

SISTEM REKOMENDASI DEPOSITO BERJANGKA MENGGUNAKAN METODE RANDOM FOREST: STUDI KASUS PADA BANK MARKETING UCI

Aurelia¹, Lely Hiryanto²

Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia

E-mail: ¹aurelia.535210031@stu.untar.ac.id, ²lelyh@fti.untar.ac.id

Abstrak

Perkembangan teknologi dan perubahan perilaku konsumen di era digital telah mendorong industri perbankan untuk mengadopsi pendekatan yang lebih inovatif dalam menawarkan produk dan layanan mereka, terutama deposito berjangka. Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi deposito berjangka menggunakan metode Random Forest, dengan studi kasus pada dataset Bank Marketing UCI. Sistem ini dirancang untuk memberikan rekomendasi yang akurat dan personal berdasarkan data historis, preferensi, dan karakteristik individu nasabah. Metode Random Forest digunakan karena kemampuannya dalam menangani data besar dan kompleks, serta mengurangi risiko pemodelan berlebihan. Dataset yang digunakan mencakup 45.211 instances dengan 17 atribut, termasuk informasi demografis nasabah dan riwayat interaksi dengan bank. Hasil penelitian menunjukkan bahwa model Random Forest yang dikembangkan mencapai akurasi 91.56%, dengan presisi 65.47%, recall 52.12%, dan F1-score 58.04%. Analisis variable importance measure mengidentifikasi 'duration', 'age', dan 'balance' sebagai faktor paling berpengaruh dalam keputusan nasabah untuk berlangganan deposito berjangka. Implementasi sistem ini diharapkan dapat meningkatkan efektivitas pemasaran produk deposito berjangka dan meningkatkan kepuasan nasabah.

Kata kunci— Bank Marketing, Machine Learning, Random Forest, Deposito Berjangka, Sistem Rekomendasi

Abstract

Technological advancements and changing consumer behavior in the digital era have pushed the banking industry to adopt more innovative approaches in offering their products and services, particularly time deposits. This research aims to develop a time deposit recommendation system using the Random Forest method, with a case study on the Bank Marketing UCI dataset. The system is designed to provide accurate and personalized recommendations based on historical data, preferences, and individual characteristics of customers. The Random Forest method is used due to its ability to handle large and complex data, as well as reduce the risk of overfitting. The dataset used includes 45,211 instances with 17 attributes, including customer demographic information and interaction history with the bank. The results show that the developed Random Forest model achieved an accuracy of 91.56%, with precision of 65.47%, recall of 52.12%, and F1-score of 58.04%. Variable importance measure analysis identified 'duration', 'age', and 'balance' as the most influential factors in a customer's decision to subscribe to a time deposit. The implementation of this system is expected to improve the effectiveness of time deposit product marketing and increase customer satisfaction.

Keywords - Bank Marketing, Machine Learning, Random Forest, Time Deposit, Recommendation System

1. PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan transformasi digital telah mengubah cara industri perbankan beroperasi dan berinteraksi dengan nasabahnya [1]. Di era digitalisasi yang pesat ini, industri keuangan dituntut untuk dapat memanfaatkan data nasabah secara efektif guna meningkatkan layanan dan produk yang ditawarkan [2]. Salah satu strategi yang kini banyak diadopsi adalah implementasi sistem rekomendasi berbasis kecerdasan buatan untuk menyesuaikan penawaran produk bank, termasuk deposito berjangka [3].

Sistem rekomendasi telah terbukti efektif dalam berbagai industri, termasuk *e-commerce* dan hiburan, untuk meningkatkan kepuasan pelanggan dan mendorong penjualan [4]. Dalam konteks perbankan, teknologi ini membantu mengidentifikasi nasabah potensial untuk produk tertentu berdasarkan profil dan perilaku mereka [5]. Hal ini tidak hanya mengoptimalkan strategi pemasaran, tetapi juga memperbaiki pengalaman nasabah dengan penawaran yang lebih relevan. Beragam algoritma *machine learning* telah diaplikasikan dalam pengembangan sistem rekomendasi perbankan, termasuk *decision trees*, *neural networks*, dan *random forests* [6]. Di antara metode tersebut, Random Forest menunjukkan potensi besar dalam memprediksi perilaku nasabah dan kesuksesan strategi pemasaran bank [7]. Kelebihan *random forest* terletak pada kemampuannya mengelola data kompleks dan multidimensi, serta ketahanannya terhadap overfitting [8].

