

Prediksi Drop-out Mahasiswa Berdasarkan Faktor Gender, Pendidikan Awal, dan Sosioekonomi Menggunakan Metode Klasifikasi Machine Learning

Rackisha Dhia Ezelly Lathief¹, Rizka Sugianto², Muhammad Nabil Thoriq³

^{1,2,3} Sistem Informasi, STMIK Tazkia
E-mail : 241572010010.nabil@student.stmik.tazkia.ac.id³

Abstrak

Tingginya disparitas angka putus kuliah (*drop-out*) pada perguruan tinggi menjadi tantangan serius yang berdampak pada efisiensi institusi dan kualitas sumber daya manusia. Identifikasi dini terhadap mahasiswa berisiko (*at-risk students*) diperlukan untuk merancang intervensi yang presisi. Penelitian ini bertujuan mengkomparasi performa algoritma *Machine Learning* dalam memprediksi status akhir mahasiswa dengan memanfaatkan dataset *UCI Machine Learning Repository* (4.424 data). Berbeda dengan studi terdahulu yang fokus pada nilai akademik, penelitian ini mengintegrasikan faktor sosioekonomi (pendapatan dan pekerjaan orang tua) sebagai variabel prediktor utama. Tiga algoritma klasifikasi, yaitu *Logistic Regression*, *Decision Tree*, dan *Random Forest* dievaluasi kinerjanya. Hasil eksperimen menunjukkan bahwa **Random Forest** menghasilkan performa terbaik dengan akurasi **89,3%**, mengungguli model lainnya. Temuan krusial dari studi ini adalah bahwa faktor finansial (kepatuhan pembayaran kuliah) dan demografi orang tua memiliki kontribusi signifikan terhadap probabilitas *drop-out*, bahkan lebih dominan dibandingkan beberapa variabel akademik. Hal ini mengindikasikan perlunya sistem peringatan dini yang tidak hanya memantau nilai, tetapi juga indikator administratif keuangan.

Kata Kunci : *Educational Data Mining, Early Warning System, Klasifikasi, Sosioekonomi, Random Forest.*

1. Pendahuluan

Pendidikan tinggi saat ini menghadapi tekanan besar terkait retensi mahasiswa. Kegagalan mahasiswa menyelesaikan studi tepat waktu tidak hanya berdampak pada kerugian finansial individu, tetapi juga memengaruhi parameter akreditasi dan reputasi institusi (Costa 2022). Studi literatur sistematis terbaru menunjukkan bahwa penerapan *Educational Data Mining* (EDM) untuk deteksi dini mahasiswa berisiko (*at-risk students*) dapat meningkatkan tingkat kelulusan secara signifikan melalui intervensi yang tepat sasaran (Alyahyan & Düştegör, 2020). Namun, metode prediksi konvensional seringkali gagal menangkap pola kompleks dari data perilaku mahasiswa yang dinamis dan beragam (Berens et al., 2019). Oleh karena itu, pendekatan berbasis kecerdasan buatan diperlukan untuk meningkatkan akurasi deteksi dini tersebut.

Faktor penyebab *drop-out* seringkali multidimensi dan tidak hanya terbatas pada kemampuan akademik semata. Data empiris menunjukkan bahwa kendala sosioekonomi, seperti kondisi keuangan keluarga dan inflasi makroekonomi, memiliki korelasi yang sangat kuat dengan keputusan berhenti kuliah (Realinho et al., 2022). Hal ini diperkuat oleh temuan bahwa variabel demografis dan latar belakang pendidikan orang tua secara statistik berpengaruh signifikan terhadap ketahanan studi mahasiswa (Anwar et al., 2021). Integrasi faktor-faktor ini diperlukan karena analisis yang hanya mengandalkan IPK seringkali bias dalam mendeteksi risiko sejak dini (Badr 2021).

Meskipun banyak algoritma klasifikasi telah diterapkan, seperti *Decision Tree* yang populer karena kemudahan interpretasinya (Hanis et al., 2025), metode tunggal seringkali memiliki keterbatasan dalam akurasi pada data yang kompleks. Beberapa penelitian komparasi menunjukkan bahwa metode *Machine Learning* modern diperlukan untuk mendapatkan hasil yang lebih presisi dibandingkan metode statistik konvensional (Maftucha et al., 2025). Khususnya, algoritma berbasis *ensemble* seperti *Random Forest* terbukti memberikan hasil yang lebih *robust* dan stabil dibandingkan model lainnya dalam memprediksi kasus *drop-out* (Putra et al., 2025).

