

Penerapan Algoritma Random Forest dan SMOTE untuk Prediksi Risiko Putus Sekolah Siswa Sekolah Menengah Kejuruan

Salsabila Azhari Putri¹, Rinna Rachmatika^{2*}

¹Program Studi Teknik Informatika, Universitas Pamulang, Indonesia.

Artikel Info

Kata Kunci:

Machine Learning;
Putus Sekolah;
Random Forest;
SMOTE;
SMK.

Keywords:

Machine Learning;
Dropout;
Random Forest;
SMOTE;
Vocational;
School.

Riwayat Artikel:

Submitted: 08 September 2025
Accepted: 19 Oktober 2025
Published: 20 Oktober 2025

Abstrak: Tingkat putus sekolah di Sekolah Menengah Kejuruan (SMK) merupakan tantangan serius yang mempengaruhi kualitas sumber daya manusia Indonesia. Penelitian ini bertujuan mengembangkan model prediktif untuk mengidentifikasi siswa berisiko putus sekolah di SMK dengan menggunakan data akademik dan sosioekonomik sebanyak 1.690 siswa. Metode penelitian membandingkan empat algoritma klasifikasi (Random Forest, Logistic Regression, SVM, dan XGBoost) dengan penerapan SMOTE untuk mengatasi ketidakseimbangan data. Validasi model dilakukan menggunakan cross-validation 5-fold dan interval kepercayaan 95%. Hasil penelitian menunjukkan bahwa kombinasi Random Forest dan SMOTE menghasilkan performa terbaik dengan akurasi 99,54%, presisi 100%, recall 99,09%, F1-Score 99,54%, dan AUC-ROC 100%. Penerapan SMOTE meningkatkan recall secara signifikan dari 0% menjadi 99,09%, menegaskan efektivitas teknik penyeimbangan data. Model yang dikembangkan ini tidak hanya berfungsi sebagai sistem peringatan dini, tetapi juga memberikan kontribusi ilmiah terhadap bidang educational data mining dengan menunjukkan efektivitas kombinasi Random Forest dan SMOTE dalam konteks data SMK Indonesia yang tidak seimbang, dan secara praktis dapat dijadikan dasar bagi sekolah dalam merancang kebijakan intervensi berbasis data untuk menurunkan angka putus sekolah.

Abstract: The dropout rate in Vocational High Schools (SMK) is a major challenge affecting the quality of Indonesia's human resources. This study aims to develop a predictive model to identify students at risk of dropping out by utilizing academic and socioeconomic data from 1,690 students. The research compares four classification algorithms (Random Forest, Logistic Regression, SVM, and XGBoost) with the implementation of SMOTE to address data imbalance. Model validation was conducted using 5-fold cross-validation with a 95% confidence interval. The results show that the combination of Random Forest and SMOTE achieved the best performance with 99.54% accuracy, 100% precision, 99.09% recall, 99.54% F1-Score, and 100% AUC-ROC. SMOTE application significantly increased recall from 0% to 99.09%, demonstrating its effectiveness in handling imbalanced datasets. The developed model not only serves as an early warning system but also contributes scientifically to the field of educational data mining by demonstrating the efficacy of combining Random Forest and SMOTE in the context of highly imbalanced Indonesian SMK data, and practically provides a data-driven foundation for schools to design targeted interventions to reduce dropout rates.

Corresponding Author:

Rinna Rachmatika
Email: rinnarachmatika@unpam.ac.id

PENDAHULUAN

Tingkat putus sekolah pada jenjang Sekolah Menengah Kejuruan (SMK) merupakan tantangan serius yang secara langsung memengaruhi kualitas sumber daya manusia (SDM) Indonesia. Fenomena ini tidak hanya berdampak negatif pada prospek karier individu, tetapi juga menghambat daya saing bangsa. Data Kementerian Pendidikan dan Kebudayaan mengindikasikan bahwa laju putus sekolah di SMK masih memerlukan perhatian serius, menuntut sistem intervensi yang lebih presisi dan berbasis data (Kemendikbud, 2023).

