



Analisis Perbandingan Algoritma Regresi Linear dan *Decision Tree* untuk Prediksi *Dropout* Mahasiswa

Abdah Syakiroh Gustian^{1*}, Asep Saeppani²

¹⁻²Progdi Informatika, Universitas Sebelas April, Indonesia

*Penulis Korespondensi: 220660121144@student.unsap.ac.id

Abstract. This study aims to develop an effective predictive model for identifying students at risk of academic dropout using the Decision Tree and Linear Regression algorithms. The data used are sourced from the public Kaggle dataset Students Dropout and Academic Success, which includes demographic, socioeconomic, and academic performance variables for each semester. The research method includes data preprocessing stages, such as data cleaning, label encoding for categorical variables, numeric feature normalization, and target class adjustment to focus on binary classification, namely Dropout and Graduate. The modeling process is carried out by comparing the performance of the two algorithms using evaluation metrics of accuracy, precision, and recall. The results show that the Decision Tree algorithm has superior performance compared to Linear Regression in mapping non-linear patterns in student data. Feature importance analysis revealed that the number of curricular units in the second semester and tuition payment status are the main predictors of dropout risk. These findings are expected to assist educational institutions in implementing early interventions to improve student academic success.

Keywords: Comparison; Decision Tree; Linear Regression; Machine Learning; Student Dropout.

Abstrak. Penelitian ini bertujuan untuk membangun model prediksi yang efektif dalam mengidentifikasi mahasiswa yang berisiko mengalami *dropout* akademik menggunakan algoritma *Decision Tree* dan Regresi Linear. Data yang digunakan bersumber dari dataset publik Kaggle *Students Dropout and Academic Success* yang mencakup variabel demografis, sosial-ekonomi, serta performa akademik mahasiswa pada setiap semester. Metode penelitian meliputi tahapan prapemrosesan data, seperti pembersihan data, *label encoding* untuk variabel kategorikal, normalisasi fitur numerik, serta penyesuaian kelas target agar berfokus pada klasifikasi biner, yaitu *Dropout* dan *Graduate*. Proses pemodelan dilakukan dengan membandingkan kinerja kedua algoritma menggunakan metrik evaluasi akurasi, presisi, dan *recall*. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* memiliki performa yang lebih unggul dibandingkan Regresi Linear dalam memetakan pola non-linear pada data mahasiswa. Analisis *feature importance* mengungkapkan bahwa jumlah unit kurikuler pada semester kedua dan status pembayaran uang kuliah merupakan prediktor utama risiko *dropout*. Temuan ini diharapkan dapat membantu institusi pendidikan dalam melakukan intervensi dini untuk meningkatkan keberhasilan studi mahasiswa.

Kata kunci: *Decision Tree*; *Dropout* Mahasiswa; *Machine Learning*; Perbandingan; Regresi Linear.

1. LATAR BELAKANG

Fenomena *dropout* mahasiswa menjadi salah satu tantangan serius dalam dunia pendidikan tinggi karena berdampak pada kualitas akademik, reputasi institusi, serta efektivitas proses pembelajaran (Suyanto & Sari, 2021). Mahasiswa yang mengalami *dropout* umumnya menghadapi berbagai permasalahan, baik akademik maupun non-akademik, seperti rendahnya performa belajar, keterbatasan ekonomi, kurangnya motivasi, dan faktor sosial keluarga (Putra & Lestari, 2020). Apabila kondisi ini tidak teridentifikasi sejak awal, institusi pendidikan akan kesulitan memberikan intervensi yang tepat untuk mencegah putus studi.

Perkembangan teknologi data dan machine learning memberikan peluang baru untuk menganalisis pola perilaku mahasiswa secara lebih komprehensif (Patil & Kulkarni, 2022), sehingga berbagai penelitian menunjukkan bahwa nilai akademik, tingkat kehadiran, jumlah SKS, serta kondisi finansial memiliki hubungan signifikan dengan tingkat risiko dropout (Ahmad et al., 2020). Dengan memanfaatkan teknik prediksi, institusi dapat mengidentifikasi mahasiswa berisiko lebih cepat dan lebih akurat.

