Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

Prediksi Risiko Putus Kuliah Mahasiswa Berdasarkan Atribut Sosio-Ekonomi Menggunakan Ensemble Learning: Studi Eksperimental pada Dataset UCI

Predicting Student Dropout Risk Based on Socioeconomic Attributes Using Ensemble Learning: An Experimental Study on the UCI Dataset

Sugeng Hendra Wijaya[1,*], Novi Andayaningtyas[2]

[1]Teknik, Teknik Informatika, Universitas Bakti Indonesia, Banyuwangi, Indonesia
Jalan Kampus Bumi Cempokosari No.40, Dusun Cempokasari, Sarimulyo, Kec. Cluring, Kabupaten Banyuwangi, Jawa Timur, Indonesia
[2]Teknik, Sistem Informasi, Universitas Bakti Indonesia, Banyuwangi, Indonesia
Jalan Kampus Bumi Cempokosari No.40, Dusun Cempokasari, Sarimulyo, Kec. Cluring, Kabupaten Banyuwangi, Jawa Timur, Indonesia
Email: [1,*]sugeng@ubibanyuwangi.ac.id, [2]novityas@ubibanyuwangi.ac.id
Email Penulis Korespondensi: sugeng@ubibanyuwangi.ac.id
Received: **06/10/2026**, Revised: **26/10/2026**, Accepted: **28/10/2026**, Published**: 31/10/2026**

Abstrak−Angka putus kuliah (dropout) merupakan salah satu indikator penting dalam evaluasi keberhasilan penyelenggaraan pendidikan tinggi. Keterlambatan dalam mengidentifikasi mahasiswa yang berpotensi mengalami dropout dapat menyebabkan intervensi akademik maupun finansial dilakukan setelah permasalahan berkembang. Penelitian ini bertujuan membandingkan kinerja tiga algoritma ensemble learning, yaitu Random Forest, AdaBoost, dan Gradient Boosted Trees (GBT), dalam memprediksi status dropout mahasiswa berdasarkan atribut sosio-ekonomi. Penelitian menggunakan dataset sekunder Predict Students Dropout and Academic Success dari UCI Machine Learning Repository. Setelah mahasiswa dengan status Enrolled dikeluarkan, diperoleh 3.630 data dengan dua kelas, yaitu Dropout dan Graduate. Pemodelan dilakukan menggunakan platform RapidMiner dengan pembagian data 80:20 dan evaluasi berdasarkan accuracy, sensitivity, specificity, dan Area Under the Curve (AUC). Hasil penelitian menunjukkan bahwa GBT memberikan kinerja terbaik dengan accuracy sebesar 84,16%, sensitivity sebesar 77,46%, specificity sebesar 88,46%, dan AUC sebesar 0,912. Analisis Information Gain menunjukkan bahwa Tuition fees up to date memiliki nilai informasi tertinggi sebesar 0,2145, diikuti Scholarship holder sebesar 0,1582 dan Debtor sebesar 0,0984. Kontribusi penelitian ini adalah memberikan perbandingan empiris kinerja tiga algoritma ensemble learning menggunakan atribut sosio-ekonomi serta mengidentifikasi atribut yang paling relevan dalam klasifikasi status dropout mahasiswa. Hasil penelitian dapat menjadi dasar pengembangan sistem peringatan dini yang membantu perguruan tinggi dalam melakukan pemantauan dan menentukan prioritas intervensi terhadap mahasiswa yang berpotensi mengalami dropout.

Kata Kunci: Ensemble Learning; Prediksi Putus Kuliah; Sosio-Ekonomi; RapidMiner; Sistem Peringatan Dini

Abstract−Student dropout is an important indicator of the success of higher education institutions. Delays in identifying students who are potentially at risk of dropping out may result in academic and financial interventions being provided only after problems have developed. This study aims to compare the performance of three ensemble learning algorithms, namely Random Forest, AdaBoost, and Gradient Boosted Trees (GBT), in predicting student dropout status based on socio-economic attributes. The study uses the secondary Predict Students Dropout and Academic Success dataset from the UCI Machine Learning Repository. After removing students with the Enrolled status, 3,630 records with two classes, Dropout and Graduate, were obtained. The modeling process was conducted using RapidMiner with an 80:20 data split and evaluated using accuracy, sensitivity, specificity, and Area Under the Curve (AUC). The results show that GBT achieved the best performance with an accuracy of 84.16%, sensitivity of 77.46%, specificity of 88.46%, and AUC of 0.912. Information Gain analysis shows that Tuition fees up to date has the highest information value of 0.2145, followed by Scholarship holder at 0.1582 and Debtor at 0.0984. The contribution of this study is to provide empirical evidence on the comparative performance of three ensemble learning algorithms using socio-economic attributes and to identify the most relevant attributes for classifying student dropout status. The findings can serve as a basis for developing an early warning system to support higher education institutions in monitoring students and prioritizing interventions for those potentially at risk of dropping out.