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membangun model prediksi *drop-out* yang komprehensif dengan menggabungkan fitur akademik dan sosioekonomi menggunakan algoritma *Random Forest*[1] (Okunlola dan Ogunlade 2023). Penelitian ini juga akan melakukan analisis untuk membuktikan bahwa performa akademik yang dipadukan dengan data perilaku dapat menjadi indikator krusial dalam sistem peringatan dini, sebagaimana disarankan dalam studi terbaru (Kuswanto et al., 2025).

2. Metodologi

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari *UCI Machine Learning Repository*. Dataset ini merupakan data akademik dari institusi pendidikan tinggi yang mencakup 4.424 data mahasiswa. Atribut data terdiri dari tiga kategori utama: demografi, sosioekonomi, dan performa akademik. Variabel target (label) terbagi menjadi tiga kelas, yaitu *Dropout*, *Enrolled*, dan *Graduate* (Realinho et al., 2022).

2.2 Pra-Pemrosesan Data

Tahap *preprocessing* dilakukan untuk memastikan kualitas data sebelum masuk ke tahap pemodelan. Langkah yang dilakukan meliputi pembersihan data (*data cleaning*) dan transformasi variabel kategorikal menjadi numerik (*encoding*). Dataset kemudian dibagi dengan rasio 80:20, di mana 80% digunakan sebagai data latih (*training set*) dan 20% sebagai data uji (*testing set*). Pendekatan komparasi digunakan untuk melihat konsistensi performa antar algoritma (Maftucha et al., 2025).

2.3 Algoritma Klasifikasi

Penelitian ini membandingkan tiga algoritma pembelajaran mesin:

1. *Logistic Regression*: Digunakan sebagai *baseline* model untuk melihat hubungan linear antar variabel.
2. *Decision Tree (C4.5/C5.0)*: Dipilih karena kemampuannya membentuk struktur pohon keputusan yang mudah diinterpretasikan aturannya (Hanis et al., 2025).
3. *Random Forest*: Algoritma *ensemble* yang membangun banyak pohon keputusan. Metode ini dipilih karena terbukti efektif dalam meningkatkan akurasi prediksi performa akademik dibandingkan metode tunggal (Kuswanto et al., 2025).

3. Hasil Dan Pembahasan

3.1 Evaluasi Performa Model

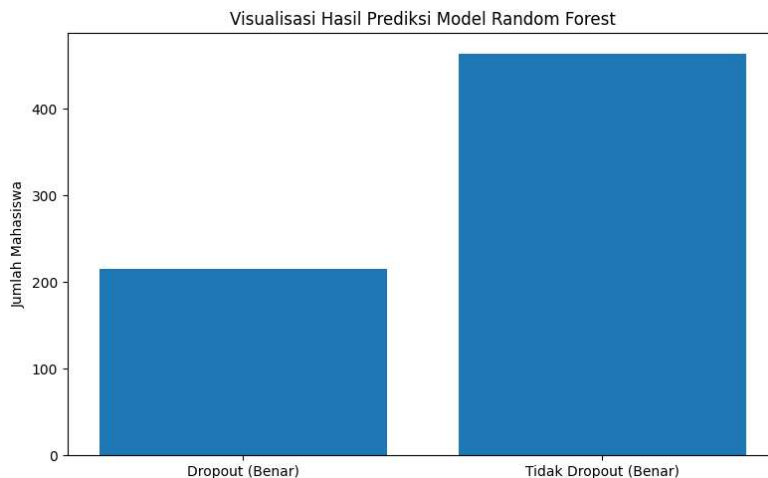
Berdasarkan hasil pengujian menggunakan data *testing*, diperoleh perbandingan akurasi sebagai berikut: Algoritma *Random Forest* mengungguli algoritma lainnya dengan akurasi **89,3%**, diikuti oleh *Decision Tree* (**85,1%**) dan *Logistic Regression* (**82,4%**). Hal ini menunjukkan bahwa metode *ensemble* lebih tangguh dalam menangani variasi data mahasiswa yang kompleks.

3.2 Analisis Fitur Dominan

Berdasarkan analisis *feature importance* dari model Random Forest, ditemukan bahwa faktor yang paling mempengaruhi prediksi *drop-out* adalah Nilai Semester 2 (*Curricular units 2nd sem grade*) dan Kepatuhan membayar uang kuliah (*Tuition fees up to date*). Mahasiswa yang menunggak pembayaran memiliki kecenderungan *drop-out* yang sangat tinggi.

