706	0,4000	0,4324	0,5682	0,4636

Hasil Optimasi *Hyperparameter* Menggunakan *GridSearchCV*

Optimasi *hyperparameter* dilakukan untuk memperoleh kombinasi parameter terbaik pada masing-masing model. Hasil *gridsearch* pada SVM memperlihatkan bahwa skor terbaik cenderung muncul pada nilai C yang lebih besar dan kernel RBF, yang mengindikasikan bahwa pemisahan non-linear lebih sesuai untuk data saham ini. Namun, peningkatan performa pada validasi belum sepenuhnya konsisten pada data uji.



Gambar 5. GridSearchCV Heatmap SVM

Hasil evaluasi setelah tuning adalah sebagai berikut:

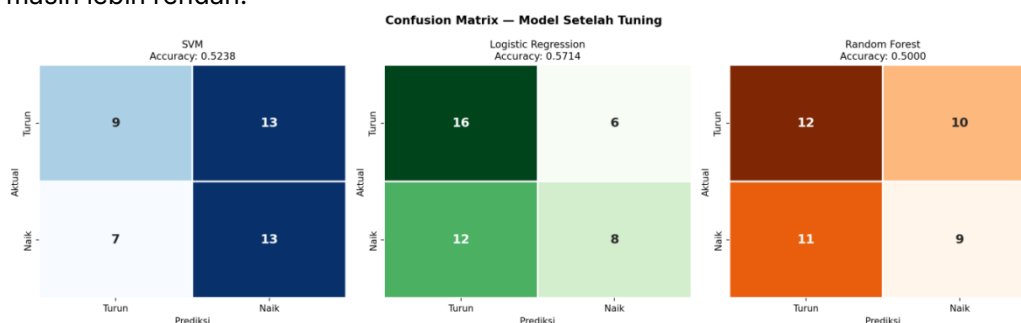
Tabel 5. Performa Model Setelah Tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV-Acc
SVM	0,5238	0,5000	0,6500	0,5652	0,4159	0,5184
Logistic Regression	0,5714	0,5714	0,4000	0,4706	0,6136	0,5970
Random Forest	0,5000	0,4737	0,4500	0,4616	0,5227	0,5182

Hasil *tuning* menunjukkan bahwa peningkatan performa tidak terjadi secara bersamaan pada seluruh metrik. SVM mengalami peningkatan *recall* dan *F1-score* tanpa perubahan *accuracy*, *Logistic Regression* menunjukkan peningkatan ROC-AUC dan *cross-validation accuracy* meskipun *accuracy test* sedikit menurun, sedangkan *Random Forest* hanya mengalami peningkatan performa yang relatif kecil.

Hasil Evaluasi Model dan Prediksi

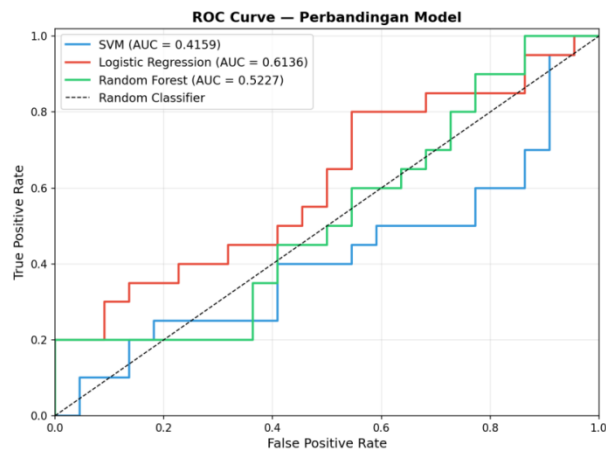
Hasil *confusion matrix* menunjukkan bahwa *Logistic Regression* memberikan keseimbangan prediksi paling stabil, sedangkan SVM cenderung lebih agresif memprediksi kelas naik, dan *Random Forest* memiliki sensitivitas yang cukup seimbang namun *accuracy* yang masih lebih rendah.