Keywords: Ensemble Learning; Dropout Prediction; Socio-Economic; RapidMiner; Early Warning System

1. PENDAHULUAN

Pendidikan tinggi memegang peranan yang sangat strategis dalam mencetak sumber daya manusia yang berkualitas, responsif terhadap perkembangan teknologi, dan mampu memajukan daya saing bangsa di kera global. Perguruan tinggi tidak hanya berfungsi sebagai wadah transfer pengetahuan, melainkan juga sebagai pusat pengembangan karakter dan kompetensi profesional mahasiswa. Oleh karena itu, keberhasilan suatu institusi pendidikan tinggi diukur dari kemampuannya dalam mengantarkan mahasiswa menyelesaikan studi mereka secara tepat waktu dan berkualitas. Namun, dalam realitas operasionalnya, perguruan tinggi sering kali dihadapkan pada tantangan besar berupa tingginya angka putus kuliah (drop out). Fenomena mahasiswa yang berhenti di tengah masa studinya merupakan masalah kompleks yang merugikan banyak pihak [1], [2], [3]. Bagi mahasiswa dan keluarganya, putus kuliah berarti hilangnya investasi waktu, tenaga, biaya yang besar, serta pupusnya peluang

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 223
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

untuk memperoleh pekerjaan dan masa depan yang lebih baik. Bagi institusi perguruan tinggi swasta maupun negeri, tingginya angka mahasiswa yang tidak menyelesaikan studi berimplikasi buruk terhadap penurunan nilai akreditasi, berkurangnya pendapatan dari sektor dana operasional pendidikan, serta rusaknya reputasi institusi di mata masyarakat luas.

Secara umum, keputusan seorang mahasiswa untuk berhenti kuliah jarang sekali disebabkan oleh satu faktor tunggal, melainkan merupakan akumulasi dari interaksi berbagai variabel internal dan eksternal yang kompleks. Selama ini, sebagian besar kebijakan institusi cenderung berfokus pada performa akademik, seperti Indeks Prestasi Kumulatif (IPK) atau nilai ujian, sebagai indikator utama penentu keberlanjutan studi. Padahal, faktor non-akademik, khususnya kondisi sosio-ekonomi keluarga mahasiswa, memegang peranan yang tidak kalah krusial dalam memahami risiko keberlanjutan studi [3], [4], [5], [6], [7]. Komponen sosio-ekonomi ini mencakup tingkat pendapatan kotor orang tua, latar belakang pendidikan ayah dan ibu, jumlah beban tanggungan anak dalam keluarga, daerah asal mahasiswa, hingga status kepemilikan beasiswa pendidikan. Mahasiswa yang berasal dari latar belakang ekonomi lemah sering kali mengalami tekanan psikologis dan finansial yang berat, yang pada akhirnya memicu ketidakmampuan untuk membayar Uang Kuliah Tunggal (UKT) atau biaya hidup, sehingga mereka terpaksa mengambil keputusan untuk melakukan mutasi program studi, mengambil cuti akademis yang berkepanjangan, atau bahkan putus kuliah sepenuhnya.

Fenomena putus kuliah di lingkungan perguruan tinggi memiliki implikasi terhadap keberlanjutan studi mahasiswa dan efektivitas pengelolaan layanan pendidikan. Oleh karena itu, identifikasi faktor yang berkaitan dengan status dropout menjadi penting dalam mendukung proses pemantauan mahasiswa. Penelitian mengenai dropout tidak hanya dapat menggunakan indikator akademik seperti Indeks Prestasi Kumulatif (IPK) atau nilai ujian, tetapi juga dapat mempertimbangkan karakteristik non-akademik yang tersedia sebelum atau selama proses studi. Atribut sosio-ekonomi seperti latar belakang pendidikan orang tua, pekerjaan orang tua, kondisi pembayaran biaya pendidikan, status beasiswa, dan karakteristik demografis dapat memberikan informasi tambahan mengenai kondisi mahasiswa. Pemilihan atribut sosio-ekonomi dalam penelitian ini didasarkan pada pertimbangan bahwa informasi tersebut merepresentasikan kondisi non-akademik mahasiswa yang dapat digunakan sebagai informasi tambahan dalam proses identifikasi status dropout. Dari sisi teknis, penggunaan atribut non-akademik memungkinkan model mengevaluasi karakteristik mahasiswa di luar capaian akademik. Dari sisi pengelolaan pendidikan, informasi tersebut dapat membantu institusi mengidentifikasi kebutuhan dukungan mahasiswa secara lebih kontekstual, khususnya dalam aspek finansial dan sosial. Meskipun demikian, penggunaan atribut sosio-ekonomi tidak dimaksudkan untuk memberikan label atau keputusan otomatis terhadap mahasiswa, melainkan sebagai informasi pendukung yang perlu dikombinasikan dengan evaluasi akademik dan verifikasi kondisi individual mahasiswa [7].

Masalah mendasar yang dihadapi oleh pihak manajemen universitas, khususnya Biro Administrasi Akademik dan Kemahasiswaan (BAAK), adalah belum adanya mekanisme deteksi dini yang bersifat proaktif dan objektif. Prosedur pemantauan dan evaluasi yang berjalan saat ini umumnya masih bersifat reaktif konvensional. Pihak birokrasi kampus biasanya baru menyadari adanya mahasiswa yang bermasalah setelah mahasiswa tersebut tidak melakukan registrasi ulang (herregistrasi) selama beberapa semester berturut-turut, atau ketika batas masa studi maksimal hampir habis. Akibat keterlambatan informasi ini, segala bentuk intervensi manajemen yang seharusnya bisa diberikan—seperti program bimbingan konseling intensif oleh dosen wali, kebijakan dispensasi atau penurunan klaster UKT, hingga pengalokasian bantuan beasiswa darurat—menjadi terlambat dan tidak tepat sasaran. Mahasiswa bersangkutan sering kali sudah terlanjur keluar dari sistem pengawasan institusi sebelum bantuan atau solusi preventif dapat diupayakan secara maksimal.