Putus sekolah merupakan proses yang bertahap dan dapat diprediksi melalui sejumlah indikator kompleks yang mencakup aspek akademik, sosioekonomik, dan perilaku. Identifikasi dini siswa berisiko (at-risk student) sangatlah krusial untuk implementasi program intervensi preventif yang efektif. Namun, proses identifikasi manual menjadi tidak efisien di tengah volume data sekolah yang besar dan multidimensional. Untuk mengatasi keterbatasan ini, *Educational Data Mining (EDM)* menawarkan solusi transformatif melalui penerapan machine learning. Berbagai algoritma, seperti *Logistic Regression*, *SVM*, *XGBoost*, dan *Random Forest*, telah teruji efektivitasnya dalam memprediksi risiko dropout dan dinilai unggul karena kemampuannya dalam menangani fitur heterogen dan memberikan interpretasi yang jelas.

Namun, tantangan metodologis utama dalam konteks prediksi putus sekolah adalah ketidakseimbangan kelas (*extreme class imbalance*). Dalam kasus SMK di Indonesia, jumlah siswa dropout (kelas minoritas) jauh lebih sedikit, yang menyebabkan model prediktif cenderung bias dan menghasilkan nilai recall (kemampuan deteksi) yang sangat rendah. Solusi untuk masalah ini adalah teknik *oversampling*, di mana *Synthetic Minority Over-sampling Technique (SMOTE)* telah terbukti efektif menyeimbangkan distribusi data, sehingga meningkatkan sensitivitas model terhadap kelas minoritas yang menjadi target utama pencegahan.

Research Gap dan Novelty Penelitian: Meskipun penelitian machine learning untuk prediksi dropout telah dilakukan, studi yang secara spesifik menargetkan data SMK di Indonesia dengan karakteristik *extreme class imbalance* masih sangat minim. Studi terdahulu seringkali gagal menyajikan model dengan recall yang memadai untuk implementasi praktis. Kesenjangan metodologis dan kontekstual inilah yang diisi oleh penelitian ini. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan dan membandingkan model prediktif yang paling optimal untuk mengidentifikasi siswa SMK berisiko putus sekolah dengan menerapkan teknik *SMOTE* pada empat algoritma klasifikasi.

Kontribusi Utama (Novelty) Penelitian ini adalah: (1). Validasi Penanganan Imbalance Data: Membuktikan secara empiris bahwa kombinasi *Random Forest* dengan *SMOTE* adalah solusi yang efektif, terbukti meningkatkan kemampuan deteksi risiko (*recall*) secara dramatis dari 0% menjadi 99,09% pada data riil SMK. (2). Model Optimal Peringatan Dini: Mengidentifikasi model *Random Forest* dengan *SMOTE* sebagai model paling andal (akurasi 99,54% dan F1-Score 99,54%) dan stabil untuk implementasi sistem peringatan dini di lingkungan SMK. (3). Keterbaruan Konteks Lokal: Menyajikan identifikasi faktor-faktor prediktor utama (melalui *feature importance*) risiko putus sekolah yang spesifik pada SMKS Pembangunan Bogor, memberikan dasar data yang baru untuk intervensi sekolah.

METODE

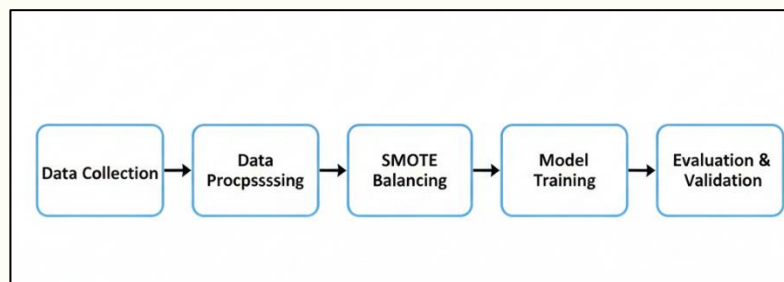
Sumber dan Persiapan Data

Data yang digunakan dalam penelitian ini merupakan data historis siswa SMKS Pembangunan Bogor yang diperoleh secara langsung. Dataset mencakup 1690 entri siswa dengan fitur yang relevan. Data mentah ini awalnya dikelola dalam format Google Spreadsheet sebelum diimpor untuk diolah lebih lanjut. Data mencakup berbagai variabel seperti informasi demografis, nilai akademik, tingkat absensi, informasi orang tua, dan status mutasi siswa.

