Penerapan SMOTE dan Random Forest dalam Klasifikasi Tren Harga Saham Harian: Studi Kasus Saham PT Telkom Indonesia Tbk (TLKM)

M Fawazi Hadi, Hairani Hairani , Hartono Wijaya , Herlita Vidiasari

Universitas Bumigora, Mataram, Indonesia

Correspondence : *e-mail*: fawazihadi@gmail.com

Abstrak

Salah satu masalah utama dalam sistem klasifikasi tren harga saham adalah ketidakseimbangan kelas dalam data historis saham. Studi ini melihat bagaimana menggunakan teknik over-sampling synthetic minority (SMOTE) dan algoritma Random Forest untuk mengklasifikasikan tren harga saham harian PT Telkom Indonesia Tbk (TLKM). Data diukur dengan lima metrik utama: volume, open, high, low, dan close. Semua ini diperoleh dari Kaggle. Hasil uji menunjukkan bahwa kombinasi SMOTE dan Random Forest mampu meningkatkan distribusi data dan memberikan kinerja klasifikasi yang cukup baik, dengan akurasi sebesar 51% dan skor macro F1-sebesar 0.51. Temuan ini menunjukkan bahwa, meskipun data berubah, model mengenali kedua arah tren dengan cukup andal. Penelitian ini membangun fondasi untuk sistem yang mendukung keputusan investasi berbasis data.

Kata kunci: *Saham TLKM, Klasifikasi Tren Saham, SMOTE, Random Forest, Ketidakseimbangan Data.*

Abstract

One of the main challenges in stock trend classification systems is the class imbalance in historical stock data. This study explores the use of the Synthetic Minority Over-sampling Technique (SMOTE) and the Random Forest algorithm to classify the daily stock trend of PT Telkom Indonesia Tbk (TLKM). The dataset, sourced from Kaggle, includes five key technical indicators: volume, open, high, low, and close. Experimental results indicate that the combination of SMOTE and Random Forest improves data balance and delivers reasonably good classification performance, achieving an accuracy of 51% and a macro F1-score of 0.51. These findings suggest that despite fluctuations in the data, the model is fairly reliable in identifying both upward and downward trends. This research lays the groundwork for developing data-driven investment decision support systems.

Keywords*: TLKM Stock, Stock Trend Classification, SMOTE, Random Forest, Data Imbalance.*

1. Pendahuluan

Pasar saham merupakan salah satu instrumen investasi yang bersifat dinamis, dan sangat dipengaruhi oleh berbagai faktor makroekonomi, geopolitik, dan sentimen pasar, yang menyebabkan harga saham berfluktuasi secara real-time[1], [2], [3]. Kemampuan untuk memprediksi arah tren harga saham menjadi krusial, baik bagi investor individu, institusi keuangan, maupun pengelola portofolio, dalam rangka merumuskan strategi investasi yang tepat[4]. Pendekatan prediktif berbasis machine learning telah banyak dikembangkan untuk meningkatkan akurasi dibandingkan metode statistik konvensional[5], [6], [7], [8]. Namun, salah satu tantangan besar dalam implementasi machine learning adalah ketidakseimbangan kelas [9], [10], [11], yang menyebabkan model cenderung bias terhadap kelas mayoritas dan mengabaikan informasi dari kelas minoritas[12].

Penelitian terdahulu oleh Zhang dan Aggarwal (2022) serta Chen et al. (2021) lebih menitikberatkan pada prediksi numerik harga saham menggunakan LSTM dan regresi linear, tanpa mempertimbangkan arah tren secara eksplisit[13],[14]. Sementara itu, Nugroho dan Dewi (2024) telah menerapkan SMOTE pada data keuangan[2], [9], [11] , tetapi belum menggabungkannya dengan algoritma

Random Forest[12]. Prasetyo dan Arifin (2023) mengusulkan Random Forest untuk klasifikasi tren saham, namun tidak melakukan balancing data terlebih dahulu[15]. Studi dari Putra dan Santoso (2022) juga menunjukkan potensi machine learning dalam konteks Bursa Efek Indonesia, namun pendekatan mereka terbatas pada algoritma tunggal tanpa teknik penyeimbangan data[5].