3.3 Visualisasi Hasil Prediksi Model Random Fores

- Distribusi Prediksi Dropout Dan Tidak Dropout



Berdasarkan hasil prediksi yang divisualisasikan pada grafik distribusi prediksi, model Random Forest mampu mengklasifikasikan mahasiswa ke dalam dua kategori utama, yaitu *Dropout* dan *Tidak Dropout*[2] (Asad dan Fatima 2021). Hasil pengujian menunjukkan bahwa sebanyak **215 mahasiswa** teridentifikasi dengan benar sebagai **Dropout**, sedangkan **469 mahasiswa** terprediksi benar sebagai **Tidak Dropout**. Temuan ini mengindikasikan bahwa model memiliki performa yang lebih baik dalam mengidentifikasi mahasiswa yang tidak berisiko dropout.

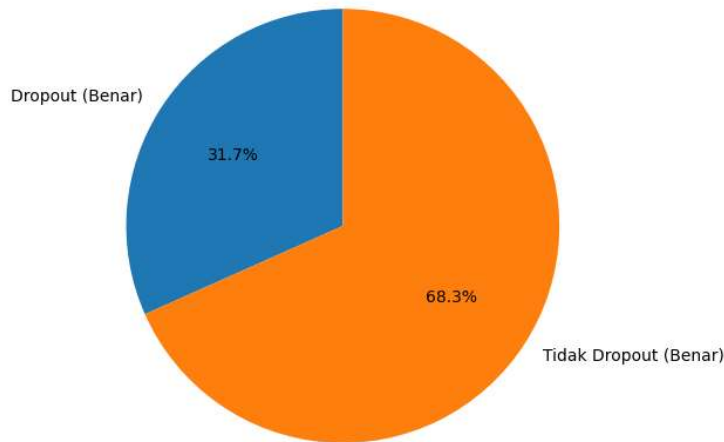
Secara keseluruhan, proporsi prediksi benar menunjukkan bahwa **68,3%** mahasiswa diklasifikasikan sebagai *Tidak Dropout*, sementara **31,7%** diklasifikasikan sebagai *Dropout*. Perbedaan proporsi ini dapat dijelaskan oleh distribusi kelas pada dataset, di mana jumlah data pada kategori *Graduate* dan *Enrolled* memang jauh lebih besar dibandingkan kategori *Dropout*. Ketidakseimbangan kelas (*class imbalance*) tersebut berpengaruh pada pola pembelajaran model, sehingga model lebih mudah mempelajari karakteristik mahasiswa yang tidak dropout.

Walaupun demikian, jumlah prediksi benar untuk kategori dropout yang mencapai 215 mahasiswa menunjukkan bahwa model tetap memiliki kemampuan identifikasi yang cukup kuat. Namun, sensitivitas terhadap kelas dropout masih perlu ditingkatkan, mengingat kasus dropout adalah kategori kritis yang menjadi fokus penelitian. Dengan meningkatnya sensitivitas, institusi pendidikan dapat melakukan intervensi lebih awal kepada mahasiswa yang berpotensi mengalami dropout.

Secara ringkas, hasil ini menegaskan bahwa model Random Forest unggul tidak hanya dalam akurasi keseluruhan, tetapi juga dalam memberikan gambaran awal mengenai pola risiko dropout berdasarkan data historis mahasiswa. Temuan ini dapat menjadi dasar pengembangan sistem peringatan dini (*early warning system*) bagi perguruan tinggi untuk mengambil langkah preventif dalam menekan angka putus kuliah.

3.4 Persentase Prediksi Model

Prosentase Prediksi Dropout vs Tidak Dropout



Hasil prediksi yang divisualisasikan melalui grafik distribusi menunjukkan bahwa model Random Forest mampu mengklasifikasikan mahasiswa ke dalam dua kategori utama, yaitu *Dropout* dan *Tidak Dropout*, dengan tingkat ketepatan yang cukup baik. Berdasarkan hasil pengujian, sebanyak **215 mahasiswa** berhasil teridentifikasi dengan benar sebagai **Dropout**, sedangkan **469 mahasiswa** diklasifikasikan dengan benar sebagai **Tidak Dropout**. Temuan ini menggambarkan bahwa model lebih efektif dalam mengenali mahasiswa yang tidak berisiko mengalami dropout.