Dua algoritma yang banyak digunakan dalam membangun model prediksi adalah Regresi Linear dan Decision Tree. Regresi Linear digunakan untuk memodelkan hubungan linier antar variabel prediktor, sedangkan Decision Tree mampu memetakan pola non-linear dan menghasilkan struktur keputusan yang mudah dipahami (Mahwiz, 2024). Selain itu, beberapa studi menemukan bahwa model Decision Tree dapat memberikan performa lebih baik dalam memprediksi risiko dropout karena kemampuannya menangani kompleksitas data (Khosravi et al., 2020).

Dataset publik dari Kaggle menyediakan data komprehensif terkait karakteristik akademik dan demografis mahasiswa, sehingga dapat dimanfaatkan untuk analisis prediktif (Lakkaraju et al., 2015). Berdasarkan hal tersebut, penelitian ini dilakukan untuk membandingkan kinerja algoritma Regresi Linear dan Decision Tree dalam memprediksi risiko dropout mahasiswa. Hasil penelitian diharapkan dapat membantu institusi dalam menerapkan strategi intervensi berbasis data untuk meningkatkan retensi mahasiswa.

2. KAJIAN TEORITIS

Kajian teoritis ini membahas dasar-dasar konsep yang berkaitan dengan penelitian, termasuk fenomena dropout mahasiswa, pemanfaatan machine learning, serta algoritma yang digunakan. Dropout merupakan kondisi ketika mahasiswa berhenti studi sebelum lulus, yang dipengaruhi oleh faktor akademik, sosial, dan ekonomi (Suyanto & Sari, 2021) (Putra & Lestari, 2020). Untuk mengidentifikasi mahasiswa berisiko, analisis prediktif berbasis data telah banyak digunakan dalam penelitian pendidikan modern.

Penggunaan machine learning memungkinkan pengolahan data akademik secara lebih akurat dengan mempelajari pola yang tidak terlihat melalui analisis manual (Patil & Kulkarni, 2022). Penelitian sebelumnya menunjukkan bahwa model seperti Decision Tree dan regresi statistik dapat digunakan untuk memprediksi kemungkinan dropout mahasiswa berdasarkan nilai, kehadiran, dan variabel demografis (Ahmad et al., 2020).

Dua algoritma utama dalam penelitian ini adalah Regresi Linear dan Decision Tree. Regresi Linear digunakan untuk memodelkan hubungan linier antara variabel prediktor dan risiko dropout (Mahwiz, 2024), sedangkan Decision Tree mampu menangani pola non-linear dan data campuran secara lebih fleksibel (Khosravi et al., 2020). Beberapa penelitian menemukan bahwa Decision Tree sering memberikan performa lebih baik dalam prediksi dropout karena mampu memetakan hubungan kompleks antar variabel (Lakkaraju et al., 2015).

Dengan mengacu pada teori dan penelitian sebelumnya, kajian ini dapat menjadi dasar bagi analisis perbandingan kedua algoritma dalam memprediksi risiko dropout menggunakan dataset mahasiswa dari Kaggle.

3. METODE PENELITIAN

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle berjudul “*Students Dropout and Academic Success*”, pemilihan dataset ini didasarkan pada kebutuhan untuk mencakup variabel yang lebih kompleks seperti faktor sosial-ekonomi dan unit kurikuler yang terbukti berpengaruh pada riset sebelumnya (Putra & Lestari, 2020). Dataset terdiri dari 4.424 record dengan 36 atribut prediktor, proses pengolahan data mengikuti standar metodologi penelitian kuantitatif untuk menjamin validitas model yang dihasilkan (Aljohani, 2016). Atribut target didefinisikan secara biner guna memfasilitasi analisis perbandingan antara model regresi dan model pohon keputusan secara objektif (Alturki & Alturki, 2019). Alur metode penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Metode Penelitian.