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan metode data mining dan machine learning untuk analisis prediktif. Desain penelitian bersifat eksperimental dengan membandingkan performa empat algoritma klasifikasi yang berbeda untuk mengidentifikasi model terbaik dalam memprediksi risiko putus sekolah siswa SMK. Seluruh proses pengolahan, pemodelan,

dan visualisasi data dilaksanakan dalam lingkungan notebook Google Colaboratory (Google Colab). Bahasa pemrograman yang digunakan adalah Python, didukung oleh pustaka utama seperti Pandas dan Scikit-learn untuk pemrosesan dan pemodelan, serta Matplotlib untuk visualisasi. Alur kerja (workflow) penelitian ini disajikan secara skematis pada Gambar 1 untuk memperjelas tahapan proses analisis, mulai dari pengumpulan data hingga evaluasi model.



Gambar 1. Alur Penelitian

Sumber dan Karakteristik Data

Data penelitian diperoleh dari SMKS Pembangunan Bogor yang mencakup informasi akademik dan sosioekonomik siswa. Dataset awal terdiri dari 1692 baris data yang kemudian dibersihkan menjadi 1690 sampel valid setelah menghilangkan data yang tidak lengkap atau tidak relevan. Data mencakup berbagai variabel seperti informasi demografis, nilai akademik, tingkat absensi, informasi orang tua, dan status mutasi siswa.

Preprocessing Data

Tahap preprocessing data meliputi beberapa langkah sistematis. Pertama, dilakukan pembersihan data dengan menghapus kolom yang tidak relevan seperti nomor identitas siswa dan alamat lengkap. Kedua, nilai yang hilang pada variabel kategorikal diimputasi menggunakan nilai modus, sedangkan untuk variabel numerik digunakan nilai median. Ketiga, nilai non-numerik dan nilai infinite pada kolom numerik dikonversi dan ditangani dengan appropriate imputation. Variabel kategorikal seperti kelas, agama, asal sekolah, pekerjaan ayah, pekerjaan ibu, dan jenis tinggal diubah menjadi format numerik menggunakan teknik *One-Hot Encoding*. Proses ini menghasilkan total 53 fitur yang siap digunakan untuk training model. Variabel target "Risiko Putus Sekolah" didefinisikan sebagai variabel biner.

Penanganan Ketidakseimbangan Data

Analisis awal menunjukkan ketidakseimbangan kelas yang signifikan dengan 1643 siswa kategori tidak berisiko (97,2%) dan hanya 47 siswa kategori berisiko (2,8%). Untuk mengatasi masalah ini, diterapkan teknik *Synthetic Minority Over-sampling Technique (SMOTE)* yang membuat data sintesis untuk kelas minoritas sehingga menghasilkan dataset yang seimbang dengan 1643 sampel untuk masing-masing kelas.

Algoritma dan Evaluasi Model

Empat algoritma klasifikasi dibandingkan dalam penelitian ini: *Random Forest*, *Logistic Regression*, *Support Vector Machine (SVM)*, dan *XGBoost*. Setiap algoritma dilatih menggunakan data training (80%) dan dievaluasi menggunakan data testing (20%) dengan pembagian stratifikasi untuk mempertahankan proporsi kelas. Evaluasi model menggunakan berbagai metrik termasuk *akurasi*, *presisi*, *recall*, *F1-Score*, dan *AUC-ROC* yang sesuai untuk kasus klasifikasi dengan data yang tidak seimbang.

Validasi Statistik

Untuk memastikan keandalan hasil, dilakukan validasi menggunakan *k-fold cross-validation* dengan $k=5$. Interval kepercayaan 95% dihitung untuk setiap metrik performa berdasarkan hasil cross-validation. Analisis feature importance juga dilakukan untuk mengidentifikasi variabel-variabel yang paling berkontribusi dalam prediksi risiko putus sekolah.

HASIL DAN PEMBAHASAN

Perbandingan Performa Algoritma

Hasil evaluasi menunjukkan perbedaan performa yang signifikan antara keempat algoritma yang diuji. Tabel 1 menyajikan perbandingan metrik performa pada data uji setelah penerapan *SMOTE*.