Meskipun demikian, pengembangan sistem rekomendasi yang efektif untuk produk perbankan, khususnya deposito berjangka, masih menghadapi beberapa tantangan. Salah satunya adalah karakteristik data perbankan yang seringkali bersifat *sparse* atau jarang, terutama untuk produk-produk tertentu [9]. Selain itu, faktor sentimen nasabah dan dinamika pasar keuangan juga perlu dipertimbangkan untuk meningkatkan akurasi rekomendasi.

Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem rekomendasi deposito berjangka menggunakan metode *random forest*, dengan mempertimbangkan berbagai faktor yang mempengaruhi keputusan nasabah. Dengan memanfaatkan dataset Bank Marketing UCI, yang telah digunakan dalam beberapa studi sebelumnya untuk memprediksi keberhasilan strategi pemasaran *telemarketing* bank [1], penelitian ini diharapkan dapat memberikan wawasan baru dalam pengembangan sistem rekomendasi yang lebih akurat dan efektif untuk industri perbankan.

2. METODE PENELITIAN

2.1 Data

Penelitian ini memanfaatkan sebuah *dataset* yang berasal dari strategi pemasaran deposito berjangka oleh sebuah institusi perbankan. *Dataset* ini, yang telah dihimpun dan diteliti oleh Moro et al. (2014), menyajikan informasi menyeluruh mengenai profil nasabah bank serta catatan interaksi mereka dengan pihak bank. Walaupun tujuan awal pengumpulan data ini adalah untuk memprediksi keberhasilan kampanye pemasaran deposito berjangka, namun cakupan informasi di dalamnya sangat bermanfaat untuk pengembangan sistem rekomendasi produk perbankan yang lebih komprehensif.

Kumpulan data pada Tabel 1 memuat beragam atribut yang signifikan untuk menganalisis tingkah laku nasabah dan kemungkinan ketertarikan mereka terhadap produk-produk perbankan. Informasi yang terkandung di dalamnya meliputi data demografis nasabah, rekam jejak finansial,

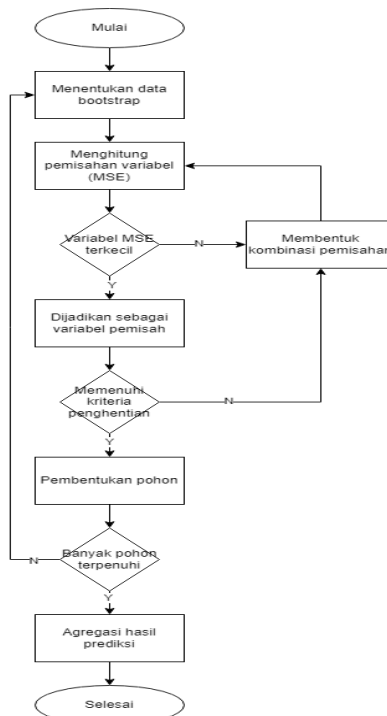
serta detail komunikasi antara nasabah dan bank selama berlangsungnya kampanye pemasaran. Keragaman dan kelengkapan data ini memungkinkan pengembangan model yang dapat mengidentifikasi pola dan preferensi nasabah dengan tingkat ketelitian yang tinggi. Berikut ini adalah penjelasan terperinci mengenai variabel-variabel yang terdapat dalam dataset tersebut:

Tabel 1 Dataset Bank Marketing UCI

Variabel	Deskripsi
Age	Usia nasabah
Job	Jenis pekerjaan
Marital	Status pernikahan
Education	Tingkat pendidikan
Default	Apakah memiliki kredit macet
Balance	Saldo rata-rata tahunan
Housing	Apakah memiliki pinjaman perumahan
Loan	Apakah memiliki pinjaman pribadi
Contact	Jenis komunikasi kontak
Day_Of_Week	Hari terakhir kontak
Month	Bulan terakhir kontak
Duration	Durasi kontak terakhir (dalam detik)
Campaign	Jumlah kontak selama kampanye ini
Pdays	Jumlah hari berlalu sejak nasabah terakhir dikontak
Previous	Jumlah kontak sebelum kampanye ini
Poutcome	Hasil kampanye pemasaran sebelumnya
Y	Apakah nasabah berlangganan deposito berjangka (variabel target)