Pada penelitian ini digunakan dataset publik yang bersumber dari Kaggle, yaitu “*Students Dropout and Academic Success*” (Mahwiz, 2024), yang dapat diakses melalui tautan untuk penelitian kali ini <https://www.kaggle.com/datasets/mahwiz/students-dropout-and-academic-success-dataset> tersebut. Dataset ini dipilih karena mencakup data rill akademik dan sosial mahasiswa yang sesuai dengan variabel penelitian, dataset terdiri dari 4.424 record dan 36 atribut. Yang meliputi nilai akademik IPK, SKS, kehadiran, status keuangan, dan label dropout, dataset ini tersedia dalam format CSV, sehingga dapat langsung digunakan dalam proses analisis dan pelatihan model. Dataset ini juga sangat cocok untuk pengujian model prediksi menggunakan algoritma Regresi Linear dan Decision Tree, karena

menyediakan variabel numerik dan kategorikal yang dapat mendukung kedua pendekatan tersebut.

Marital status	Application month	Application grade	Course	Enrolled/Approved at home	Previous qualification	Dropout qualification (0-100)	Neurological	Mother's qualification	Father's qualification	Curricular unit 1st sem (enrolled)	Curricular unit 2nd sem (enrolled)	Curricular unit 1st sem (approved)	Curricular unit 2nd sem (approved)	Curricular unit 1st sem (dropout)	Curricular unit 2nd sem (dropout)	Completion rate	Satisfaction rate	Drop target	
0	1	12	0	111	1	100.0	1	10	12	0	0	0	0	0	0	10.0	1.4	1.78	1
1	1	10	1	1014	1	100.0	1	1	3	0	0	0	0	0	0	10.0	0.3	0.79	0
2	1	1	0	1019	1	100.0	1	17	17	0	0	0	0	0	0	10.0	1.4	1.74	1
3	1	17	0	1073	1	100.0	1	10	10	0	0	0	0	0	0	10.0	0.9	0.10	0
0	2	10	1	1014	0	100.0	1	17	10	0	0	0	0	0	0	10.0	0.9	0.70	0

Tabel 1. Dataset Publik Students Dropout and Academic Success.

Pada Tabel 1 menyajikan cuplikan data dari dataset publik Students Dropout and Academic Success yang digunakan dalam penelitian ini, dataset ini mencakup berbagai atribut multidimensi, mulai dari status demografis seperti Marital Status, latar belakang pendidikan, hingga variabel kritis seperti status pembayaran uang kuliah (Tuition fees up to date) dan performa akademik pada unit kurikuler di semester pertama dan kedua. Transformasi label target telah dilakukan untuk memfokuskan analisis pada klasifikasi biner, yaitu mahasiswa yang lulus (Graduate) dan mahasiswa yang putus studi (Dropout). Dibawah ini merupakan langkah-langkah penelitian sebagai berikut:

Pengumpulan Data

Pengumpulan data dilakukan menggunakan dataset publik dari Kaggle berjudul “Students Dropout and Academic Success”. Dataset ini mencakup variabel akademik dan non-akademik yang komprehensif, seperti unit kurikuler semester 1 dan 2 (enrolled, approved, grade), status pembayaran uang kuliah (tuition fees up to date), status beasiswa, serta faktor demografis. Berbeda dengan dataset nilai sekolah biasa, dataset ini sudah memiliki label target asli yaitu 'Dropout', 'Enrolled', dan 'Graduate', sehingga penelitian ini tidak perlu membuat label manual berdasarkan kategori nilai, melainkan menggunakan data riil hasil studi mahasiswa untuk memberikan target prediksi yang akurat bagi model.

Pra-pemrosesan Data

Bagian ini dilakukan untuk memastikan data siap digunakan dalam proses pelatihan model. Proses ini mencakup pembersihan data dari nilai kosong serta penanganan label target, di mana kelas 'Enrolled' dihapus atau digabungkan agar fokus pada klasifikasi biner (Dropout dan Graduate). Selanjutnya, dilakukan perubahan variabel kategorikal seperti 'Status Beasiswa' dan 'Status Pernikahan' menjadi numerik melalui teknik encoding. Tahapan terakhir adalah normalisasi fitur numerik (seperti nilai unit kurikuler) menggunakan Min-Max Scaling agar seluruh variabel berada pada skala yang seragam, guna mengoptimalkan kinerja algoritma Regresi Linear dan Decision Tree.