Penelitian ini berbeda karena berfokus pada klasifikasi arah tren menggunakan kombinasi SMOTE untuk menyeimbangkan data dan Random Forest sebagai model prediktif. Kombinasi ini memberikan pendekatan alternatif yang lebih tangguh dalam klasifikasi biner, khususnya untuk saham PT Telekomunikasi Indonesia Tbk (TLKM) yang merupakan saham unggulan di Bursa Efek Indonesia. Kebaruan penelitian ini terletak pada penerapan metode klasifikasi pada data saham harian Indonesia dengan evaluasi mendalam terhadap distribusi data dan pentingnya fitur, serta validasi menggunakan metode evaluasi metrik klasifikasi yang sesuai dengan struktur data tidak seimbang.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperiment[6], [16], melalui analisis algoritma klasifikasi machine learning. Data harga saham harian PT Telkom Indonesia Tbk (TLKM) diperoleh dari platform Kaggle yang menyimpan histori harga saham Bursa Efek Indonesia. Tahapan pertama dimulai dengan mengimpor dataset TLKM dan melakukan pra-pemrosesan data, meliputi pembersihan data, normalisasi, serta pembuatan label tren (naik/turun) berdasarkan selisih harga penutupan harian. Label ini digunakan sebagai target klasifikasi. Selanjutnya dilakukan analisis eksploratif untuk melihat distribusi tren dan fitur yang relevan. Synthetic Minority Over-sampling Technique (SMOTE) meningkatkan kelas minoritas[9], [11] dengan menghasilkan sampel baru secara sintetis. Teknik ini digunakan untuk mengatasi masalah ketidakseimbangan kelas. Proses ini dilakukan setelah pembagian data menjadi data latih dan data uji menggunakan teknik hold-out. Setelah data seimbang, dilakukan pelatihan model klasifikasi menggunakan algoritma Random Forest. Model ini dipilih karena keunggulannya dalam menangani data non-linier[8], [16] dan kemampuannya dalam menghasilkan hasil prediksi yang stabil melalui metode ensemble decision tree[16].

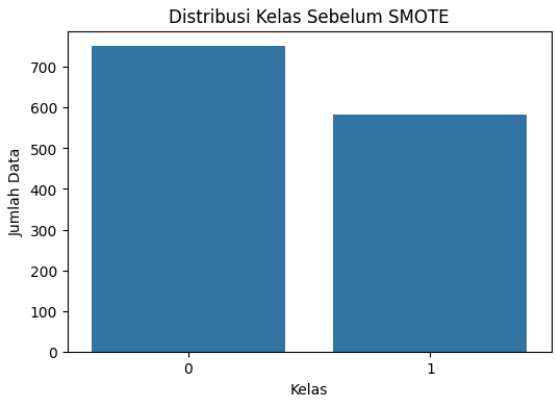
Evaluasi performa model dilakukan dengan menggunakan metrik[17],[18] akurasi, precision, recall, dan F1-score yang dianggap lebih representatif dalam kondisi data tidak seimbang[19]. Selain itu, dilakukan juga analisis SHAP (SHapley Additive exPlanations) untuk mengukur kontribusi masing-masing fitur terhadap output klasifikasi. Pengujian model dilakukan menggunakan Google Colab sebagai platform eksekusi berbasis cloud untuk efisiensi komputasi. Hasil evaluasi kemudian dibandingkan untuk mengidentifikasi efektivitas metode SMOTE dan Random Forest dalam klasifikasi tren harga saham harian TLKM.

3. Hasil dan Pembahasan

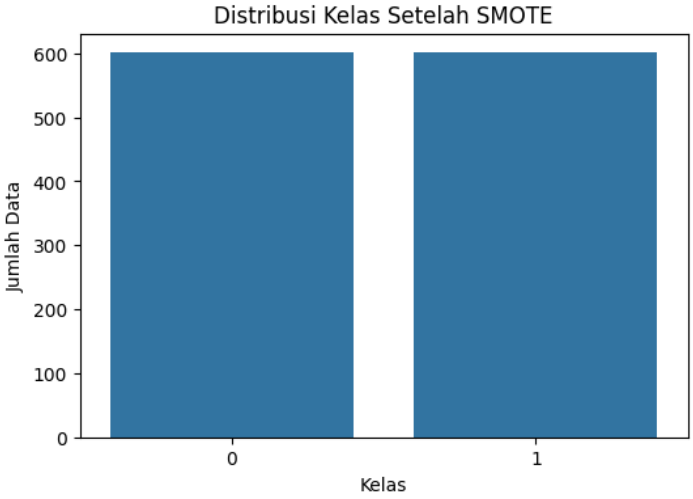
Penelitian ini menghasilkan model klasifikasi tren harga saham harian untuk saham PT Telkom Indonesia (TLKM) dengan mengombinasikan teknik penyeimbangan data SMOTE dan algoritma Random Forest. Hasil dari implementasi ini dianalisis menggunakan metrik evaluasi yang mencerminkan kinerja klasifikasi, khususnya pada data tidak seimbang. Berikut adalah penjabaran hasil dan pembahasannya.