2.2 Random Forest

Random forest adalah suatu metode klasifikasi yang terdiri dari gabungan beberapa pengklasifikasi yang berbentuk pohon keputusan $\{h(x, \theta_k), k=1, \dots\}$. Di dalamnya, $\{\theta_k\}$ adalah sekumpulan vektor acak yang memiliki distribusi identik dan bersifat independen, di mana setiap pohon memberikan suara untuk kelas yang paling banyak dipilih berdasarkan input x [8].



Gambar 1 Flowchart Random Forest

Langkah-langkah dalam algoritma *Random Forest* ditunjukkan pada Gambar 1 dengan penjelasan sebagai berikut:

1. Menentukan parameter utama, yaitu *mtry*, yang menentukan jumlah variabel prediktor yang diambil secara acak.
2. Menentukan jumlah pohon (*N_tree*) yang ingin dibuat dalam model.
3. Menetapkan kriteria penghentian.
4. Pembagian data menjadi data latih dan uji.
5. Melakukan prediksi dengan membuat model dari kumpulan data baru.
6. Pembentukan pohon keputusan, menggunakan persamaan (1)

$$MSE_n = \frac{1}{N} \sum_{i=1}^N (Y_i - \underline{Y}_n)^2 \quad (1)$$

Keterangan:

MSE_n = Kesalahan kuadrat rata-rata pohon ke-n.

N = Jumlah data pada pohon ke-n

Y_i = Nilai sampel dari data ke-I pada pohon ke-n.

$\underline{Y}_n$ = Nilai rata-rata sampel dari pohon ke-n.

7. Pengulangan proses hingga terbentuk *N_tree* pohon yang diinginkan.
8. Menentukan hasil prediksi dengan menggabungkan hasil dari setiap pohon, menggunakan persamaan (2)

$$\hat{Y}_i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{Y}_n \quad (2)$$

Keterangan:

$\hat{Y}_i$ = Prediksi akhir untuk data ke-i

N_{tree} = Jumlah total pohon dalam *random forest*

$\hat{Y}_n$ = Prediksi dari pohon keputusan ke-n

2.3 CART (*Classification and Regression Tree*)

Metode pohon klasifikasi dan regresi (CART) merupakan salah satu teknik statistik tradisional yang diaplikasikan untuk mengatasi permasalahan klasifikasi dan regresi [10]. Metode ini tergolong dalam pendekatan non-parametrik dan non-linier, di mana ruang fitur terbagi menjadi beberapa segmen. Pada setiap segmen tersebut, diterapkan model yang sederhana, seperti konstanta [11].

2.4 *Bootstrap Aggregating (Bagging)*

Bagging merupakan merupakan sebuah metode ensemble yang mengombinasikan beragam model untuk meningkatkan ketepatan prediksi. Metode ini beroperasi dengan cara mengambil sampel secara acak dari dataset awal dengan pengembalian, sehingga menghasilkan dataset baru [12]. Proses ini bertujuan untuk memperkuat atau memperbaiki akurasi hasil prediksi melalui penggabungan berbagai model.

2.5 Validasi Model

Validasi model merupakan proses yang dilakukan untuk memastikan bahwa model yang dibangun memiliki kinerja yang baik. Salah satu ukuran akurasi yang paling umum digunakan untuk menilai keakuratan sebuah prediksi atau ramalan adalah *confusion matrix*, dengan menggunakan persamaan (3), (4), (5), dan (6). Metrik-metrik seperti *precision*, *recall*, *accuracy*, dan *F-score* untuk setiap kelas dari matriks kekeliruan ini dan digunakan untuk mengevaluasi kinerja keseluruhan pengkategorian [13].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (6)$$

Keterangan:

TP (*True Positive*) = Model benar memprediksi positif

TN (*True Negative*) = Model benar memprediksi negatif

FP (*False Positive*) = Model salah memprediksi positif (seharusnya negatif)