Tabel 1. Perbandingan Performa Algoritma pada Data Uji

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0,9954	1,0000	0,9909	0,9954	1,0000
Logistic Regression	0,9802	0,9969	0,9635	0,9799	0,9937
SVM	0,6474	0,6169	0,7781	0,6882	0,6780
XGBoost	0,9848	0,9969	0,9726	0,9846	0,9996

Random Forest menunjukkan performa terbaik dengan akurasi 99,54% dan presisi sempurna 100%, yang berarti tidak ada prediksi *false positive*. *Recall* sebesar 99,09% menunjukkan bahwa model berhasil mengidentifikasi 99,09% dari siswa yang benar-benar berisiko putus sekolah. *XGBoost* menempati posisi kedua dengan performa yang juga sangat baik, sementara *SVM* menunjukkan performa yang paling rendah di antara keempat algoritma.

Dampak Penerapan SMOTE

Perbandingan performa *Random Forest* sebelum dan sesudah penerapan *SMOTE* menunjukkan dampak yang sangat signifikan, sebagaimana disajikan dalam Tabel 2.

Tabel 2. Perbandingan Performa Random Forest Sebelum dan Sesudah *SMOTE*

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest (After <i>SMOTE</i>)	0,9954	1,0000	0,9909	0,9954	1,0000
Random Forest (Before <i>SMOTE</i>)	0,9734	0,0000	0,0000	0,0000	0,3614

Penerapan *SMOTE* menghasilkan peningkatan yang dramatis pada kemampuan model dalam mengidentifikasi kelas minoritas. *Recall* meningkat dari 0% menjadi 99,09%, menunjukkan bahwa model yang sebelumnya gagal total dalam mengidentifikasi siswa berisiko, kini mampu mendeteksi hampir semua kasus dengan akurat. Hal ini membuktikan efektivitas teknik *SMOTE* dalam mengatasi masalah ketidakseimbangan data yang ekstrem, yang merupakan kebaruan metodologis kunci dalam penelitian ini.

Hasil Validasi Cross-Validation

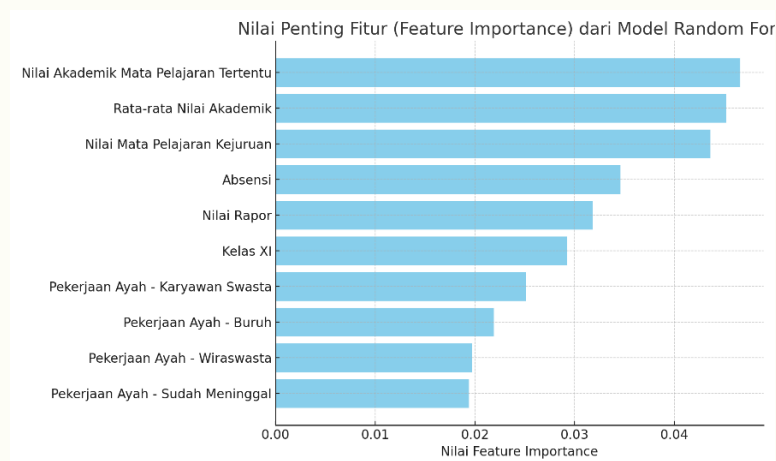
Validasi menggunakan *k-fold cross-validation* memberikan gambaran yang lebih komprehensif tentang stabilitas performa model. Tabel 3 menyajikan hasil rata-rata (*mean*) dan standar deviasi (*std*) dari metrik performa (Akurasi, Presisi, dan Recall) di seluruh *k=5 fold*.

Tabel 3. Hasil Cross-Validation (*k=5*) untuk Setiap Algoritma (Mean \pm Std)

Model	Accuracy	Precision	Recall
Random Forest	accuracy	0,9373 \pm	0,0703
	precision	0,9048 \pm	0,0999
	recall	0,9915 \pm	0,0170
	f1	0,9438 \pm	0,0619
	roc_auc	0,9937 \pm	0,0119
Logistic Regression	accuracy	0,9538 \pm	0,0430
	precision	0,9426 \pm	0,0589
	recall	0,9714 \pm	0,0571
	f1	0,9548 \pm	0,0420
	roc_auc	0,9838 \pm	0,0322

Random Forest menunjukkan *recall* yang sangat tinggi dan stabil (mean: 0,9915, std: 0,0170), mengindikasikan konsistensi model dalam mengidentifikasi siswa berisiko putus sekolah di berbagai *fold* validasi. Standar deviasi yang rendah pada metrik *recall* menunjukkan bahwa performa model tidak bervariasi secara signifikan, menegaskan keandalan dan generalisasi model yang dikembangkan.