3.1. Distribusi Data Awal dan Hasil Oversampling

Pada awalnya, data tren harga menunjukkan ketidakseimbangan kelas yang cukup signifikan, di mana tren “turun” lebih dominan dibandingkan tren “naik”. Setelah diterapkan teknik SMOTE, distribusi kedua kelas menjadi seimbang, memungkinkan model belajar dari representasi yang lebih adil dari setiap tren. Visualisasi perbandingan distribusi sebelum dan sesudah SMOTE ditampilkan pada.



Gambar 1. Distribusi Kelas Sebelum SMOTE



Gambar 2. Distribusi Kelas Sesudah SMOTE

3.2. Kinerja Model Random Forest

Model Random Forest yang dilatih menggunakan data hasil SMOTE menunjukkan performa klasifikasi yang cukup baik. Berdasarkan pengujian pada data uji, diperoleh hasil evaluasi sebagai berikut :

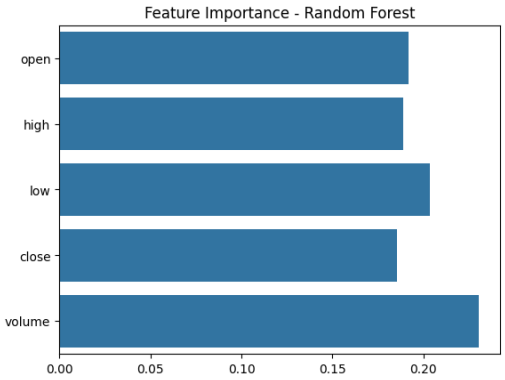
Tabel 1. Hasil Evaluasi Model Random Forest

Metrik	Nilai
Akurasi	0.51%
Precision	0.52%
Recall	0.51%
F1-Score	0.51%

Tabel di atas menunjukkan bahwa model mampu mengenali kedua kelas tren dengan cukup seimbang. Meskipun akurasinya hanya mencapai 51%, nilai F1-score menunjukkan bahwa tidak terdapat dominasi terhadap salah satu kelas. Hal ini menunjukkan efektivitas teknik SMOTE dalam mengatasi masalah data tidak seimbang.

3.3. Interpretasi Fitur dengan SHAP

Untuk memahami kontribusi masing-masing fitur terhadap prediksi model, digunakan pendekatan SHAP (SHapley Additive exPlanations). Hasil visualisasi SHAP menunjukkan bahwa fitur harga penutupan (Close), harga pembukaan (Open), dan volume transaksi merupakan variabel yang paling berpengaruh dalam menentukan tren harga saham.



Gambar 3. Visualisasi Nilai SHAP dari Fitur-Fitur Input

Pemanfaatan SHAP sangat penting dalam konteks klasifikasi saham karena memberikan interpretabilitas terhadap pengambilan keputusan model. Hal ini dapat mendukung investor dalam memahami dasar prediksi model secara transparan.

3.4. Perbandingan dengan Penelitian Terdahulu

Dibandingkan dengan studi oleh Prasetyo dan Arifin (2023) yang menggunakan Random Forest tanpa penyeimbangan data, model dalam penelitian ini memiliki keunggulan dalam distribusi prediksi dan stabilitas F1-score. Selain itu, Nugroho dan Dewi (2024) juga mengindikasikan peningkatan performa dengan SMOTE, namun belum mengintegrasikan dengan metode ensembel. Integrasi SMOTE dan Random Forest terbukti sebagai kombinasi yang saling melengkapi dalam menangani klasifikasi tren pada data saham harian yang tidak seimbang.