Gambar 6. Confusion Matrix Model Setelah Tuning

Berdasarkan *confusion matrix*, *Logistic Regression* menunjukkan performa paling stabil dengan *accuracy* 0,5714, karena mampu memprediksi 16 data turun dan 8 data naik secara benar serta memiliki keseimbangan *precision* dan *recall* yang lebih baik dibandingkan model lain. SVM lebih baik dalam mendeteksi kelas naik dengan 13 prediksi benar, tetapi menghasilkan *false positive* yang cukup tinggi sehingga *accuracy* tetap berada pada 0,5238, sedangkan *Random Forest* menghasilkan distribusi prediksi yang lebih merata dengan *accuracy* 0,5000, namun jumlah kesalahan klasifikasinya masih lebih besar dibandingkan *Logistic Regression*. Hasil tersebut diperkuat oleh ROC Curve Perbandingan Model, di mana Logistic Regression memperoleh ROC-AUC sebesar 0,6136, lebih tinggi dibandingkan *Random Forest* 0,5227 dan SVM 0,4159. Nilai ROC-AUC yang lebih tinggi menunjukkan bahwa *Logistic Regression* memiliki kemampuan pemisahan kelas yang paling baik dalam membedakan

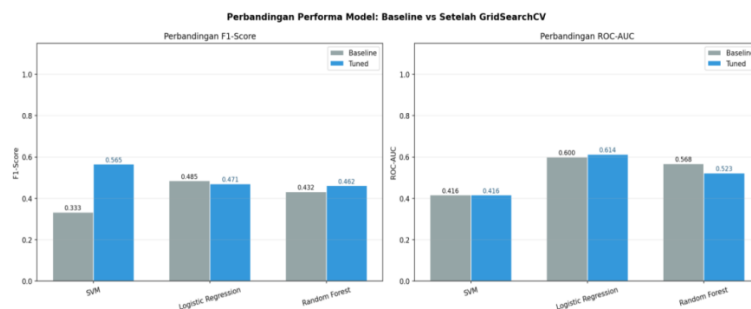
pergerakan saham naik dan turun secara probabilistik, sehingga model ini dinilai paling seimbang dan paling optimal untuk klasifikasi pergerakan saham TINS tahun 2025.



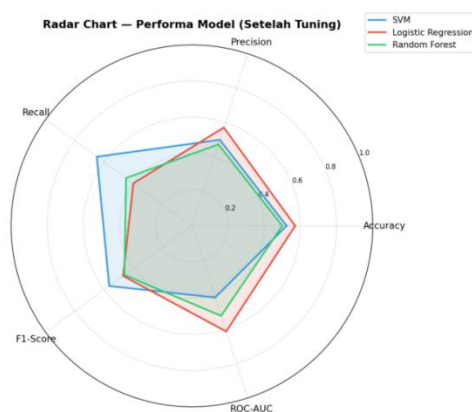
Gambar 7. ROC Curve Perbandingan Model

Komparasi Model

Hasil komparasi menunjukkan bahwa masing-masing algoritma memiliki karakteristik performa yang berbeda. Jika dilihat dari *accuracy* dan ROC-AUC, *Logistic Regression* menjadi model paling baik. Jika dilihat dari *recall*, SVM menunjukkan peningkatan paling besar setelah tuning. Sementara itu, *Random Forest* lebih stabil tetapi belum unggul dalam keseluruhan metrik.



Gambar 8. Perbandingan Performa Model: Baseline vs Setelah GridSearchCV



Gambar 9. Radar Chart Performa Model Setelah Tuning

Ringkasan perbandingan performa dapat dilihat pada tabel berikut.

Tabel 6. Perbandingan Model Terbaik

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	0,5238	0,5000	0,6500	0,5652	0,4159
Logistic Regression	0,5714	0,5714	0,4000	0,4706	0,6136
Random Forest	0,5000	0,4737	0,4500	0,4616	0,5227

Berdasarkan tabel tersebut, *Logistic Regression* merupakan model dengan kinerja paling seimbang karena memberikan kombinasi terbaik antara *accuracy*, *precision*, dan ROC-AUC. SVM unggul pada *recall*, tetapi performa diskriminatifnya masih rendah. *Random Forest* memberikan hasil yang cukup stabil, namun masih belum cukup kuat untuk mengungguli *Logistic Regression* pada data uji.

Pembahasan

Hasil penelitian menunjukkan bahwa saham TINS tahun 2025 memiliki karakteristik yang volatil, kompleks, dan cenderung non-linear, terlihat dari pola *sideways* pada awal tahun dan lonjakan harga tajam hingga sekitar 3.500 pada kuartal terakhir. Dari hasil *exploratory data analysis* dan *feature engineering*, fitur berbasis volatilitas dan *range* harga, seperti High_Low_Range, Volatility_5d, Volatility_10d, Return_5d, RSI_14, Open_Close_Gap, MACD_Hist, dan BB_Pos, terbukti lebih informatif dibandingkan fitur lainnya, sehingga pergerakan saham lebih banyak dipengaruhi oleh dinamika harga jangka pendek daripada tren panjang atau volume perdagangan.