FN (*False Negative*) = Model salah memprediksi negatif (seharusnya positif)

2.6 Variable Importance Measure

Metode *Variable Importance Measure* (VIM) digunakan untuk mengidentifikasi variabel lingkungan yang signifikan bagi algoritma *random forest*, dengan menggunakan persamaan (7). Metode ini menilai pentingnya variabel lingkungan dalam proses pemodelan dengan mengukur seberapa besar variabel tersebut dapat mengurangi akurasi model [14].

$$\%IncMSE = \frac{(MSE_{permuted} - MSE)}{MSE} \times 100\% \quad (7)$$

Keterangan:

$MSE_{permuted}$ = Kesalahan kuadrat rata-rata setelah variabel dipermutasi

MSE = Kesalahan kuadrat rata-rata dari model asli

$\%IncMSE$ = Peningkatan persentase kesalahan setelah permutasi variabel

3. HASIL DAN PEMBAHASAN

3.1 Persiapan Data

Penelitian ini menggunakan dataset *Bank Marketing* dari *UCI Machine Learning Repository*, yang berisi 45.211 *instances* dengan 17 atribut. Dataset ini mencakup informasi tentang kampanye pemasaran deposito berjangka yang dilakukan oleh sebuah institusi perbankan di Portugal.

Langkah pertama dalam analisis adalah melakukan pra-pemrosesan data, termasuk penanganan nilai yang hilang dan encoding variabel kategorikal. Setelah itu, *dataset* dibagi menjadi data latih (75%) dan data uji (25%) untuk membangun dan mengevaluasi *model random forest*.

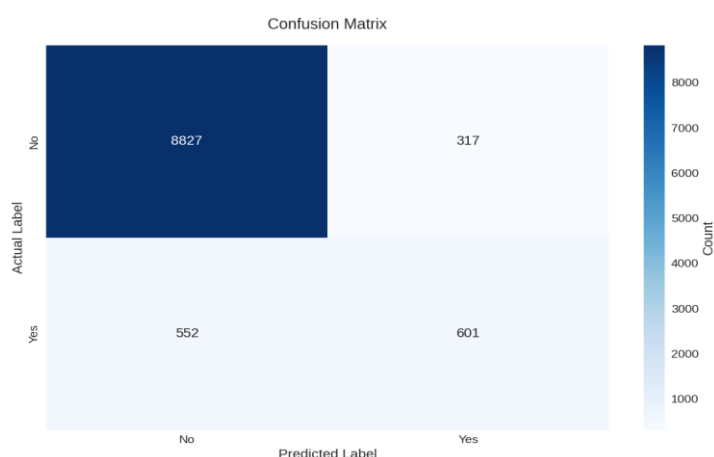
3.2 Pembangunan Model *Random Forest*

Pada tahap ini, pembangunan model *random forest* dilakukan dengan tujuan untuk menghasilkan prediksi yang akurat dalam memberikan rekomendasi berdasarkan data yang tersedia. Model ini dibangun menggunakan 100 pohon keputusan atau dikenal sebagai $n_estimators=100$. Semakin banyak pohon yang digunakan, semakin baik model dapat mengurangi varians dan menghindari *overfitting*, sehingga hasil prediksi menjadi lebih stabil.

Selain itu, parameter *mtry* yang menentukan jumlah fitur yang dipertimbangkan pada setiap *split* diatur sebagai akar kuadrat dari jumlah total fitur. Pemilihan ini didasarkan pada rekomendasi umum untuk model *random forest*, yang menyarankan penggunaan akar kuadrat dari total fitur sebagai pendekatan optimal untuk menjaga keseimbangan antara bias dan varians. Dengan mempertimbangkan hanya sebagian dari fitur pada setiap *split*, model dapat menghindari dominasi fitur tertentu.