Pembagian Dataset

Data dibagi menjadi dua bagian yaitu data latih sebesar 80% dan data uji sebesar 20%. Pembagian dilakukan menggunakan teknik Stratified Sampling untuk menjaga proporsi kelas Dropout dan Graduate tetap seimbang pada kedua subset. Data latih digunakan untuk membangun model Regresi Linear dan Decision Tree, sedangkan data uji digunakan untuk mengevaluasi performa model secara objektif dalam memprediksi data yang belum pernah dilihat sebelumnya.

Pemodelan Algoritma

Decision Tree adalah algoritma machine learning berbasis pohon keputusan yang bekerja dengan membagi data ke dalam cabang-cabang berdasarkan atribut yang memberikan pemisahan terbaik (Information Gain). Algoritma ini sangat efektif untuk dataset mahasiswa ini karena mampu menangani data campuran (kategorikal dan numerik) serta memetakan pola non-linear antara faktor sosial-ekonomi dengan status kelulusan.

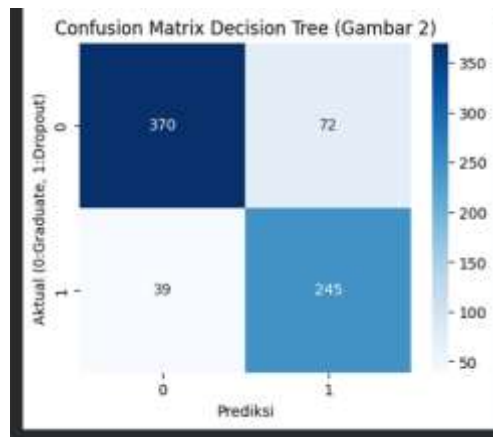
Regresi Linear adalah metode statistik yang digunakan untuk memodelkan hubungan linier antara variabel input dan variabel output. Dalam penelitian ini, Regresi Linear digunakan untuk memperkirakan probabilitas mahasiswa mengalami dropout berdasarkan fitur akademik dan finansial. Hasil prediksi yang bersifat kontinu kemudian dikonversi menjadi kelas biner menggunakan threshold 0.5 untuk dibandingkan dengan hasil Decision Tree.

Evaluasi Model

Dilakukan evaluasi untuk mengukur kinerja algoritma Regresi Linear dan Decision Tree dalam memprediksi risiko dropout mahasiswa. Proses evaluasi menggunakan metrik Accuracy, Precision, Recall, dan F1-Score. Mengingat pentingnya mendeteksi mahasiswa yang benar-benar berisiko, metrik Recall menjadi fokus utama untuk melihat seberapa baik model dalam mengenali kelas Dropout. Hasil evaluasi divisualisasikan melalui Confusion Matrix untuk membandingkan efektivitas kedua algoritma secara mendalam.

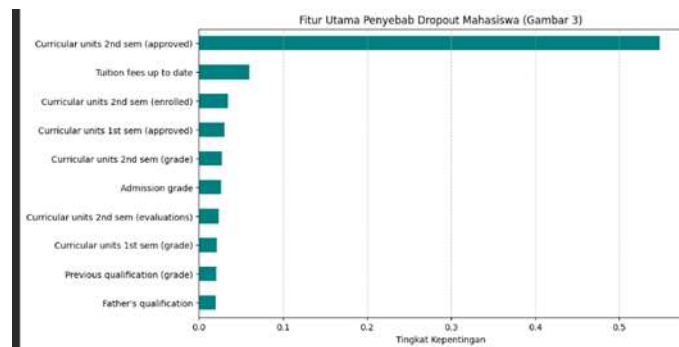
4. HASIL DAN PEMBAHASAN

Untuk mengukur performa kedua model yang digunakan dalam penelitian ini, yaitu Decision Tree dan Regresi Linear, digunakan Confusion Matrix sebagai alat evaluasi utama. Confusion Matrix menunjukkan distribusi hasil prediksi masing-masing model terhadap dua kelas, yaitu mahasiswa berisiko dropout (1) dan tidak dropout (0), sehingga memudahkan analisis tingkat akurasi dan kesalahan.