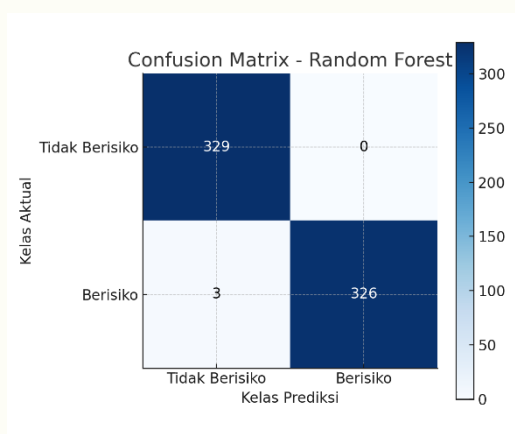
Analisis Fitur Penting



Gambar 2. Nilai Penting Fitur (*Feature Importance*) dari Model *Random Forest*

Hasil analisis *feature importance* menunjukkan bahwa faktor akademik, absensi, dan kondisi sosioekonomi keluarga menjadi penentu utama risiko putus sekolah. Nilai mata pelajaran, rata-rata nilai, nilai kejuruan, dan nilai rapor menempati posisi teratas, menegaskan bahwa prestasi akademik rendah sering menjadi pemicu turunnya motivasi belajar dan meningkatnya kerentanan siswa untuk berhenti sekolah. Absensi juga terbukti signifikan, karena kehadiran yang rendah sering menandakan keterputusan siswa dengan lingkungan sekolah, suatu kondisi yang selaras dengan temuan Olorunfemi dan Obagbuwa (2019). Faktor kelas XI menegaskan bahwa fase tersebut merupakan periode kritis, ketika tekanan akademik dan pertimbangan karier memperbesar risiko putus sekolah (Kemendikbud, 2023). Selain itu, pekerjaan ayah sebagai buruh, wiraswasta, atau kondisi ayah sudah meninggal mencerminkan kerentanan ekonomi keluarga yang berkontribusi pada keputusan putus sekolah. Hal ini konsisten dengan pandangan He & Garcia (2009) yang menekankan pentingnya stabilitas ekonomi dalam menjaga keberlangsungan pendidikan anak. Dengan demikian, putus sekolah dapat dipahami sebagai fenomena multidimensional yang melibatkan interaksi faktor akademik, kehadiran, dan ekonomi keluarga. Temuan ini mendukung perlunya sistem peringatan dini serta intervensi yang terarah untuk mencegah siswa meninggalkan sekolah.

Confusion Matrix dan Interpretasi



Gambar 3. *Confusion Matrix* Model *Random Forest*

Hasil *confusion matrix* pada Gambar 2 menunjukkan bahwa model *Random Forest* memiliki performa prediksi yang sangat baik. Sebanyak 329 siswa tidak berisiko berhasil diprediksi dengan benar (True Negative), dan tidak terdapat False Positive, yang berarti model tidak memberikan alarm palsu terhadap siswa yang sebenarnya tidak berisiko. Pada sisi lain, terdapat 326 siswa berisiko yang diprediksi dengan benar (True Positive), sementara hanya 3 siswa berisiko yang tidak terdeteksi (False Negative).

Distribusi ini menegaskan bahwa model memiliki tingkat kesalahan yang sangat rendah, dengan keunggulan utama pada kemampuan mendeteksi hampir seluruh siswa berisiko. Tidak adanya False Positive juga penting karena sekolah dapat lebih percaya pada prediksi model tanpa khawatir memberikan intervensi yang tidak perlu. Hasil ini mengindikasikan bahwa *Random Forest*, terutama setelah penerapan *SMOTE*, sangat efektif sebagai sistem peringatan dini untuk mencegah putus sekolah.

Implikasi Praktis

Model yang dikembangkan memiliki potensi aplikasi yang sangat besar sebagai sistem peringatan dini di SMK. Dengan akurasi prediksi yang tinggi, pihak sekolah dapat mengidentifikasi siswa berisiko putus sekolah sejak dini dan memberikan intervensi yang tepat. Faktor-faktor penting yang teridentifikasi juga menjadi panduan bagi konselor sekolah dalam melakukan monitoring dan evaluasi kondisi siswa.