Guna mengatasi keterbatasan sistem pemantauan konvensional tersebut, beberapa peneliti terdahulu telah mulai memanfaatkan perkembangan teknologi informasi, khususnya di bidang penambangan data (data mining) dan pembelajaran mesin (machine learning), untuk melakukan klasifikasi dan prediksi terhadap potensi kelulusan mahasiswa. Penggunaan algoritma klasifikasi tunggal seperti Decision Tree (C4.5), Naive Bayes, K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM) telah banyak diimplementasikan dalam riset-riset analitik pendidikan terdahulu untuk memetakan performa mahasiswa [8], [9], [10]. Akurasi model tunggal ini umumnya sangat bergantung pada homogenitas sebaran data awal, sehingga sering kali menghasilkan bias interpretasi yang tinggi jika diterapkan pada lingkungan institusi dengan latar belakang mahasiswa yang heterogen. Melalui algoritma-algoritma tersebut, data historis mahasiswa diolah sedemikian rupa untuk menghasilkan pola-pola keputusan tersembunyi yang menggambarkan karakteristik mahasiswa yang lulus tepat waktu atau terlambat. Meskipun metode-metode tunggal ini mampu memberikan hasil klasifikasi awal yang cukup baik, penerapannya pada data riil di institusi pendidikan sering kali menghadapi kendala serius terkait ketidakseimbangan kelas (imbalanced data), di mana jumlah sampel mahasiswa yang aktif jauh lebih mendominasi dibandingkan jumlah sampel mahasiswa yang putus kuliah. Akibatnya, model klasifikasi tunggal cenderung menghasilkan tingkat kesalahan (error rate) yang tinggi pada kelas minoritas (mahasiswa berisiko drop out), memiliki tingkat sensitivitas yang rendah, serta dapat menghadapi tantangan ketika distribusi kelas tidak seimbang dan prediktor bersifat heterogen [11], [12], [13].

Melihat adanya keterbatasan (research gap) pada model klasifikasi tunggal tersebut, penelitian ini menawarkan solusi inovatif berupa pendekatan komputasi cerdas berbasis ensemble learning. Prinsip dasar dari ensemble learning adalah mengombinasikan prediksi dari beberapa algoritma pembelajaran mesin dasar (base

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 224
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

classifiers) secara simultan untuk membentuk satu kesatuan model prediksi baru yang jauh lebih kuat, stabil, dan memiliki akurasi yang lebih tinggi [10], [14], [15]. Pendekatan multi-algoritma ini bekerja dengan mereduksi bias (boosting) dan varians (bagging) secara simultan, menjadikannya standar baru dalam komputasi analitik pendidikan modern. Melalui teknik penambangan data seperti bagging (Bootstrap Aggregating) dan boosting, beberapa model tangguh seperti Random Forest, Gradient Boosting, dan AdaBoost akan digabungkan untuk saling menutupi kelemahan masing-masing algoritma dasar saat memproses dataset yang tidak seimbang. Penelitian ini secara khusus menitikberatkan fokus pada pemanfaatan atribut sosio-ekonomi mahasiswa sebagai variabel prediktor utama dalam membangun model early warning system (sistem peringatan dini) [11], [16]. Pemilihan platform RapidMiner sebagai alat bantu eksperimen didasarkan pada keandalannya dalam melakukan pemrosesan data, kemudahan integrasi operator ensemble, serta kapabilitas visualisasi hasil evaluasi model yang mendalam.

Berdasarkan permasalahan dan kesenjangan penelitian tersebut, penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma ensemble learning, yaitu Random Forest, AdaBoost, dan Gradient Boosted Trees (GBT), dalam memprediksi status dropout mahasiswa berdasarkan atribut sosio-ekonomi. Penelitian ini memberikan beberapa kontribusi. Pertama, penelitian memberikan perbandingan empiris kinerja tiga algoritma ensemble learning pada dataset mahasiswa dengan memanfaatkan atribut sosio-ekonomi sebagai prediktor. Kedua, penelitian mengidentifikasi atribut sosio-ekonomi yang memiliki relevansi paling tinggi terhadap status dropout menggunakan metode Information Gain. Ketiga, penelitian memberikan implikasi praktis berupa pemanfaatan hasil pemodelan sebagai dasar pengembangan sistem peringatan dini (Early Warning System) yang dapat membantu perguruan tinggi dalam melakukan pemantauan dan menentukan prioritas intervensi terhadap mahasiswa yang berpotensi mengalami dropout. Dengan demikian, penelitian ini tidak hanya berfokus pada pencapaian kinerja prediksi, tetapi juga menghasilkan informasi yang dapat digunakan sebagai dasar pengambilan keputusan dalam pengelolaan retensi mahasiswa.

