



Prediksi Kelulusan Mahasiswa Menggunakan Random Forest Dan Smote Untuk Mengatasi Imbalanced Data

Fahrialdy Febriansyah Putra^{1*}, Sulthan Fakhri Wibowo², Jelita Rahmah Zebua³

¹⁻³Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Indonesia

Email: fahrialdyfebrians@gmail.com, sulthanfakhri1234@gmail.com, jelitarahmahzebua@gmail.com

*Penulis Korespondensi : Fahrialdy Febriansyah Putra

Abstract. Higher education institutions face challenges in identifying student graduation risks early due to academic data complexity and imbalanced data constraints. This study aims to build an accurate graduation prediction model by combining the Random Forest algorithm and the SMOTE technique to handle sample gaps between majority and minority classes. The method employed is a quantitative Data Mining experiment using Python in a Google Colab environment. A dataset of 1,500 academic records was processed through preprocessing, SMOTE balancing, and Random Forest ensemble classification. Experimental results demonstrate outstanding performance with a 98.75% accuracy and stable evaluation metrics. Feature importance analysis reveals that Assignment Scores and E-Learning Activity are the most influential predictors of graduation status. This model holds significant potential as an Early Warning System to help institutions mitigate graduation delays automatically and efficiently.

Keywords: Imbalanced Data, Random Forest, SMOTE, Student Graduation.

Abstrak. Institusi pendidikan tinggi masih menghadapi tantangan dalam mengidentifikasi risiko keterlambatan kelulusan mahasiswa secara dini, terutama akibat kompleksitas data akademik dan kendala imbalanced data. Penelitian ini bertujuan untuk membangun model prediksi kelulusan mahasiswa yang akurat dengan mengombinasikan algoritma Random Forest dan teknik SMOTE guna menangani kesenjangan jumlah sampel antara kelas mayoritas dan minoritas. Metode yang digunakan adalah eksperimen kuantitatif berbasis Data Mining dengan memanfaatkan bahasa pemrograman Python di lingkungan Google Colab. Dataset yang terdiri dari 1.500 catatan akademik diproses melalui tahap pra-pemrosesan, penyeimbangan data menggunakan SMOTE, hingga klasifikasi menggunakan mekanisme ensemble learning pada Random Forest. Hasil eksperimen menunjukkan performa yang sangat memuaskan dengan tingkat akurasi mencapai 98,75% serta nilai precision, recall, dan f1-score yang stabil. Analisis kepentingan atribut mengungkapkan bahwa Nilai Tugas dan Aktivitas E-Learning merupakan indikator prediktor yang paling berpengaruh dalam menentukan status kelulusan. Model ini memiliki potensi besar untuk diimplementasikan sebagai sistem peringatan dini (Early Warning System) guna membantu institusi dalam memitigasi angka keterlambatan kelulusan secara otomatis dan efisien..

Kata kunci: Kelulusan Mahasiswa; Ketidakseimbangan Data; Random Forest; SMOTE.

1. LATAR BELAKANG

Kelulusan mahasiswa tepat waktu merupakan salah satu indikator utama dalam menilai kualitas, efektivitas, dan efisiensi suatu institusi pendidikan tinggi. Tingkat kelulusan yang tinggi mencerminkan keberhasilan proses pembelajaran, sistem akademik, dan layanan pendidikan yang berjalan secara optimal (Wijyaningrum et al., 2024). Sebaliknya, fenomena keterlambatan kelulusan berisiko menurunkan kualitas akademik, menghambat perolehan akreditasi program studi, hingga memperburuk

reputasi perguruan tinggi di mata masyarakat (Anju Fauziah & Julan Hernadi, 2025). Oleh karena itu, diperlukan sebuah mekanisme pemantauan yang mampu mengidentifikasi risiko keterlambatan mahasiswa sedini mungkin.