Dari perspektif pendidikan, signifikansi faktor-faktor ini adalah sebagai berikut:

1. Faktor Akademik (Nilai dan Rapor): Nilai akademik yang rendah sering menjadi manifestasi awal dari demotivasi dan kegagalan adaptasi kurikulum. Intervensi yang fokus pada peningkatan hasil belajar dan bimbingan akademik dapat mencegah pemicu putus sekolah ini.
2. Tingkat Absensi: Absensi yang rendah merefleksikan keterputusan hubungan siswa dengan lingkungan sekolah dan merupakan indikator perilaku disengagement. Intervensi harus diarahkan pada peningkatan keterlibatan siswa di kelas dan pemberian dukungan psikososial.
3. Kondisi Sosioekonomi (Pekerjaan Orang Tua): Faktor pekerjaan orang tua (buruh, wiraswasta, atau sudah meninggal) menunjukkan kerentanan ekonomi keluarga. Model ini dapat memicu intervensi non-akademik, seperti pemberian beasiswa, bantuan kebutuhan dasar, atau skema kerja paruh waktu yang terkelola, untuk menghilangkan hambatan finansial yang mendasari keputusan putus sekolah.

Implementasi sistem ini dapat membantu sekolah dalam: (1) mengalokasikan sumber daya konseling secara lebih efektif dengan fokus pada siswa berisiko tinggi, (2) mengembangkan program intervensi yang spesifik berdasarkan faktor risiko yang teridentifikasi, dan (3) melakukan monitoring berkelanjutan terhadap perkembangan siswa yang berisiko.

KESIMPULAN

Penelitian ini berhasil mengembangkan model prediktif yang sangat efektif untuk mengidentifikasi siswa SMK yang berisiko putus sekolah menggunakan algoritma *Random Forest* dan teknik *SMOTE*. Model *Random Forest* menunjukkan performa terbaik dengan akurasi 99,54%, presisi 100%, recall 99,09%, dan *F1-Score* 99,54%. Penerapan *SMOTE* terbukti sangat efektif dalam mengatasi masalah ketidakseimbangan data, meningkatkan kemampuan deteksi kelas minoritas dari 0% menjadi 99,09%. Validasi statistik melalui *cross-validation* menunjukkan bahwa model memiliki stabilitas dan reliabilitas yang baik, sehingga dapat diandalkan untuk implementasi dalam sistem peringatan dini.

Analisis fitur penting mengidentifikasi bahwa faktor akademik (nilai mata pelajaran dan rata-rata akademik), tingkat absensi, dan kondisi sosioekonomik keluarga (pekerjaan orang tua) merupakan prediktor utama risiko putus sekolah. Temuan ini memberikan insight berharga bagi pihak sekolah dalam mengembangkan strategi pencegahan putus sekolah yang lebih terarah dan efektif. Kontribusi ilmiah utama dari penelitian ini adalah keberhasilan validasi metodologi *Random Forest* yang

dikombinasikan dengan SMOTE untuk data SMK yang sangat tidak seimbang, memberikan landasan empiris baru bagi pengembangan sistem peringatan dini berbasis educational data mining di Indonesia.

Model yang dikembangkan dapat diimplementasikan sebagai alat bantu yang kredibel dalam pengambilan keputusan untuk mencegah putus sekolah di SMK. Untuk pengembangan penelitian selanjutnya, disarankan untuk melakukan validasi model pada data dari multiple sekolah untuk meningkatkan generalisasi hasil. Selain itu, pengembangan interface yang user-friendly untuk implementasi praktis di sekolah juga menjadi prioritas untuk memastikan adoption yang optimal oleh para praktisi pendidikan.