3.3 Evaluasi Model

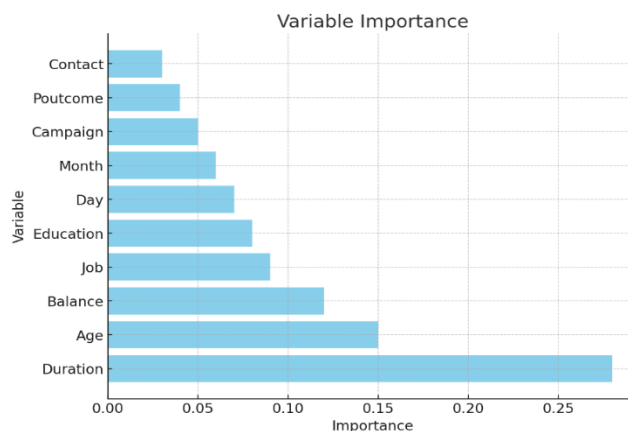
Setelah pelatihan model, evaluasi dilakukan menggunakan data uji. Metrik utama yang digunakan adalah akurasi, presisi, *recall*, dan *F1-score*. Berdasarkan hasil evaluasi pada Gambar 2, menunjukkan bahwa model memiliki akurasi sebesar 91.56%, presisi 65.47%, *recall* 52.12%, dan *F1-score* 58.04%. Ini menunjukkan bahwa model memiliki kinerja yang baik dalam memprediksi apakah seorang nasabah akan berlangganan deposito berjangka atau tidak.



Gambar 2 Evaluasi Model dengan *Confusion Matrix*

3.4 Analisis *Variable Importance Measure* (VIM)

Selanjutnya, analisis *Variable Importance Measure* (VIM) dilakukan untuk mengidentifikasi variabel yang paling berpengaruh dalam prediksi. Berdasarkan grafik pada Gambar 3, menunjukkan hasil bahwa '*duration*' (durasi panggilan terakhir), '*age*' (usia nasabah), dan '*balance*' (saldo rekening) adalah tiga variabel teratas yang mempengaruhi keputusan nasabah untuk berlangganan deposito berjangka.



Gambar 3 Grafik Pengaruh Variabel Terhadap Rekomendasi

Grafik di atas menunjukkan tingkat kepentingan relatif dari berbagai variabel dalam model. Terlihat bahwa '*duration*' memiliki pengaruh yang sangat signifikan, diikuti oleh '*age*' dan '*balance*'.

3.5 Interpretasi Hasil

Berdasarkan hasil analisis, sistem rekomendasi yang dikembangkan dapat memberikan saran yang personal kepada nasabah mengenai produk deposito berjangka. Misalnya, nasabah dengan durasi panggilan yang lebih lama, usia yang lebih tua, dan saldo rekening yang lebih tinggi cenderung lebih mungkin tertarik pada produk deposito berjangka. Sistem dapat menggunakan informasi ini untuk menyesuaikan strategi pemasaran dan meningkatkan efektivitas kampanye.

3.6 Analisis Hasil Penelitian

Hasil penelitian menunjukkan bahwa sistem rekomendasi deposito berjangka yang dikembangkan menggunakan metode *Random Forest* memberikan kinerja yang sangat memuaskan. Model yang dibangun dengan 100 pohon keputusan dan menggunakan dataset Bank Marketing UCI yang terdiri dari 45,211 *instances* dengan 17 atribut, berhasil mencapai akurasi sebesar 91.56%. Metrik evaluasi lainnya juga menunjukkan hasil yang baik, dengan presisi 65.47%, recall 52.12%, dan F1-score 58.04%.

Analisis *Variable Importance Measure* (VIM) mengungkapkan bahwa tiga faktor yang paling berpengaruh dalam keputusan nasabah adalah durasi panggilan terakhir (*duration*), usia nasabah (*age*), dan saldo rekening (*balance*). Temuan ini memberikan wawasan berharga bagi institusi perbankan dalam merancang strategi pemasaran yang lebih efektif dan personal. Nasabah dengan durasi panggilan yang lebih lama, usia yang lebih matang, dan saldo rekening yang lebih tinggi cenderung menunjukkan ketertarikan lebih besar terhadap produk deposito berjangka.