Dalam beberapa tahun terakhir, transformasi digital di sektor pendidikan telah mendorong pemanfaatan Educational Data Mining (EDM). Melalui pendekatan machine learning, EDM mampu menggali pola tersembunyi dari data akademik seperti Indeks Prestasi Kumulatif (IPK), jumlah SKS yang diambil, nilai mata kuliah kunci, hingga tingkat kehadiran (Mukherjee & Khushi, 2021). Pemanfaatan data ini memungkinkan pihak manajemen perguruan tinggi untuk melakukan intervensi dini dan mengambil keputusan strategis guna membantu mahasiswa yang memiliki potensi keterlambatan lulus. Machine learning merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk belajar dari data tanpa harus diprogram secara eksplisit (Siswara et al., 2024).

Sejumlah studi literatur menunjukkan bahwa algoritma Random Forest memiliki performa yang sangat baik dalam klasifikasi data pendidikan. Keunggulan utamanya terletak pada kemampuan menangani data yang kompleks, mengurangi risiko overfitting, serta memiliki kemampuan generalisasi yang tinggi (Pratiwi & Utami, 2025). Namun, implementasi Random Forest pada data kelulusan sering kali terbentur oleh kendala ketidakseimbangan data (imbalanced data) (Hamdani et al., 2025). Pada dataset pendidikan, jumlah mahasiswa yang lulus tepat waktu biasanya jauh lebih besar dibandingkan mahasiswa yang terlambat, sehingga model cenderung bias dan gagal memprediksi mahasiswa yang berisiko terlambat dengan akurat.

Untuk mengatasi permasalahan tersebut, teknik Synthetic Minority Oversampling Technique (SMOTE) hadir sebagai solusi untuk menyeimbangkan distribusi dataset melalui pembangkitan data sintesis pada kelas minoritas (Sidik & Ansawarman, 2022). Integrasi SMOTE diharapkan dapat meningkatkan performa model, khususnya pada metrik precision, recall, dan F1-score, sehingga prediksi terhadap mahasiswa yang berisiko terlambat menjadi lebih presisi (Tedut et al., 2025).

Berdasarkan urgensi tersebut, penelitian ini berfokus pada pengembangan model prediksi kelulusan mahasiswa dengan mengombinasikan algoritma Random Forest dan teknik SMOTE (Anjani et al., 2023). Pemilihan Python di lingkungan Google Colab

didasarkan pada ketersediaan ekosistem pustaka machine learning yang komprehensif, seperti scikit-learn dan imbalanced-learn, yang memungkinkan pembangunan alur kerja otomatisasi data secara stabil dan akurat (Alnassar et al., 2026). Penelitian ini bertujuan untuk menguji sejauh mana penanganan imbalanced data melalui skrip pemrograman tersebut dapat meningkatkan akurasi dan objektivitas model dalam memprediksi status kelulusan (Hairani, 2021). Hasil dari penelitian ini diharapkan dapat menjadi rujukan bagi institusi pendidikan dalam memitigasi angka keterlambatan kelulusan secara otomatis, efisien, dan akurat.

2. KAJIAN TEORITIS

Tinjauan pustaka dilakukan untuk memperoleh pemahaman menyeluruh mengenai teknologi yang digunakan serta konteks permasalahan. Studi ini mencakup eksplorasi terhadap faktor akademik yang memengaruhi kelulusan mahasiswa, prinsip kerja klasifikasi data menggunakan machine learning, dasar teori algoritma Random Forest, serta efektivitas teknik SMOTE dalam menangani imbalanced data (Diantika et al., 2024).

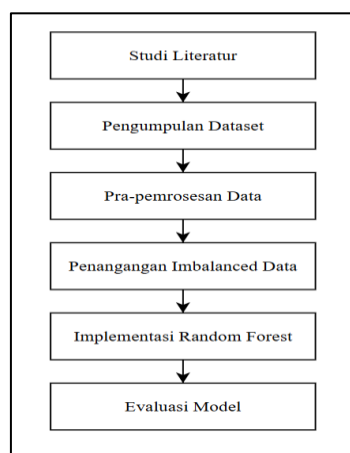
Penggunaan algoritma klasifikasi konvensional pada data kelulusan mahasiswa yang tidak seimbang sering kali menghasilkan performa yang kurang optimal. Namun, setelah diintegrasikan dengan teknik SMOTE, akurasi model meningkat signifikan sebesar 3%, dan nilai sensitivity serta precision juga menunjukkan perbaikan yang stabil dalam mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu (Anggrawan et al., 2023). Pendekatan berbasis Random Forest untuk prediksi kelulusan di perguruan tinggi Indonesia mampu mencapai akurasi sebesar 87,5% dengan memanfaatkan fitur utama seperti IPK, tingkat kehadiran, dan nilai mata kuliah inti (Andrianof et al., 2025). Keunggulan model ini terletak pada kemampuannya menangani data yang rumit dan memberikan stabilitas prediksi melalui mekanisme ensemble learning (Nurhalizah et al., 2024).