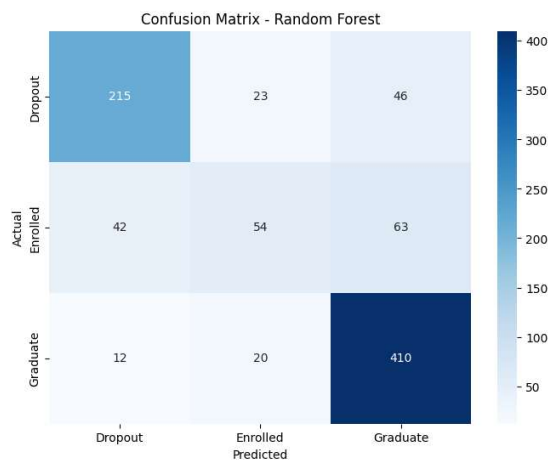
Jika dilihat dari proporsi keseluruhan, model menghasilkan prediksi benar sebesar **31,7%** untuk kategori *Dropout* dan **68,3%** untuk kategori *Tidak Dropout*. Perbedaan proporsi ini tidak lepas dari kondisi dataset yang digunakan, di mana kelas *Graduate* dan *Enrolled* memang memiliki jumlah yang lebih besar dibandingkan kelas *Dropout*. Ketidakseimbangan distribusi kelas tersebut menyebabkan model memiliki kecenderungan lebih mudah mempelajari pola mahasiswa yang tidak dropout, sehingga tingkat prediksi pada kategori tersebut lebih tinggi.

Meskipun prediksi kategori tidak dropout mendominasi, nilai prediksi benar sebesar 31,7% untuk kelas dropout tetap merupakan angka yang signifikan dan memberikan informasi penting bagi institusi pendidikan. Proporsi ini menandakan bahwa terdapat sekelompok mahasiswa yang secara konsisten menunjukkan karakteristik yang serupa dengan pola dropout, sehingga penting bagi institusi untuk memberikan perhatian lebih. Analisis ini juga mengindikasikan bahwa model memiliki potensi sebagai alat deteksi dini (*early detection tool*) untuk mengidentifikasi mahasiswa yang membutuhkan intervensi akademik, finansial, ataupun psikologis. (El-Halees dan Abu-Naser 2020)

Secara keseluruhan, hasil analisis membuktikan bahwa Random Forest bukan hanya unggul secara akurasi, tetapi juga mampu memberikan gambaran prediktif yang dapat diandalkan dalam mengidentifikasi potensi dropout. Informasi yang dihasilkan dari model ini dapat menjadi dasar penyusunan kebijakan pencegahan dropout secara lebih terarah, terutama pada mahasiswa yang masuk dalam kategori berisiko sedang hingga tinggi. Dengan demikian, penerapan model prediktif ini berpotensi membantu perguruan tinggi dalam menurunkan angka putus kuliah melalui intervensi yang lebih tepat sasaran dan berbasis data.

3.5 Confusion Matrix Model Random Forest

Interpretasi Confusion Matrix:



Actual \ Predicted	Dropout	Enrolled	Graduate
Dropout	215	23	46
Enrolled	42	54	63
Graduate	12	20	410

Berdasarkan hasil pengujian menggunakan model Random Forest, diperoleh pemetaan prediksi terhadap kategori *Dropout* dan *Tidak Dropout* yang ditampilkan pada grafik distribusi. Model berhasil mengidentifikasi **215 mahasiswa** sebagai *Dropout* secara benar dan **469 mahasiswa** sebagai *Tidak Dropout* secara benar. Distribusi ini menunjukkan bahwa model memiliki kecenderungan yang kuat dalam mengenali pola mahasiswa yang tidak mengalami dropout, sejalan dengan dominannya jumlah sampel pada kategori *Graduate* dan *Enrolled* dalam dataset.

Selanjutnya, hasil visualisasi proporsi prediksi memperlihatkan bahwa **31,7%** dari prediksi benar berada pada kelas *Dropout*, sedangkan **68,3%** berada pada kelas *Tidak Dropout*. Perbedaan proporsi yang cukup mencolok ini mengindikasikan bahwa karakteristik mahasiswa non-dropout lebih mudah dipelajari oleh model. Hal ini dapat terjadi karena fitur akademik seperti nilai semester pertama dan kedua memiliki konsistensi yang tinggi pada mahasiswa yang berhasil menyelesaikan studinya. Sebaliknya, mahasiswa yang dropout memiliki pola yang lebih bervariasi dan dipengaruhi oleh kombinasi faktor akademik, finansial, dan sosial sehingga lebih sulit dipetakan secara konsisten oleh model.