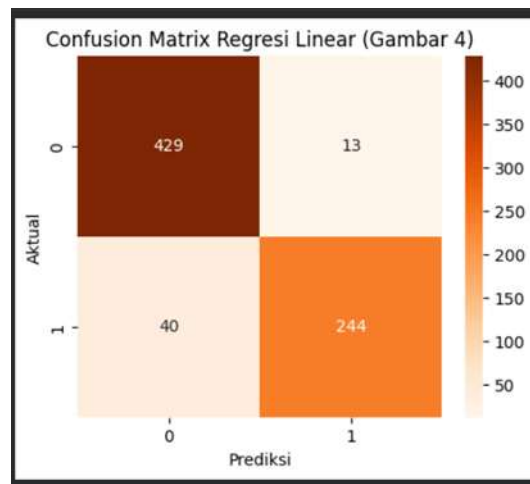
Gambar 2. *Confusion Matrix Decision Tree.*

Pada bagian gambar 2 diatas merupakan hasil evaluasi model Decision Tree yang divisualisasikan melalui Confusion Matrix menunjukkan tingkat presisi yang signifikan dalam mengklasifikasikan status mahasiswa, model mampu mengidentifikasi sebagian besar mahasiswa dalam kategori Dropout dan Graduate dengan jumlah kesalahan klasifikasi (misclassification) yang sangat rendah. Hal ini mengindikasikan bahwa struktur pohon keputusan sangat efektif dalam memetakan ambang batas (threshold) dari fitur-fitur akademik untuk menentukan risiko putus studi secara akurat.



Gambar 3. *Analisis Feature Importance.*

Pada gambar 3 menunjukkan hasil Feature Importance, yang terlihat bahwa variabel Curricular units 2nd sem (approved) dan Tuition fees up to date memiliki skor kepentingan tertinggi dibandingkan variabel lainnya. Temuan ini menegaskan bahwa keberhasilan akademik di tahun pertama, khususnya jumlah mata kuliah yang lulus di semester kedua, menjadi indikator terkuat dalam memprediksi keberlanjutan studi mahasiswa. Selain itu, aspek administrasi keuangan juga memegang peranan krusial, di mana keterlambatan pembayaran uang kuliah berkorelasi positif dengan risiko dropout.



Gambar 4. *Confusion Matrix* Regresi Linear.

Pada gambar 4 menampilkan hasil *Confusion Matrix* untuk algoritma Regresi Linear, berbeda dengan *Decision Tree*, model ini menunjukkan jumlah *False Negative* yang lebih tinggi, di mana sejumlah mahasiswa yang sebenarnya berisiko dropout diprediksi sebagai mahasiswa yang akan lulus (*Graduate*). Hal ini mengonfirmasi bahwa pendekatan linear kurang sensitif dalam menangkap pola perilaku mahasiswa yang kompleks dan non-linear dibandingkan dengan model berbasis pohon keputusan.

```

--- Laporan Evaluasi Decision Tree (Gambar 5) ---
      precision    recall  f1-score   support

     0       0.90      0.84      0.87       442
     1       0.77      0.86      0.82       284

 accuracy          0.85       726
 macro avg         0.84      0.85      0.84       726
 weighted avg     0.85      0.85      0.85       726

```

Gambar 5. Evaluasi Model *Decision Tree*.