Eksplorasi efektivitas penanganan data tidak seimbang juga dibuktikan melalui komparasi algoritma Random Forest dan Support Vector Classifier (SVC) (Rohman et al., 2025). Berdasarkan pengujian pada 1.298 catatan mahasiswa, Random Forest menunjukkan performa yang unggul dalam hal akurasi dan metrik ROC-AUC, membuktikan bahwa algoritma ini sangat tangguh dalam mengklasifikasikan pola akademik yang kompleks (Ngaeni, 2025). Penerapan SMOTE pada klasifikasi nilai

mahasiswa memberikan hasil yang sangat akurat serta berhasil menyeimbangkan distribusi data kelas minoritas tanpa kehilangan informasi penting (Rekha et al., 2023). Melalui kajian teoretis ini, integrasi penanganan data tidak seimbang dengan seleksi fitur pada algoritma Random Forest terbukti sangat krusial dalam domain pendidikan untuk mendeteksi kelas minoritas secara presisi (Kusuma & Firmansyah, 2022).

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen kuantitatif dengan menerapkan teknik Data Mining dan algoritma pembelajaran mesin. Tujuan utama dari penelitian ini adalah untuk membangun serta menguji model prediksi kelulusan mahasiswa dengan mengombinasikan algoritma Random Forest dan teknik SMOTE guna mengatasi kendala ketidakseimbangan data. Seluruh tahapan eksperimen diimplementasikan menggunakan bahasa pemrograman Python pada lingkungan Google Colab dengan memanfaatkan pustaka Scikit-learn, Imbalanced-learn, Pandas, dan NumPy. Alur penelitian disusun secara sistematis melalui beberapa tahapan penting yang meliputi studi literatur, pengumpulan dataset, pra-pemrosesan data, penanganan imbalanced data, implementasi Random Forest, hingga evaluasi model.



Gambar 1. Metode Penelitian

Data yang digunakan berasal dari dataset publik bertajuk “Prediksi Status Akademik” dari situs Kaggle dengan total 1.500 catatan akademik mahasiswa. Fitur-fitur yang digunakan mencakup persentase kehadiran, skor partisipasi diskusi, rata-rata nilai tugas, skor aktivitas e-learning, dan Indeks Prestasi Kumulatif (IPK). Hasil analisis awal menunjukkan data tidak seimbang dengan jumlah kelas “Lulus” sebanyak 1.399 data dan

kelas “Tidak” sebanyak 101 data. Pra-pemrosesan data dilakukan melalui validasi missing values, outliers, dan penghapusan data duplikat. Fitur kategorikal dikonversi menjadi numerik menggunakan label encoding (Ngaeni, 2025). Dataset kemudian dibagi dengan proporsi 80% sebagai data latih (1.200 data) dan 20% sebagai data uji (300 data) menggunakan stratified sampling untuk menjaga konsistensi proporsi kelas.

Penanganan imbalanced data direalisasikan lewat teknik SMOTE yang bekerja dengan membangkitkan sampel sintesis baru pada kelas minoritas berdasarkan prinsip tetangga terdekat (k-nearest neighbors). Setelah data seimbang, algoritma Random Forest diterapkan melalui mekanisme ensemble learning. Prediksi akhir ditentukan lewat majority voting dari sekumpulan pohon keputusan (decision trees) yang dibangun dari bootstrap sampling. Hasil prediksi akhir model merupakan pembagian dari jumlah total pohon di dalam hutan terhadap akumulasi penjumlahan hasil prediksi dari seluruh pohon keputusan yang terbentuk (Jawad et al., 2022). Evaluasi performa model diukur berdasarkan metrik akurasi, precision, recall, dan F1-score melalui confusion matrix..