Pada tahap *baseline*, *Logistic Regression* menunjukkan performa awal terbaik. Setelah optimasi *hyperparameter* dengan *GridSearchCV*, SVM memang meningkat pada *recall* dan *F1-score*, tetapi ROC-AUC-nya tetap rendah, sedangkan *Random Forest* hanya mengalami peningkatan yang relatif kecil. *Logistic Regression* menjadi model paling seimbang karena meskipun *accuracy* test sedikit menurun setelah *tuning*, nilai ROC-AUC dan *cross-validation accuracy* justru meningkat, yang menunjukkan generalisasi model lebih baik. Hasil *confusion matrix* dan ROC *curve* juga menguatkan bahwa *Logistic Regression* memiliki keseimbangan *precision-recall* paling stabil serta kemampuan pemisahan kelas terbaik, sehingga model ini paling sesuai untuk klasifikasi pergerakan saham TINS pada dataset penelitian ini.

KESIMPULAN

Secara keseluruhan, penelitian ini menunjukkan bahwa klasifikasi pergerakan saham PT Timah Tbk (TINS) tahun 2025 dapat dilakukan secara lebih efektif melalui kombinasi *feature engineering* dan optimasi *hyperparameter*, terutama karena data yang dianalisis memiliki pola harga yang cepat berubah dan tidak sepenuhnya linier. Dari 227 data harian yang diproses menjadi 218 observasi siap model, fitur-fitur teknikal berbasis volatilitas, rentang harga, dan momentum jangka pendek terbukti paling berperan dalam membentuk keputusan klasifikasi dan memberikan informasi yang lebih relevan dibandingkan fitur lainnya. Hasil pengujian terhadap tiga algoritma menunjukkan bahwa optimasi dengan *GridSearchCV* memberikan dampak yang berbeda pada tiap model, namun *Logistic Regression* tetap menjadi pilihan paling unggul karena mampu menghasilkan performa yang paling seimbang pada data uji, dengan *accuracy* 0,5714, *precision* 0,5714, *F1-score* 0,4706, *cross-validation accuracy* 0,5970, dan ROC-AUC tertinggi sebesar 0,6136. Sementara itu, SVM lebih kuat dalam mendeteksi kelas naik setelah *tuning*, sedangkan *Random Forest* menunjukkan hasil yang cukup stabil tetapi belum mampu melampaui *Logistic Regression* secara keseluruhan. Dengan demikian, penelitian ini menegaskan bahwa *Logistic Regression* paling layak digunakan sebagai model akhir untuk klasifikasi arah pergerakan saham TINS pada dataset 2025 karena memberikan keseimbangan terbaik antara ketepatan prediksi dan kemampuan generalisasi model.

REFERENSI

- Asmara, I. Y., & Rudianto, D. (2024). Analysis of Factors Affecting Stock Price Movement: Case Study on PT Timah Tbk. *Co-Value: Jurnal Ekonomi, Koperasi & Kewirausahaan*, 15(2).
- Cahyani, R. D., & Prasetyaningrum, P. T. (2026). Sentiment Analysis of User Reviews for AI Applications: Evaluating SVM, Logistic Regression, and Random Forest. *Journal of Information Systems and Informatics*, 8(1), 1–27. <https://doi.org/10.63158/journalisi.v8i1.1366>
- Elsayed, S., Salah, A., Elhenawy, I., & Abdellah, M. (2025). Predicting stock prices using ensemble learning techniques. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(2), 1783–1792. <https://doi.org/10.11591/ijece.v15i2.pp1783-1792>