2. METODOLOGI PENELITIAN

Penelitian ini menerapkan metodologi eksperimental berbasis kuantitatif dengan pendekatan Knowledge Discovery in Databases (KDD), sebagaimana pendekatan data mining banyak digunakan dalam penelitian analitik pendidikan dan prediksi dropout [8], [1], [3]. Proses KDD dijalankan secara sistematis melalui beberapa tahapan utama, meliputi pemilihan data (data selection), pembersihan data (data preprocessing), transformasi data (data transformation), implementasi penambangan data (data mining), serta evaluasi hasil (evaluation). Alur komputasi dan eksperimen dalam penelitian ini seluruhnya dibangun dan dieksekusi menggunakan perangkat lunak RapidMiner.

2.1 Sumber Data dan Karakteristik Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset sekunder bertajuk "Predict Students Dropout and Academic Success" yang diperoleh dari UCI Machine Learning Repository [17], [18]. Dataset ini merekam data historis mahasiswa dari sebuah institusi pendidikan tinggi dengan total sampel sebanyak 4.424 data (instances). Dari keseluruhan atribut yang tersedia, penelitian ini secara spesifik mengisolasi dan berfokus pada subset atribut non-akademik yang merepresentasikan kondisi sosio-ekonomi mahasiswa. Atribut-atribut sosio-ekonomi yang digunakan sebagai variabel prediktor meliputi: Marital Status, Mother's Qualification, Father's Qualification, Mother's Occupation, Father's Occupation, Displaced, Debtor, Tuition fees up to date, dan Scholarship holder. Variabel target atau label dalam dataset ini adalah atribut Target yang memiliki tiga kelas awal, yaitu Dropout (Putus Kuliah), Enrolled (Masih Aktif Kuliah), dan Graduate (Lulus).

2.2 Pra-pemrosesan Data (Data Preprocessing)

Sebelum memasuki tahap pemodelan, dataset terlebih dahulu melalui tahapan pra-pemrosesan di RapidMiner. Tahap pertama adalah pembersihan data (data cleaning) untuk memastikan tidak terdapat nilai yang hilang (missing values). Selanjutnya, dilakukan reduksi data menggunakan operator Select Attributes guna mengeliminasi atribut akademik sehingga model yang terbentuk murni berbasis prediktor sosio-ekonomi.

Karena penelitian ini berfokus pada deteksi dini risiko putus kuliah, dilakukan penyaringan data (data filtering) menggunakan operator Filter Examples. Data dengan label Enrolled dieliminasi dari dataset untuk membentuk skema klasifikasi biner (binary classification) yang hanya menyisakan dua kelas tegas, yaitu kelas Dropout dan kelas Graduate. Terakhir, operator Set Role diterapkan untuk mengubah status atribut Target dari atribut biasa menjadi variabel label. Untuk menguji keandalan model prediksi secara objektif, dataset yang telah bersih dibagi menjadi dua bagian terpisah menggunakan operator Split Data dengan rasio pembagian data 80:20 (80% data latih dan 20% data uji).

2.3 Formulasi Matematika Algoritma Ensemble Learning

Inti dari arsitektur komputasi dalam penelitian ini adalah penerapan metode ensemble learning [10], [14], [15]. Pendekatan ini dipilih untuk mengatasi kelemahan akurasi model tunggal konvensional serta menangani masalah

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 225
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

ketidakseimbangan kelas (imbalanced data) yang melekat pada dataset mahasiswa putus kuliah. Tiga algoritma berbasis ensemble learning yang diuji di platform RapidMiner diformulasikan secara matematis sebagai berikut:

2.3.1 Random Forest

Random Forest adalah metode ensemble berbasis bagging yang membangun B buah pohon keputusan (T_1, T_2,...., T_B) secara acak dari sampel bootstrap data latih [14], [19]. Untuk membagi setiap simpul (node) pada pohon, algoritma memilih subset acak dari m prediktor sosio-ekonomi dari total M atribut yang tersedia. Prediksi akhir untuk data uji baru x ditentukan berdasarkan voting mayoritas dari seluruh pohon yang terbentuk, yang dinyatakan dalam persamaan:

$$\hat{C}_{RF}(x) = \arg\max_c \sum_{b=1}^{B} I\,(T_b(x) = c) \quad (1)$$

Dimana I (.) adalah fungsi indikator yang bernilai 1 jika prediksi pohon $T_b(x)$ sama dengan kelas c, dan bernilai 0 untuk kondisi lainnya.