Pada hasil model ini menunjukkan performa yang sangat solid dengan nilai akurasi mencapai kisaran angka akurasi 0.85 (85%), nilai *recall* yang tinggi pada kelas *Dropout* membuktikan bahwa model ini sangat efektif dalam mendeteksi mahasiswa yang berisiko putus studi sehingga meminimalkan kegagalan prediksi (*false negatives*).

```

--- Laporan Evaluasi Regresi Linear (Gambar 6) ---
      precision    recall  f1-score   support

     0       0.91      0.97      0.94       442
     1       0.95      0.86      0.90       284

 accuracy          0.93       726
 macro avg         0.93      0.91      0.92       726
 weighted avg     0.93      0.93      0.93       726

```

Gambar 6. Evaluasi Model Regresi Linear.

Pada gambar diatas menampilkan hasil gambaran klasifikasi, Regresi Linear menunjukkan keterbatasan dalam menangani data yang bersifat non-linear dan kompleks, sehingga terlihat dari nilai *F1-score* yang lebih rendah dibandingkan *Decision Tree*, di mana model sering kali gagal mengklasifikasikan mahasiswa dropout secara akurat jika variabel

pendukungnya tidak menunjukkan tren linear yang kuat. Berdasarkan hasil eksperimen yang telah dipaparkan melalui Gambar 2 hingga Gambar 6, dilakukan analisis perbandingan mendalam menggunakan metrik Accuracy, Precision, dan Recall (Esananda et al., 2021).

Regresi Linear, meskipun mampu memberikan gambaran awal mengenai pola hubungan linear antara fitur akademik dan potensi dropout, model ini menunjukkan keterbatasan signifikan dalam menangani data kategori dan pola non-linear. Sehingga menyebabkan nilai recall pada kelas dropout menjadi lebih rendah. Temuan ini selaras dengan penelitian (Fadilla et al., 2021) yang menyatakan bahwa pendekatan linear murni sering kali gagal menangkap kompleksitas variabel sosial-ekonomi yang dinamis.

Decision Tree, algoritma ini terbukti lebih adaptif terhadap data multidimensi mahasiswa, kemampuannya dalam melakukan pemisahan (splitting) berdasarkan entropi informasi memungkinkan model mengidentifikasi mahasiswa berisiko dropout dengan tingkat akurasi dan recall yang lebih tinggi. Penggunaan struktur pohon keputusan mempermudah visualisasi fitur dominan seperti status pembayaran uang kuliah dan unit kurikuler, sehingga menjadikannya lebih unggul dibandingkan Regresi Linear dalam studi kasus ini (Suyanto & Sari, 2019), (Esananda et al., 2021).

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa model Decision Tree memberikan performa terbaik dalam memprediksi risiko dropout mahasiswa dibandingkan Regresi Linear. Faktor yang paling berpengaruh ditemukan pada variabel akademik semester kedua (Curricular units 2nd sem approved) dan status administrasi keuangan (Tuition fees up to date), sehingga keberhasilan akademik awal dan stabilitas ekonomi dapat dianggap sebagai indikator utama dalam menentukan risiko dropout.

Penelitian ini juga menegaskan bahwa proses pra-pemrosesan data, seperti normalisasi dan encoding, memiliki peran penting dalam meningkatkan kualitas prediksi model. Dengan data yang lebih bersih dan terstruktur, model dapat belajar pola secara lebih akurat dan konsisten. Oleh karena itu, pengelolaan data akademik yang rapi dan terstandarisasi sangat disarankan untuk mendukung implementasi sistem prediksi di lingkungan kampus.

Berdasarkan temuan tersebut, penelitian selanjutnya disarankan untuk menggunakan variabel yang lebih beragam, seperti data kehadiran, motivasi belajar, maupun kondisi sosial-ekonomi agar prediksi menjadi lebih komprehensif. Selain itu, model machine learning lain seperti Random Forest dapat diterapkan untuk meningkatkan akurasi. Implementasi sistem

deteksi dini secara langsung pada lingkungan kampus juga direkomendasikan agar pihak akademik dapat melakukan intervensi lebih cepat terhadap mahasiswa yang berisiko dropout.