4. HASIL DAN PEMBAHASAN

4.1 Hasil

Penelitian klasifikasi status akademik ini menghasilkan model prediksi menggunakan algoritma Random Forest yang diintegrasikan dengan teknik penyeimbangan data SMOTE. Proses eksperimen dilakukan untuk mengevaluasi performa model dalam menangani ketidakseimbangan data riil mahasiswa.

4.1.1 Evaluasi Performa Model

Berdasarkan hasil pengujian terhadap data uji, model menunjukkan performa yang sangat baik. Hasil metrik evaluasi utama dari laporan klasifikasi menunjukkan pencapaian tingkat akurasi model sebesar 98,75%.

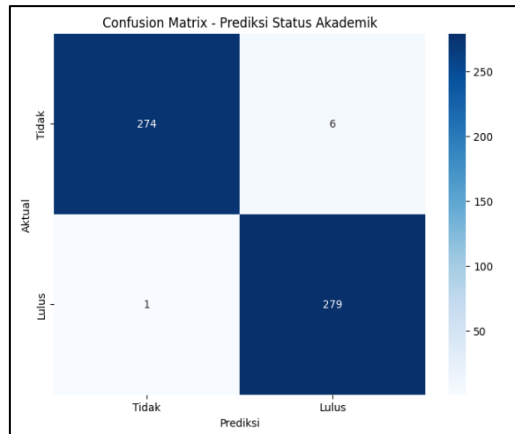
Akurasi Model: 98.75%					
Classification Report:					
	precision	recall	f1-score	support	
0	1.00	0.98	0.99	280	
1	0.98	1.00	0.99	280	
accuracy			0.99	560	
macro avg	0.99	0.99	0.99	560	
weighted avg	0.99	0.99	0.99	560	

Gambar 2. Klasifikasi Model Random Forest

Nilai precision sebesar 1.00 pada Kelas 0 menandakan bahwa seluruh mahasiswa yang diprediksi tidak lulus memang benar-benar berstatus tidak lulus, sehingga tidak ditemukan false positive pada kelas minoritas awal.

4.1.2 Analisa Confusion Matrix

Distribusi kesalahan klasifikasi antar kelas secara detail dapat dilihat melalui visualisasi matriks kebingungan.

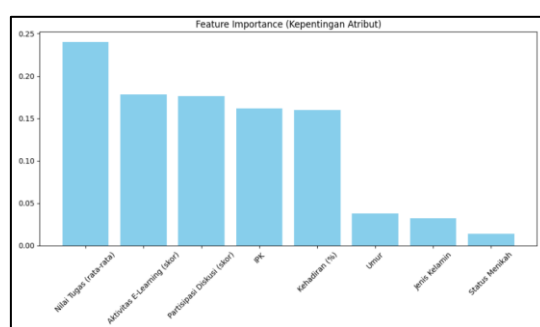


Gambar 3. Confusion Matrix Prediksi Status Akademik

Dari total 280 data aktual mahasiswa "Tidak Lulus", sebanyak 274 berhasil diprediksi secara benar dan hanya 6 sampel yang meleset. Sementara itu, dari total 280 data aktual mahasiswa "Lulus", sebanyak 279 berhasil diprediksi benar dengan hanya 1 kesalahan prediksi. Kesalahan minor ini menunjukkan adanya kemiripan fitur pada beberapa sampel mahasiswa, namun secara keseluruhan model memiliki kemampuan pemisahan kelas yang tajam.

4.1.3 Analisa Pentingnya Atribut (*Feature Importance*)

Melalui fitur feature importance pada Random Forest, tingkat pengaruh masing-masing atribut dalam menentukan status kelulusan mahasiswa dapat dipetakan secara jelas.



Gambar 4. Feature Importance Kepentingan Atribut

Variabel Nilai Tugas menempati posisi tertinggi sebagai prediktor utama kesuksesan akademik. Aktivitas E-Learning memiliki pengaruh signifikan kedua, disusul oleh Partisipasi Diskusi di peringkat ketiga. Indikator akumulatif tradisional seperti IPK dan Kehadiran berada di peringkat bawah, yang membuktikan bahwa karakteristik perilaku belajar harian jauh lebih menentukan status kelulusan mahasiswa.