2.3.2 AdaBoost (Adaptive Boosting)

AdaBoost bekerja secara sekuensial dengan memberikan bobot pada setiap data latih [12], [15]. Misalkan dataset latih berukuran n dinyatakan sebagai (x_1, y_1),...., (x_n, y_n) dengan label $y_i \in \{-1, 1\}$. Pada iterasi awal t=1, bobot data diinisialisasi secara merata: $D_1(i) = \frac{1}{n}$. Untuk setiap iterasi t = 1,....., T, algoritma melatih weak classifier $h_t(x)$ dan menghitung nilai kesalahan (error) ϵ_t :

$$\epsilon_t = \sum_{i=1}^{n} D_t\,(i)\; I\,(h_t\,(x_i) \neq y_i) \quad (2)$$

Bobot dari classifier h_t dihitung berdasarkan nilai α_t :

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon t}{\epsilon t}\right) \epsilon_t \quad (3)$$

Selanjutnya, bobot data diperbarui untuk iterasi berikutnya menggunakan persamaan:

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t} \quad (4)$$

Dimana Z_t adalah faktor normalisasi agar $\sum D_{t+1}(i) = 1$. Prediksi akhir dari model AdaBoost merupakan kombinasi linier berbobot:

$$H(x) = \text{sign}\,(\sum_{t=1}^{T} \alpha_t h_t(x)) \quad (5)$$

2.3.3 Gradient Boosted Trees (GBT)

Gradient Boosted Trees meminimalkan fungsi kerugian (loss function) L(y, f(x)) secara berulang menggunakan konsep gradient descent [12], [15]. Pada fungsi klasifikasi biner, model diinisialisasi dengan sebuah nilai konstan:

$$f_0\,(x) = \arg\min\gamma \sum_{i=1}^{n} L(y_i, \gamma) \quad (6)$$

Untuk setiap iterasi m = 1,....., M, algoritma menghitung pseudo-residuals (r_{im}), yang merupakan kemiringan negatif (negative gradient) dari fungsi kerugian:

$$r_{im} = -\,[\,\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\,]\; f(x) = f_{m-1}\,(x) \quad (7)$$

Sebuah pohon keputusan baru kemudian dilatih untuk mencocokkan nilai r_{im} tersebut. Model diperbarui dengan menjumlahkan hasil prediksi pohon baru yang dikalikan dengan laju pembelajaran (learning rate) υ:

$$f_m\,(x) = f_{m-1}\,(x) + \upsilon \sum_{j=1}^{Jm} \gamma j_m 1\,(x \in Rj_m \quad (8)$$

Dimana Rj_m adalah wilayah terminal dari pohon ke-m.

2.4 Konfigurasi Parameter Global (Hyperparameter Tuning)

Akurasi dan konvergensi dari ketiga model ensemble learning yang diuji dalam penelitian ini sangat bergantung pada pengaturan parameter internal (hyperparameters) pada masing-masing operator di RapidMiner Studio. Untuk menjamin keadilan uji komparatif (fair comparison), eksperimen ini melakukan tuning parameter secara terstruktur guna meminimalkan risiko overfitting. Konfigurasi arsitektur komputasi untuk masing-masing algoritma ditetapkan sebagai berikut:

1. Random Forest: Operator dikonfigurasi dengan parameter number of trees sebesar 100 pohon keputusan independen. Kriteria pemisahan simpul (criterion) menggunakan gain ratio untuk menangani variasi data kategori sosio-ekonomi yang tinggi, dengan maximal depth dibatasi pada nilai 10 level untuk mencegah kompleksitas struktural pohon yang berlebihan.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 226
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

2. AdaBoost: Operator ensemble ini menggunakan algoritma dasar (re-weighted base learner) berupa Decision Tree. Jumlah iterasi pengulangan sekuensial ditetapkan melalui parameter number of iterations sebesar 50, dengan kriteria pemisahan information gain.
3. Gradient Boosted Trees (GBT): Parameter number of trees ditentukan sebanyak 80 pohon secara sekuensial. Nilai laju pembelajaran (learning rate) atau υ diatur pada tingkat moderat sebesar 0,1 untuk menjamin stabilitas koreksi pseudo-residuals. Kedalaman maksimum pohon (maximal depth) disetel pada angka 5 level dengan mengaktifkan fitur early stopping guna menghentikan iterasi secara otomatis apabila tidak terjadi penurunan fungsi kerugian (loss function minimization) secara jelas pada 10 iterasi berturut-turut.

Penetapan hyperparameter dalam penelitian ini dilakukan secara manual sebagai konfigurasi eksperimen untuk menjaga kompleksitas model dan memungkinkan perbandingan yang konsisten antar algoritma. Pada Random Forest digunakan 100 pohon dengan kriteria pemisahan Gain Ratio dan kedalaman maksimum 10. Pada AdaBoost digunakan decision tree sebagai base learner dengan 50 iterasi dan kriteria Information Gain. Sementara itu, Gradient Boosted Trees menggunakan 80 pohon, learning rate sebesar 0,1, kedalaman maksimum 5, serta mekanisme early stopping setelah 10 iterasi tanpa peningkatan loss yang berarti. Konfigurasi tersebut digunakan secara tetap pada eksperimen sehingga perbandingan performa antar algoritma dilakukan pada kondisi pengujian yang sama. Penelitian ini belum menerapkan optimasi hyperparameter secara otomatis seperti Grid Search atau Evolutionary Algorithm. Oleh karena itu, konfigurasi yang digunakan diposisikan sebagai konfigurasi eksperimen dan bukan sebagai hasil optimasi hyperparameter global.

2.5 Evaluasi Performa Model

Evaluasi performa dari ketiga model ensemble learning diukur secara kuantitatif menggunakan operator Apply Model dan Performance (Classification) di RapidMiner. Parameter penilaian didasarkan pada matriks kekacauan (Confusion Matrix) yang disajikan pada Tabel 1.