- Fitria, V. A., & Widayanti, L. (2024). Enhancing Accuracy in Stock Price Prediction : The Power of Optimization Algorithms. *Matrik: Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer*, 23(2), 405–418. <https://doi.org/10.30812/matrik.v23i2.3785>
- Hakim, L., Sari, Z., Aristyo, A. R., & Pangestu, S. (2024). Optimizing android program malware classification using GridSearchCV optimized random forest. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 4(2). <https://kinetik.umm.ac.id/index.php/kinetik/article/view/1944>
- Islam, A., Sikder, R., Ishtiaq, S. M., & Sattar, A. (2023). Stock market prediction of Bangladesh using multivariate long short-term memory with sentiment identification. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(5), 5696–5706. <https://doi.org/10.11591/ijece.v13i5.pp5696-5706>
- Jaladri, N. D., & Mulyani, S. (2020). Analysis of Green Accounting Implementation at PT Timah (Company) Tbk. *Jurnal Ilmiah Ilmu Akuntansi Keuangan Dan Pajak*, 4(2), 66–77.
- Kurniawan, P. D., Surjono, H. D., Jati, H., Yogyakarta, U. N., Korespondensi, P., Regression, L., & Forest, R. (2025). Pengembangan Model Hybrid Arima-Machine Learning Untuk Prediksi Harga Saham BCA. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, 12(6), 1425–1432.
- Labiad, B., Berrado, A., & Benabbou, L. (2023). Predicting extreme events in the stock market using generative adversarial networks. *International Journal of Advances in Intelligent Informatics*, 9(2), 218–230. <https://doi.org/10.26555/ijain.v9i2.898>
- Latif, A., Astuti, D. K., & Arianty, R. (2025). Optimasi Model Support Vector Machine (SVM) Menggunakan GridSearchCV untuk Prediksi Harga Penutupan Saham PT Aneka Tambang Tbk (ANTAM). *DECODE: Jurnal Pendidikan Teknologi Informasi*, 5(3), 1132–1144. <https://doi.org/10.51454/decode.v5i3.1416>
- Mazya, S., Tyas, P., Sarno, R., & Rintyarna, B. S. (2024). Analisis Perbandingan Metode Klasifikasi Sentimen Berita Saham: Pendekatan Machine Learning, Deep Learning, Transfer Learning, dan Graf. *Jurnal Penelitian Ipteks*, 9(1), 58–64.
- Misriati, T., & Aryanti, R. (2024). Optimalisasi Random Forest dan Support Vector Machine dengan Hyperparameter GridSearchCV untuk Analisis Sentimen Ulasan PrimaKu. *Journal of Information System Research (JOSH)*, 5(4), 1333–1341. <https://doi.org/10.47065/josh.v5i4.5347>
- Prasetyo, R. A., Saputro, W. T., & Chirzah, D. (2024). Perbandingan Logistic Regression, SVM, dan Random Forest untuk Analisis Sentimen Ulasan Aplikasi Gopay. *Jurnal Informatika: Jurnal Pengembangan IT*, 10(4), 1176–1188. <https://doi.org/10.30591/jpit.v10i4.8796>
- Putu, D., & Nilakusmawati, E. (2024). Penerapan Metode Support Vector Regression (SVR) dengan Algoritma Grid Search Dalam Peramalan Harga Saham. *E-Jurnal Matematika*, 13(2), 94–104. <https://doi.org/10.24843/MTK.2024.v13.i02.p447>
- Raghavendra, S., & J, S. K. (2020). Performance evaluation of random forest with feature selection methods in prediction of diabetes. *International Journal of Electrical and Computer Engineering (IJECE)*, 10(1), 353–359. <https://doi.org/10.11591/ijece.v10i1.pp353-359>
- S, A. R. M., & Nirmala, C. R. (2020). A computing model for trend analysis in stock data stream classification. *Indonesian Journal of Electrical Engineering and Computer Science*, 17(2), 1602–1609. <https://doi.org/10.11591/ijeecs.v17i2.pp1602-1609>
- Saepudin, A., Faqih, A., & Dwilestari, G. (2024). Perbandingan Algoritma Klasifikasi Support Vector Machine, Random Forest dan Logistic Regression Pada Ulasan Shopee. *Jurnal TEKNO KOMPAK*, 18(1), 178–192.
- Sandag, G. A., Klabat, U., & Mononutu, J. A. (2020). Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. *Cogito Smart Journal*, 6(2), 167–178.
- Saud, A. S., & Shakya, S. (2022). Directional movement index based machine learning strategy for predicting stock trading signals. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(4), 4185–4194. <https://doi.org/10.11591/ijece.v12i4.pp4185-4194>
- Sidek, Z., Sakinah, S., Ahmad, S., Hasimah, N., & Teo, I. (2023). Associating deep learning and the news headlines sentiment for Bursa stock price prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, 31(2), 1041–1049. <https://doi.org/10.11591/ijeecs.v31i2.pp1041-1049>