4.2 Pembahasan

Pencapaian akurasi sebesar 98,75% membuktikan bahwa sinergi antara algoritma Random Forest dan SMOTE mampu menjawab tantangan ketidakseimbangan data pada domain Educational Data Mining. Tanpa SMOTE, model cenderung mengabaikan kelas minoritas, namun dengan teknik ini, model berhasil belajar secara adil dari kedua kelas. Dominasi atribut aktivitas harian menunjukkan bahwa intervensi dini dapat dilakukan oleh institusi pendidikan hanya dengan melihat keaktifan harian mahasiswa, tanpa harus menunggu nilai IPK keluar di akhir semester. Hal ini memungkinkan diterapkannya sistem peringatan dini (Early Warning System) yang lebih responsif dan akurat untuk memitigasi angka keterlambatan kelulusan.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan dari penelitian ini menunjukkan bahwa penerapan teknik SMOTE terbukti secara signifikan mampu mengatasi permasalahan imbalanced data pada data akademik mahasiswa. Model kombinasi Random Forest dan SMOTE ini menghasilkan performa yang sangat memuaskan dengan capaian akurasi sebesar 98,75% serta nilai precision, recall, dan f1-score yang konsisten berada pada rentang 0.99. Temuan ilmiah dari analisis feature importance menegaskan bahwa variabel Nilai Tugas dan Aktivitas E-Learning merupakan fitur yang paling berpengaruh dalam proses klasifikasi kelulusan.

Hal ini membuktikan secara empiris bahwa perilaku akademik harian mahasiswa memiliki korelasi yang jauh lebih kuat terhadap prediktabilitas kelulusan dibandingkan dengan indikator akumulatif tradisional.

5.2 Saran

Saran yang dapat diberikan berdasarkan hasil penelitian ini adalah agar institusi pendidikan mengadopsi model prediksi berbasis Random Forest ini ke dalam sistem manajemen akademik sebagai instrumen Early Warning System (EWS) guna mendeteksi mahasiswa yang memiliki risiko kendala akademik secara dini. Untuk pengembangan penelitian selanjutnya, disarankan melakukan komparasi performa dan efisiensi waktu komputasi menggunakan algoritma ensemble learning lainnya pada dataset yang lebih masif guna memvalidasi ketangguhan teknik SMOTE dalam berbagai skenario data. Selain itu, integrasi variabel non-akademik perlu dipertimbangkan guna memperkaya fitur input, sehingga model dapat memberikan hasil prediksi kelulusan yang lebih holistik dari berbagai dimensi kehidupan mahasiswa..

DAFTAR REFERENSI

- Alnassar, F. M., Blackwell, T., Homayounvala, E., & Yee-king, M. (2026). Addressing Class Imbalance in Predicting Student Performance Using SMOTE and GAN Techniques. *Applied Sciences (Switzerland)*, 16(7), 1–22. <https://doi.org/10.3390/app16073274>
- Andrianof, H., Gusman, A. P., & Universitas, O. A. P. (2025). Implementasi Algoritma Random Forest untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus di Perguruan Tinggi Indonesia. *Jurnal Sains Informatika Terapan (JSIT)*, 4(1), 187–201.
- Anggrawan, A., Hairani, H., & Satria, C. (2023). Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE. *International Journal of Information and Education Technology*, 13(2), 289–295. <https://doi.org/10.18178/ijiet.2023.13.2.1806>
- Anjani, A. F., Anggraeni, D., & Tirta, I. M. (2023). Implementasi Random Forest Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi Sister for Students UNEJ. *Jurnal Nasional Teknologi Dan Sistem Informasi*, 9(2), 163–172. <https://doi.org/10.25077/teknosi.v9i2.2023.163-172>
- Anju Fauziah, & Julan Hernadi. (2025). Klasifikasi Data Tak Seimbang Menggunakan Algoritma Random Forest dengan SMOTE dan SMOTE-ENN. *Teknomatika: Jurnal Informatika Dan Komputer*, 17(2), 38–47. <https://doi.org/10.30989/teknomatika.v17i2.1530>
- Diantika, S., Nalatissifa, H., Maulidah, N., Supriyadi, R., & Fauzi, A. (2024). Penerapan Teknik Random Oversampling Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma Random Forest. *Computer Science (CO-SCIENCE)*, 4(1),