Tabel 1. Confusion Matrix

	Aktual Positif (Dropout)	**Aktual Negatif (Graduate)**
Prediksi Positif	True Positive (TP)	False Positive (FP)
Prediksi Negatif	False Negative (FN)	True Negative (TN)

Berdasarkan Tabel 1, Confusion Matrix digunakan untuk menggambarkan kesesuaian antara hasil prediksi model dan kondisi aktual mahasiswa. Matriks tersebut terdiri atas empat komponen, yaitu True Positive (TP) yang menunjukkan mahasiswa berstatus Dropout dan berhasil diprediksi sebagai Dropout, False Positive (FP) yang menunjukkan mahasiswa berstatus Graduate tetapi diprediksi sebagai Dropout, False Negative (FN) yang menunjukkan mahasiswa berstatus Dropout tetapi diprediksi sebagai Graduate, serta True Negative (TN) yang menunjukkan mahasiswa berstatus Graduate dan berhasil diprediksi sebagai Graduate. Keempat komponen tersebut menjadi dasar perhitungan metrik evaluasi untuk mengukur kemampuan klasifikasi masing-masing algoritma. Metrik evaluasi utama dihitung berdasarkan formulasi matematika berikut:
Akurasi (Accuracy): Mengukur rasio prediksi benar terhadap keseluruhan data uji.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

Sensitivas (Sensitivity / Recall): Mengukur kemampuan model dalam mendeteksi secara akurat kelompok mahasiswa yang benar-benar berisiko putus kuliah (True Positive Rate).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{10}$$

Area Under the Curve (AUC) [20], [21]: Mengukur kinerja diskriminasi model melalui kurva Receiver Operating Characteristic (ROC) yang memplot hubungan antara True Positive Rate (TPR) pada sumbu Y dan False Positive Rate (FPR) pada sumbu X, di mana FPR = $\frac{FP}{TN+FP}$. Nilai AUC berkisar antara 0,0 hingga 1,0, di mana nilai yang mendekati 1,0 mengindikasikan bahwa model memiliki tingkat keandalan yang sempurna (excellent classification).

Untuk menjaga interpretasi hasil tetap proporsional, perlu dicatat bahwa eksperimen menggunakan pembagian data latih dan data uji dengan rasio 80:20 pada satu skema eksperimen. Penelitian ini belum menerapkan repeated cross-validation atau validasi eksternal menggunakan dataset dari institusi yang berbeda. Dengan demikian, hasil performa yang diperoleh merepresentasikan kinerja model pada dataset dan konfigurasi eksperimen yang digunakan serta masih memerlukan pengujian lebih lanjut sebelum digeneralisasikan pada konteks perguruan tinggi tertentu.

3. HASIL DAN PEMBAHASAN

Bagian ini menguraikan seluruh temuan eksperimental dan analisis mendalam mengenai implementasi pendekatan ensemble learning untuk memprediksi risiko putus kuliah mahasiswa berbasis atribut sosio-ekonomi. Pemaparan

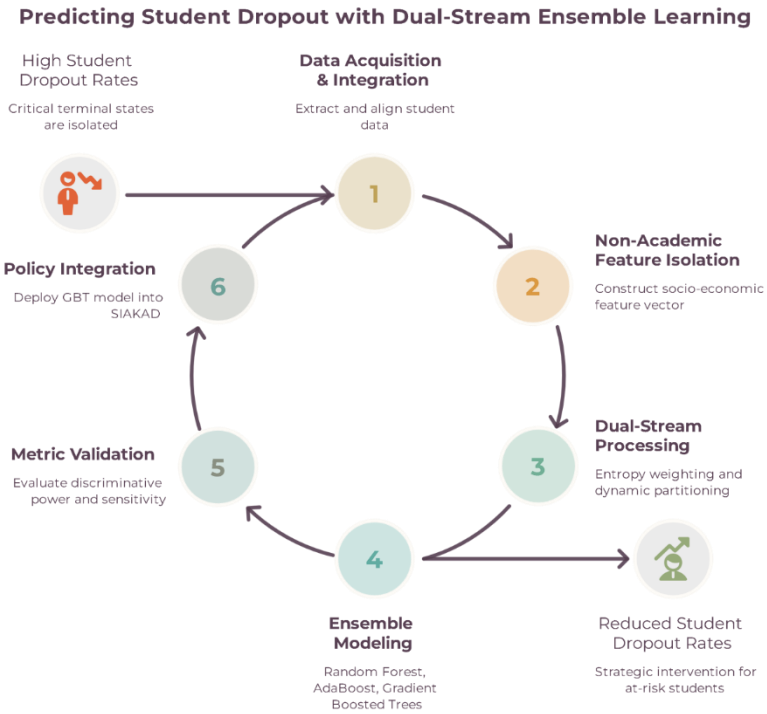
Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 227
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

disusun secara terstruktur, diawali dengan visualisasi alur metodologi eksperimen yang mendasari proses penambangan data (data mining). Selanjutnya, disajikan analisis deskriptif dataset, hasil pengujian pembobotan fitur menggunakan Information Gain, komparasi performa tiga algoritma ensemble (Random Forest, AdaBoost, dan Gradient Boosted Trees), serta diakhiri dengan diskusi mengenai implikasi manajerial temuan ini terhadap sistem tata kelola di perguruan tinggi.