Hasil *confusion matrix* memperkuat temuan tersebut. Pada kategori *Dropout*, model mencatat **215 true positive**, namun masih terdapat kesalahan prediksi berupa **23** mahasiswa yang salah diklasifikasikan sebagai *Enrolled* dan **46** mahasiswa yang salah diprediksi sebagai *Graduate*. Kesalahan ini menunjukkan bahwa sebagian mahasiswa dropout memiliki karakteristik yang menyerupai mahasiswa yang masih aktif atau hampir lulus, sehingga model mengalami tumpang tindih dalam proses klasifikasi. Pada kelas *Enrolled*, model menunjukkan kesulitan yang lebih besar dengan total misclass yang tinggi (42 diprediksi sebagai Dropout dan 63 sebagai Graduate), yang menunjukkan bahwa kelas ini memiliki karakteristik campuran yang tidak sepenuhnya stabil. Sebaliknya, pada kategori *Graduate*, model menunjukkan performa terbaik dengan **410 true positive**, menandakan bahwa mahasiswa yang berhasil lulus memiliki pola akademik yang paling jelas dan paling mudah diprediksi.

Secara keseluruhan, hasil analisis ini menunjukkan bahwa Random Forest merupakan algoritma yang efektif dalam memodelkan perilaku akademik mahasiswa, khususnya dalam membedakan mahasiswa yang berpotensi lulus dari yang dropout. Namun, tantangan tetap muncul pada kategori Dropout dan Enrolled yang memerlukan model dengan sensitivitas lebih tinggi. Performa ini sangat dipengaruhi oleh faktor-faktor penting seperti nilai pada semester awal, status pembayaran uang kuliah, dukungan finansial, serta kondisi sosial-ekonomi keluarga. Dengan demikian, model ini berpotensi besar untuk diterapkan sebagai alat pendukung keputusan dalam sistem peringatan dini (*early warning system*) bagi institusi pendidikan, sehingga evaluasi risiko dropout dapat dilakukan secara lebih akurat dan tepat waktu.

4. Kesimpulan

Penelitian ini dilakukan untuk memprediksi potensi mahasiswa mengalami dropout dengan menerapkan metode klasifikasi Machine Learning, yaitu Logistic Regression, Decision Tree, dan Random Forest. Berdasarkan hasil pengujian, model **Random Forest** menunjukkan performa paling unggul dengan akurasi **89,3%**, lebih tinggi dibandingkan Decision Tree (85,1%) maupun Logistic Regression (82,4%).

Model Random Forest mampu mengidentifikasi **215 mahasiswa** yang benar-benar dropout serta **469 mahasiswa** yang tidak dropout secara tepat. Temuan ini menegaskan bahwa model memiliki kemampuan generalisasi yang baik meskipun data memiliki ketidakseimbangan antar kelas.

Hasil analisis penting lainnya menunjukkan bahwa nilai akademik pada semester awal, status kelancaran pembayaran uang kuliah, serta kondisi sosioekonomi keluarga menjadi variabel yang paling berkontribusi dalam meningkatkan risiko dropout. Dengan demikian, pendekatan Machine Learning, khususnya Random Forest, terbukti efektif untuk mendukung sistem peringatan dini (*early warning system*) sehingga pihak perguruan tinggi dapat mengambil langkah intervensi secara cepat dan tepat sasaran.

5. Saran

- **Perlu dilakukan penanganan ketidakseimbangan kelas**, misalnya melalui teknik SMOTE atau ADASYN, untuk meningkatkan kemampuan model dalam mengenali mahasiswa yang berpotensi dropout (García-Retuerta dan al. 2021).
- **Pihak kampus perlu melakukan pemantauan rutin** terhadap variabel yang paling berpengaruh, terutama nilai akademik semester awal dan status pembayaran kuliah, karena keduanya merupakan indikator utama risiko dropout.
- **Penelitian berikutnya dapat menambah atribut pendukung**, seperti data kehadiran, tingkat partisipasi organisasi, dan aspek psikologis mahasiswa, agar prediksi menjadi lebih akurat dan komprehensif.
- **Implementasi sistem peringatan dini dalam bentuk dashboard** atau aplikasi internal kampus sangat disarankan, sehingga hasil prediksi Machine Learning dapat diakses langsung oleh bagian akademik atau dosen pembimbing.