11–18. <https://doi.org/10.31294/coscience.v4i1.1996>

- Hairani. (2021). PENINGKATAN KINERJA METODE SVM MENGGUNAKAN METODE KNN IMPUTASI DAN K-MEANS-SMOTE UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS BUMIGORA. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIK)*, 8(4), 713–718. <https://doi.org/10.25126/jtik.202183428>
- Hamdani, Susanti, Anam, M. K., Pradipta, R., & Lathifah. (2025). Optimization of Machine Learning Models in Student Graduation Prediction Systems Using Ensemble Learning with PSO and SMOTE. *Sinkron*, 9(4), 3149–3158. <https://doi.org/10.33395/sinkron.v9i4.15335>
- Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing. *Sustainability (Switzerland)*, 14(22). <https://doi.org/10.3390/su142214795>
- Kusuma, P. A., & Firmansyah, A. U. (2022). Deteksi Penyebaran Penyakit Tuberkulosis dengan Algoritma K-Means Clustering Menggunakan Rapid Miner Abstrak Penyebaran Penyakit TBC di Riau Menggunakan Rapid Miner ". Dimana menurutnya mining memakai Metode Algoritma K-Means Clustering terbukti efektif u. *Jurnal Teknologi Informatika Dan Komputer MH. Thamrin*, 8(2), 41–54.
- Mukherjee, M., & Khushi, M. (2021). Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 4(1), 1–15. <https://doi.org/10.3390/asi4010018>
- Ngaeni, N. S. (2025). Comparative Analysis of Random Forest and Support Vector Classifier for Predicting Students' on-Time Graduation. *Jurnal Pilar Nusa Mandiri*, 21(2), 247–255. <https://doi.org/10.33480/pilar.v21i2.7048>
- Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, 4(1), 61–72. <https://doi.org/10.54082/jiki.168>
- Pratiwi, A. N., & Utami, E. (2025). Prediksi Kinerja Akademik Matematika Siswa berdasarkan Kepribadian Big Five menggunakan Random Forest dengan Teknik Synthetic Minority Over - Sampling Personality Traits using Random Forest with Synthetic Minority Over - Sampling. *Sistemasi: Jurnal Sistem Informasi*, 14(2), 985–1000. <https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/5102/920>
- Rekha, M. S., Kumar, A., & Swamy, K. (2023). A Method of Classification in Student Grade Prediction using SMOTE Oversampling Technique. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal*, 3(2), 130–136. <https://doi.org/10.48175/IJARSCT-12719>
- Rohman, M. G., Abdullah, Z., Kasim, S., & Rasyidah. (2025). Hybrid Logistic Regression Random Forest on Predicting Student Performance. *International Journal on Informatics Visualization*, 9(2), 859–866.

<https://doi.org/10.62527/joiv.9.2.3487>

- Sidik, A. D., & Ansawarman, A. (2022). Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning. *Formosa Journal of Multidisciplinary Research*, 1(3), 559–568. <https://doi.org/10.55927/fjmr.v1i3.745>
- Siswara, D., M. Soleh, A., & Hamim Wigena, A. (2024). Classification Modeling with RNN-based, Random Forest, and XGBoost for Imbalanced Data: A Case of Early Crash Detection in ASEAN-5 Stock Markets. *Scientific Journal of Informatics*, 11(3), 569–582. <https://doi.org/10.15294/sji.v11i3.4067>
- Tedut, E. M., Wahyuni, W., & Adytia, P. (2025). *Early Warning System Kelulusan of Student Graduation Using the Random Forest Algorithm*. 29(1), 1–10. <https://doi.org/10.46984/sebatik.v29i1.0000>
- Wijayaningrum, V. N., Kirana, A. P., & Putri, I. K. (2024). Student Academic Performance Prediction Framework With Feature Selection and Imbalanced Data Handling. *Jurnal Ilmiah Kursor*, 12(3), 123–134. <https://doi.org/10.21107/kursor.v12i3.356>