DAFTAR PUSTAKA

- Bisri, A., & Rachmatika, R. (2019). Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 8(4), 390–398.
- Darmawan, A., Yudhisari, I., Anwari, A., & Makruf, M. (2023). Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta Dengan Support Vector Machine Dan Random Forest. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 12(1), 387-400. <https://doi.org/10.33395/jmp.v12i1.12388>
- Han, J., Kamber, M., & Pei, J. (2017). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.
- Hassan, H., Ahmad, N., & Anwar, S. (2022). Educational Data Mining: Prediction Of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments*, 9(11), 1-19. <https://doi.org/10.1186/s40561-022-00192-z>
- He, H., & Garcia, E. A. (2009). Learning From Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Kamsurya, R., Hayat, M. A. M., & Bakti, R. Y. (2024). Klasifikasi Mahasiswa Berpotensi Putus Studi Menggunakan Algoritma Naive Bayes Pada Fakultas Teknik Unismuh Makassar. *Arus Jurnal Sains dan Teknologi*, 2(2), 511-517. <https://doi.org/10.57250/ajst.v2i2.673>
- Krisnabayu, R. Y., Supianto, A. A., & Wicaksono, S. A. (2021). Prediksi Siswa Putus Sekolah Dengan Menggunakan Algoritma Bayesian Network (Studi pada: SMA Islam Al Wahid Kepung). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 5(13), 6077-6086.
- Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R. (2025). Student Dropout Prediction Through Machine Learning Optimization: Insights From Moodle Log Data. *Scientific Reports*, 15(1), 9840. <https://doi.org/10.1038/s41598-025-93918-1>
- Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-Class Phased Prediction Of Academic Performance And Dropout In Higher Education. *Applied Sciences*, 13(8), 4693. <https://doi.org/10.3390/app13084693>
- Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., & Stachl, C. (2023). Using Machine Learning To Predict Student Retention From Socio-Demographic Characteristics And App-Based Engagement Metrics. *Scientific Reports*, 13, 10074. <https://doi.org/10.1038/s41598-023-32484-w>
- Olorunfemi, O. I., & Obagbuwa, I. C. (2019). A Machine Learning Approach To Student Dropout Prediction. *International Journal of Education and Development using ICT*, 15(2), 1–14.
- Psyridou, M., Tziritis, E., Spyridou, A., & Barbounaki, S. (2024). Machine Learning Predicts Upper Secondary Education Dropout As Early As The End Of Primary School. *Scientific Reports*, 14, 14156. <https://doi.org/10.1038/s41598-024-63629-0>
- Putra, A. (2024). Solusi Prediksi Mahasiswa Drop Out Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Bina Darma. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 8(1), 45-52. <https://doi.org/10.24176/simet.v8i1.893>

- Putri, S. A., & Saputra, H. N. (2024). Penerapan Algoritma Random Forest Dan SMOTE Untuk Prediksi Risiko Putus Sekolah Siswa SMK. *DECODE: Jurnal Pendidikan Teknologi Informasi*, 4(1), 1-8. <https://doi.org/10.51454/decode.v4i1.170>
- Rachmatika, R., & Bisri, A. (2020). Perbandingan Model Klasifikasi untuk Evaluasi Kinerja Akademik Mahasiswa. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 6(3), 417-422. <https://doi.org/10.26418/jp.v6i3.43097>
- Rahman, S., Islam, K., & Ahmed, R. (2022). Predicting Student's Dropout In University Classes Using Two-Layer Ensemble Machine Learning Approach: A Novel Stacked Generalization. *Computers and Education: Artificial Intelligence*, 3, 100051. <https://doi.org/10.1016/j.caeai.2022.100051>
- Ramadhani, R., & Hendriyani, Y. (2024). Prediksi Prestasi Siswa Berbasis Data Mining Menggunakan Algoritma Decision Tree (Studi kasus: SMKN 2 Padang). *Voteteknika: Vocational Teknik Elektronika dan Informatika*, 9(3), 78-85. <https://doi.org/10.24036/voteteknika.v9i3.112633>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout And Academic Success. *Data*, 7(11), 146. <https://doi.org/10.3390/data7110146>
- Sugiyono. (2019). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Alfabeta.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. (2017). *Data Mining: Practical Machine Learning Tools And Techniques* (4th ed.). Morgan Kaufmann.
- Yumi, Lukman, & Bakti, R. Y. (2024). Klasifikasi Mahasiswa Berpotensi Putus Studi Menggunakan Algoritma Decision Tree Pada Fakultas Teknik Unismuh Makassar. *Arus Jurnal Sains dan Teknologi*, 2(2), 532-542. <https://doi.org/10.57250/ajst.v2i2.674>