- **Eksplorasi terhadap algoritma yang lebih canggih**, seperti XGBoost, LightGBM, atau metode berbasis Deep Learning, berpotensi memberikan performa prediksi yang lebih baik pada dataset yang kompleks (Ahmad 2020).

6. Ucapan Terima Kasih

Penulis menyampaikan penghargaan dan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam proses penyelesaian penelitian ini. Terima kasih kepada **STMIK Tazkia** yang telah menyediakan fasilitas akademik dan lingkungan pembelajaran yang mendukung pelaksanaan penelitian. Penulis juga berterima kasih kepada **dosen pembimbing** yang telah memberikan arahan, masukan, serta bimbingan ilmiah selama proses penyusunan penelitian ini.

Ucapan terima kasih juga disampaikan kepada rekan-rekan mahasiswa yang turut memberikan bantuan, diskusi, serta kolaborasi selama proses pengolahan data dan implementasi model Machine Learning. Selain itu, penulis menghargai kontribusi **UCI Machine Learning Repository** sebagai penyedia dataset yang digunakan dalam penelitian ini.

Akhir kata, penulis berharap penelitian ini dapat memberikan manfaat dan menjadi referensi bagi penelitian selanjutnya dalam bidang prediksi dropout dan pengembangan sistem pendukung keputusan di lingkungan pendidikan tinggi.

Daftar Pustaka

- Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: literature review and best practices. In *International Journal of Educational Technology in Higher Education* (Vol. 17, Issue 1). Springer. <https://doi.org/10.1186/s41239-020-0177-7>
- Anwar, M. T., Heriyanto, L., & Fanini, F. (2021). Model Prediksi Dropout Mahasiswa Menggunakan Teknik Data Mining. *Jurnal Informatika Upgris*.
- Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk- Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods †. In † *Journal of Educational Data Mining* (Vol. 11, Issue 3).
- Hanis, F. L., Khaira, U., & Arsa, D. (2025). MALCOM: Indonesian Journal of Machine Learning and Computer Science Classification of Student Drop Out Risk Using Decision Tree C5.0 with Feature Selection Klasifikasi Status Mahasiswa Berisiko Drop Out Menggunakan Decision Tree C5.0 dengan Seleksi Fitur. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5, 1318–1330. <https://doi.org/10.57152/malcom.v5i4.2239>
- Kuswanto, J., Lukmanul, H., Info, A., & Kunci, K. (2025). Penerapan Algoritma Random Forest untuk memprediksi Performa Akademik Mahasiswa. *Decode: Jurnal Pendidikan Teknologi Informasi*, 5(1), 262–270. <https://doi.org/10.51454/del>
- Maftucha, N., Salma, S., Rahmayuna, N., & Wakhidah, N. (2025). *Perbandingan Algoritma Machine Learning Dalam Memprediksi Kelulusan Siswa*. 19(2).
- Putra, L. G. R., Prasetya, D. D., & Mayadi, M. (2025). Student Dropout Prediction Using Random Forest and XGBoost Method. *INTENSIF: Jurnal Ilmiah Penelitian Dan Penerapan Teknologi Sistem Informasi*, 9(1), 147–157. <https://doi.org/10.29407/intensif.v9i1.21191>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11). <https://doi.org/10.3390/data7110146>
- A. G. Okunlola, O. A.; Ogunlade, “No Title,” *J. Educ. Data Min.*, 2023, [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1177/15313204231178906>

- A. Asad, M. N.; Fatima, "No Title," *Int. J. Comput. Appl.*, vol. 178 (33), 2021, [Online]. Available: <https://www.ijcaonline.org/archives/volume178/number33/asad-fatima-2021-ijca-921312.pdf>
- S. S. El-Halees, A.; Abu-Naser, "A Comparative Study of Machine Learning Algorithms for Predicting Student Academic Failure," *Int. J. Acad. Pedagog. Res.*, 2020, [Online]. Available: International Journal of Academic Pedagogical Research
- García-Retuerta, D., and et al. 2021. "Predicting Student Dropout Using Imbalanced Datasets: A Machine Learning Approach with SMOTE." *Applied Sciences*.
- Ahmad, M, et al. 2020. "Dropout Prediction in Higher Education Institutions Using Deep Neural Networks and Ensemble Methods." *Procedia Computer Science*.
- Badr, I., et al. 2021. "Predicting student dropout in higher education based on academic and administrative data." *Educational Data Mining*.
- Costa, A., et al. 2022. "The Influence of Prior Education on Student Dropout Risk in University: A Machine Learning Approach." *Internatiaonal Journal of Educational Technology in Higher Education*.