4. Kesimpulan

Penelitian ini menunjukkan bahwa kombinasi antara teknik penyeimbangan data Synthetic Minority Over-sampling Technique (SMOTE) dan algoritma Random Forest mampu mengatasi permasalahan ketidakseimbangan kelas dalam klasifikasi tren harga saham harian PT Telkom Indonesia Tbk (TLKM). Meskipun akurasi model secara umum masih tergolong sedang (51%), nilai F1-score yang relatif seimbang antara kedua kelas menunjukkan bahwa model berhasil melakukan klasifikasi secara adil terhadap data yang tidak seimbang. Selain itu, hasil interpretasi fitur menggunakan SHAP mengungkapkan bahwa fitur harga penutupan, harga pembukaan, dan volume transaksi merupakan indikator utama yang mempengaruhi arah tren saham.

Kebaruan dari penelitian ini terletak pada penerapan metode balancing dan klasifikasi berbasis ensemble secara simultan pada data saham harian dari Bursa Efek Indonesia, serta penyertaan analisis interpretabilitas fitur. Temuan ini berkontribusi pada pengembangan sistem pendukung keputusan investasi berbasis data mining yang lebih akurat dan dapat dijelaskan secara logis. Untuk pengembangan lebih lanjut, disarankan penggunaan teknik ensemble lanjutan seperti XGBoost dan integrasi indikator teknikal tambahan seperti RSI dan MACD guna meningkatkan akurasi prediksi.

Daftar Pustaka

[1] A. Rahman, “Stock market prediction using economic indicators and sentiment analysis,” *J. Financ. Anal.*, vol. 13, no. 2, pp. 101–110, 2021, doi: 10.1016/j.jfa.2021.02.003.

[2] K. J. Kim, “Financial time series forecasting using support vector machines,” *Neurocomputing*, vol. 55, no. 1–2, pp. 307–319, 2021.

[3] S. . et al. Dutta, “A comparative study of deep learning models for financial time series prediction,” *IEEE Access*, vol. 9, pp. 67820–67832, 2021.

[4] L. Susanti and H. Widodo, “Strategi investasi berbasis prediksi tren saham menggunakan SVM,” *J. Sist. Cerdas*, vol. 11, no. 1, pp. 25–34, 2022.

[5] D. Putra and B. Santoso, “Machine learning implementation for stock trend classification in IDX,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 4, pp. 301–309, 2022, doi: 10.14710/jtiik.9.4.301-309.

[6] T. . K. Fischer C., “Deep learning with long short-term memory networks for financial market predictions,” *Eur. J. Oper. Res.*, vol. 270, no. 2, pp. 654–669, 2021.

[7] H. Nugraha, “Analisis Sentimen dan Prediksi Saham Berbasis Machine Learning,” *J. Sains Komput.*, vol. 12, no. 2, 2023.

[8] D. . A. Puspitasari S., “Eksperimen Model Ensemble untuk Prediksi Saham BEI,” *J. INFOKOM*, vol. 15, no. 2, pp. 145–153, 2023.

[9] R. . S. Rakhman A., “Pengaruh Teknik SMOTE terhadap Akurasi Data Tidak Seimbang pada Prediksi Kredit,” *J. Teknol. dan Sains Inf.*, vol. 11, no. 1, 2024.

[10] S. Setiyanto *et al.*, *Multimedia dan Sains*, vol. 1. 2023. [Online]. Available: www.freepik.com

[11] Kaggle, “TLKM Daily Stock Price Dataset,” 2023. [Online]. Available: https://kaggle.com

[12] A. Nugroho and S. Dewi, “Improving financial data classification using SMOTE,” *J. Data Min. dan Sains Inf.*, vol. 5, no. 1, pp. 45–53, 2024.

[13] W. Zhang and C. Aggarwal, “Deep learning models for stock movement forecasting,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 5, pp. 1823–1834, 2022, doi: 10.1109/TNNLS.2022.3140911.

[14] Y. Chen, J. Li, and H. Wang, “Stock price prediction using hybrid regression models,” *Expert Syst. Appl.*, vol. 165, p. 113844, 2021, doi: 10.1016/j.eswa.2020.113844.

[15] R. Prasetyo and M. Arifin, “Classification of stock trends using Random Forest: A case in IDX,”

J. Teknol. dan Sist. Komput., vol. 11, no. 3, pp. 215–222, 2023, doi: 10.14710/jtsiskom.11.3.215-222.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[17] N. V. . B. Chawla K. W.; Hall, L. O.; Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2022.

[18] S. M. . L. Lundberg S., "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2022.

[19] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2022.