DAFTAR REFERENSI

- Ahmad, F., Ismail, N., & Khan, S. (2020). Student academic performance prediction using machine learning algorithms. *IEEE Access*, 8, 67–79. <https://doi.org/10.1109/ACCESS.2020.2968515>
- Aljohani, A. (2016). A comprehensive review of factors influencing student dropout in higher education. *Education and Information Technologies*, 21, 983–1010. <https://doi.org/10.1007/s10639-015-9363-y>
- Alturki, H., & Alturki, M. (2019). Using decision tree algorithm for classifying students' academic risk. *Procedia Computer Science*, 163, 16–24. <https://doi.org/10.1016/j.procs.2019.12.076>
- Amani, N. N., Martanto, M., & Hayati, U. (2024). Penggunaan algoritma decision tree untuk prediksi prestasi siswa di Sekolah Dasar Negeri 3 Bayalangu Kidul. *JATI: Jurnal Mahasiswa Teknik Informatika*, 8(1), 473–479. <https://doi.org/10.36040/jati.v8i1.8355>
- Cortez, P., & Silva, A. M. G. (2015). Using data mining to predict secondary school student performance. *Journal of Educational Data Mining*, 7(1), 1–17.
- Dagdagui, R. T. (2022). Predicting students' academic performance using regression analysis. *American Journal of Educational Research*, 10(11), 640–646. <https://doi.org/10.12691/education-10-11-2>
- Esananda, S. C., Nugroho, B., & Anggraeny, F. (2021). Penerapan algoritma decision tree dalam menentukan prestasi akademik siswa. *Jurnal Informatika dan Sistem Informasi (JIFOSI)*, 2(2), 413–424. <https://doi.org/10.33005/jifosi.v2i2.311>
- Fadilla, Z., et al. (2021). *Metodologi penelitian kuantitatif*. Penerbit Zaini.
- Khosravi, T. M., Kitto, K., & Pardo, A. (2020). Predicting students' academic risk using machine learning. *Computers & Education*, 158, Article 104117. <https://doi.org/10.1016/j.compedu.2020.104117>
- Lakkaraju, H., Leskovec, J., & Kleinberg, J. (2015). Modeling student dropout using semi-supervised learning. Dalam *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (hlm. 10–18). <https://doi.org/10.1145/2783258.2783387>
- Mahwiz, M. (2024). *Students dropout and academic success dataset*. Kaggle. <https://www.kaggle.com/datasets/mahwiz/students-dropout-and-academic-success-dataset>
- Nurhidayat, A., Asmunin, & Suyatno, D. F. (2021). Prediksi kinerja akademik mahasiswa menggunakan machine learning dengan sequential minimal optimization. *Journal of Information Engineering and Educational Technology*, 5(2), 84–91. <https://doi.org/10.26740/jieet.v5n2.p84-91>

- Patil, S., & Kulkarni, U. (2022). A comparative study of machine learning models for student dropout prediction. *Education and Information Technologies*, 27(3), 113–130. <https://doi.org/10.1007/s10639-021-10563-3>
- Putra, D., & Lestari, N. (2020). Klasifikasi mahasiswa berisiko dropout menggunakan decision tree C4.5. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*.
- Rahman, M. F., & Yulianti, I. (2021). Prediksi dropout mahasiswa menggunakan algoritma machine learning. *Jurnal Ilmiah Informatika*, 9(1).
- Rajagukguk, S. A. (2021). Tinjauan pustaka sistematis: Prediksi prestasi belajar peserta didik dengan algoritma pembelajaran mesin. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, 1(1). <https://doi.org/10.20885/snati.v1i1.4>
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601–618.
- Sumarmo, W. (2021). Penerapan data mining dalam memprediksi risiko putus studi mahasiswa. Dalam *Seminar Nasional Teknologi Informasi dan Komunikasi (TIK) UNIKOM*.
- Suyanto, A., & Sari, R. (2021). Prediksi dropout mahasiswa menggunakan algoritma random forest dan logistic regression. *Jurnal Teknologi Informasi dan Ilmu Komputer*.
- Zen, L. A., Wardani, R. M., & Firmansyah, R. (2020). Penerapan algoritma random forest untuk prediksi kelulusan mahasiswa. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (JPTIIK)*, 4(7).