3.1 Alur Metodologi Eksperimen (Experimental Framework)

Untuk menjamin validitas dan replikasi penelitian, eksperimen komputasi ini bersandar pada kerangka kerja Knowledge Discovery in Databases (KDD) yang dijalankan melalui platform RapidMiner Studio. Proses ekstraksi pengetahuan ini mengikuti standardisasi industri yang ketat agar hasil pemodelan bersifat konsisten dan bebas dari bias interpretasi. Sebelum data dianalisis, dirancang sebuah alur metodologi sistematis yang mengintegrasikan setiap tahapan pemrosesan. Tahapan tersebut digambarkan secara konseptual melalui fase-fase berikut:



Gambar 1. Arsitektur Komputasi Dual-Stream Sensitive Ensemble Learning untuk Prediksi Risiko Dropout Mahasiswa Berbasis Atribut Sosio-Ekonomi

Sebagaimana ditunjukkan pada Gambar 1, proses eksperimen diawali dengan tahapan pemilihan dan pra-pemrosesan data, dilanjutkan dengan transformasi atribut sosio-ekonomi, penerapan algoritma ensemble learning, serta evaluasi performa model. Seluruh tahapan tersebut membentuk alur komputasi yang sistematis untuk menghasilkan model prediksi status dropout mahasiswa. Implementasi alur penelitian dilakukan menggunakan RapidMiner Studio guna memastikan setiap tahapan pemrosesan dan pengujian model dapat dijalankan secara terstruktur.

3.2 Analisis Deskriptif Data Sosio-Ekonomi Mahasiswa

Eksperimen ini memanfaatkan dataset sekunder "Predict Students Dropout and Academic Success" dari UCI Machine Learning Repository [18], [17]. Dataset riil ini pada awalnya memiliki 4.424 data sampel dengan 3 kelas pada atribut target. Setelah dilakukan data preprocessing menggunakan operator Filter Examples untuk mengeliminasi mahasiswa yang berstatus masih aktif (Enrolled), diperoleh total dataset biner sebanyak 3.630 data sampel mahasiswa yang siap diolah.

Dataset biner ini terdiri dari dua kelas tegas yang merepresentasikan status akhir mahasiswa, yaitu kelas Dropout (Putus Kuliah) sebagai kelas positif (target deteksi) dan kelas Graduate (Lulus) sebagai kelas negatif. Komposisi sebaran data akhir menunjukkan adanya kondisi ketidakseimbangan kelas (imbalanced data) yang cukup laten, di mana jumlah mahasiswa yang berhasil lulus (Graduate) mendominasi sebanyak 2.209 data sampel (sekitar 60,85%), sedangkan mahasiswa yang putus kuliah (Dropout) berjumlah 1.421 data sampel (sekitar 39,15%). Kondisi ketidakseimbangan kelas inilah yang menurut literatur data mining menjadi penyebab utama mengapa algoritma tunggal konvensional gagal melakukan klasifikasi secara sensitif pada kelas minoritas (Dropout). Ketika dihadapkan pada data minoritas, algoritma tunggal cenderung mengalami overfitting pada kelas mayoritas, sehingga pendekatan ensemble learning menjadi pilihan solusi yang paling rasional, objektif, dan teruji.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 228
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

Dari perspektif karakteristik fitur, subset data non-akademik yang mengeksplorasi kondisi sosio-ekonomi keluarga menunjukkan variasi pola yang kontras antara mahasiswa yang lulus dengan yang putus kuliah. Sebagai contoh, pada fitur Scholarship holder (status penerima beasiswa), terdapat korelasi visual di mana mayoritas mahasiswa yang berakhir dropout merupakan kelompok mahasiswa mandiri yang tidak disokong oleh bantuan finansial eksternal. Sebaliknya, kelompok mahasiswa yang berhasil menyelesaikan studinya hingga lulus didominasi oleh para pemegang beasiswa. Temuan ini selaras dengan studi makro pendidikan yang menyatakan bahwa intervensi finansial berupa beasiswa bertindak sebagai pendorong stabilitas psikologis utama bagi mahasiswa. Indikasi serupa juga ditemukan pada atribut Debtor (kepemilikan utang kuliah) dan Tuition fees up to date (kelancaran pembayaran UKT). Mahasiswa yang memiliki catatan kendala finansial akut atau menunggak biaya kuliah menempati porsi persentase yang signifikan di dalam klaster mahasiswa Dropout. Pengamatan awal ini menegaskan batasan hipotesis bahwa faktor stabilitas ekonomi keluarga bertindak sebagai fondasi krusial bagi keberlanjutan masa studi mahasiswa di perguruan tinggi.

3.3 Hasil Pengujian Pembobotan Fitur (Feature Weighting)

Guna mengidentifikasi variabel sosio-ekonomi mana saja yang memegang pengaruh paling dominan dalam memicu risiko putus kuliah, dilakukan eksperimen pembobotan menggunakan operator Weight by Information Gain [20], [7] di RapidMiner. Operator ini bekerja dengan mengukur tingkat pengurangan entropi (ketidakpastian informasi) dalam dataset ketika suatu fitur digunakan untuk memisahkan data berdasarkan label targetnya. Semakin tinggi nilai bobot Information Gain suatu atribut, maka semakin sensitif dan signifikan atribut tersebut dalam membedakan karakteristik mahasiswa yang akan bertahan dengan mahasiswa yang berpotensi putus kuliah. Hasil kalkulasi pembobotan nilai atribut sosio-ekonomi tersebut disajikan secara rinci pada Tabel 2.