Meskipun penelitian ini menunjukkan hasil yang menjanjikan, terdapat beberapa keterbatasan yang perlu diperhatikan. Dataset yang digunakan berasal dari satu institusi perbankan di Portugal dan model ini berbasis data historis yang memerlukan pembaruan berkala. Untuk pengembangan ke depan, beberapa aspek yang dapat dieksplorasi meliputi integrasi data *real-time*, perbandingan dengan metode *machine learning* lainnya, dan adaptasi model untuk produk perbankan lain.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem rekomendasi deposito berjangka menggunakan metode *Random Forest* dengan kinerja yang sangat baik, ditunjukkan dengan akurasi 91.56%. Analisis *Variable Importance Measure* (VIM) berhasil mengidentifikasi tiga faktor utama yang mempengaruhi keputusan nasabah dalam berlangganan deposito berjangka yaitu durasi panggilan terakhir, usia nasabah, dan saldo rekening.

Dataset yang digunakan berasal dari satu institusi perbankan di Portugal, sehingga mungkin tidak sepenuhnya mewakili karakteristik nasabah di lokasi atau konteks perbankan lain. Selain itu, model ini didasarkan pada data historis dan mungkin perlu penyesuaian berkala untuk mencerminkan perubahan dalam tren pasar atau perilaku konsumen.

Untuk pengembangan selanjutnya, beberapa arah penelitian yang dapat dipertimbangkan adalah mengintegrasikan data *real-time* ke dalam model, membandingkan kinerja *Random Forest* dengan metode *machine learning* lainnya, serta mengadaptasi model untuk produk perbankan lain. Secara keseluruhan, penelitian ini memberikan kontribusi signifikan dalam pengembangan

system rekomendasi yang dapat meningkatkan efektivitas pemasaran produk deposito berjangka dan berpotensi meningkatkan kepuasan nasabah dalam industri perbankan.

DAFTAR PUSTAKA

- [1] P. C. P. R. Sergio Moro, "A data-driven approach to predict the success of bank telemarketing," *Decision Support System*, vol. 62, pp. 22-31, 2014.
- [2] Y. M. M. U. N. Peter Appiahene, "Predicting Bank Operational Efficiency Using Machine Learning Algorithm: Comparative Study of Decision Tree, Random Forest, and Neural Networks," *Advances in Fuzzy Systems*, vol. 2020, no. 1, 2020.
- [3] T. C. Jia-Lang Seng, "An analytic approach to select data mining for business decision," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8042-8057, 2010.
- [4] X. C. Yongfeng Zhang, "Explainable Recommendation: : A Survey and New Perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1-101, 2020.
- [5] S. N. A.-S. Lesly Alejandra Gonzalez Camacho, "Social network data to alleviate cold-start in recommender system: A systematic review," *Information Processing & Management*, vol. 54, no. 4, pp. 529-544, 2018.
- [6] S. S. Mamta Kalra, "Machine Learning and Deep Learning Techniques for Recommendation Systems: A Comprehensive Review," *Harbin Engineering University*, vol. 45, no. 5, pp. 158-172, 2024.
- [7] S. S. K. M. Martin Leo, "Machine Learning in Banking Risk Management: A Literature Review," *Risks*, vol. 7, no. 1, pp. 1-22, 2019.
- [8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [9] P. Y.-x. X. J.-b. Xu Yong, "The Improvement of Personalized Recommendation System Based on Sparse Data of Financial Products," *World Journal of Research and Review*, vol. 7, no. 2, pp. 19-25, 2018.
- [10] J. F. R. O. C. J. S. Leo Breiman, *Classification and Regression Trees*, New York: Taylor & Francis, 1984.
- [11] R. T. J. F. Trevor Hastie, Ruppert, David. "The elements of statistical learning: data mining, inference, and prediction, New York: Springer, 2017.
- [12] S. P. A. C. Yogo Aryo Jatmiko, "Analisis Perbandingan Kinerja CART Konvensional, Bagging dan Random Forest Pada Klasifikasi Objek: Hasil Dari Dua Simulasi," *Media Statistika*, vol. 12, no. 1, pp. 1-12, 2019.
- [13] M.-L. Z. Zhi-Hua Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819-1837, 2014.
- [14] R. H. Z. S. W. Z. Shufang Song, "Variable Importance Measure System Based on Advanced Random Forest," *Computer Modeling in Engineering & Sciences*, vol. 128, no. 1, pp. 65-85, 2021.