Tabel 2. Hasil Pembobotan Atribut Sosio-Ekonomi Menggunakan Information Gain

Peringkat	Atribut Sosio-Ekonomi	Nilai Information Gain	Kategori Signifikansi
1	Tuition fees up to date	0,2145	Sangat Tinggi
2	Scholarship holder	0,1582	Tinggi
3	Debtor	0,0984	Sedang
4	Mother's Qualification	0,0421	Rendah
5	Father's Qualification	0,0398	Rendah
6	Mother's Occupation	0,0312	Rendah
7	Father's Occupation	0,0287	Rendah
8	Displaced	0,0154	Sangat Rendah
9	Marital Status	0,0089	Sangat Rendah

Berdasarkan data yang tersaji pada Tabel 2, ditemukan bahwa atribut Tuition fees up to date menempati peringkat pertama dengan nilai bobot tertinggi sebesar 0,2145. Temuan ini menunjukkan bahwa atribut Tuition fees up to date memiliki nilai Information Gain tertinggi dibandingkan atribut lainnya dalam dataset yang digunakan. Hasil tersebut menunjukkan bahwa status pembayaran uang kuliah memiliki relevansi informasi paling tinggi dalam membedakan mahasiswa berstatus Dropout dan Graduate pada model penelitian ini. Namun demikian, nilai Information Gain menunjukkan tingkat relevansi atribut terhadap kelas target dan tidak secara langsung menunjukkan hubungan kausal antara suatu atribut dengan kejadian dropout. Mahasiswa yang tercatat tidak mampu melunasi atau membayar uang kuliah tepat waktu pada awal-awal semester memiliki probabilitas yang jauh lebih tinggi untuk tereliminasi dari sistem akademik. Hal ini sejalan dengan regulasi operasional di sebagian besar perguruan tinggi, di mana kegagalan dalam melakukan herregistrasi finansial secara otomatis memblokir hak akses mahasiswa terhadap pengisian Kartu Rencana Studi (KRS) dan keikutsertaan dalam perkuliahan formal.

Di posisi kedua, atribut Scholarship holder mencatatkan nilai bobot sebesar 0,1582. Status kepemilikan beasiswa merupakan variabel penentu yang berfungsi sebagai jaring pengaman (safety net) bagi mahasiswa, khususnya yang berasal dari kelas ekonomi menengah ke bawah. Riset sosiologi pendidikan kontemporer menunjukkan bahwa mahasiswa penerima beasiswa memiliki tingkat kecemasan akademik yang jauh lebih rendah, karena beban pikiran mereka tidak lagi terbagi untuk memikirkan tunggakan biaya hidup. Sementara itu, variabel Debtor berada di peringkat ketiga dengan nilai 0,0984, yang semakin mempertegas dominasi faktor keuangan langsung (direct financial factors) sebagai pemicu utama fenomena dropout.

Di sisi lain, atribut latar belakang orang tua, yang meliputi pendidikan (Mother's/Father's Qualification) dan pekerjaan (Mother's/Father's Occupation), berada pada kisaran bobot yang relatif rendah (0,02 hingga 0,04). Meskipun secara teoritis tingkat pendidikan orang tua kerap diasosiasikan dengan tingkat literasi, pola asuh, dan motivasi belajar anak, dalam pemodelan prediktif ini pengaruhnya tidak bersifat instan dan langsung jika dibandingkan dengan variabel likuiditas keuangan seperti kemampuan membayar UKT. Atribut seperti Displaced (apakah mahasiswa merantau/luar kota) dan Marital Status berada di peringkat terbawah dengan kontribusi yang sangat minim. Pola ketiadaan pengaruh signifikan dari variabel geografis ini mengonfirmasi penelitian serupa yang menyatakan bahwa teknologi informasi dan sistem perkuliahan modern telah mereduksi hambatan adaptasi bagi

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 229
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

mahasiswa perantau. Temuan pembobotan ini memberikan panduan berharga bagi pihak pengambil kebijakan untuk menyusun prioritas variabel pengawasan dalam basis data kemahasiswaan.

3.4 Analisis Performa Model Klasifikasi Ensemble Learning

Setelah dataset dipisahkan menggunakan operator Split Data dengan proporsi 80% untuk data latih (2.904 sampel) dan 20% untuk data uji (726 sampel), proses pemodelan dan pengujian tiga algoritma ensemble dijalankan secara simultan. Evaluasi performa model dilakukan dengan menguji ketangguhan masing-masing metode melalui parameter matriks kekacauan (Confusion Matrix). Hasil komparasi performa dari ketiga model ensemble yang bersaing dirangkum pada Tabel 3.

Tabel 3. Perbandingan Performa Algoritma Ensemble Learning

Metrik Evaluasi Performa	Random Forest (RF)	AdaBoost (AD)	Gradient Boosted Trees (GBT)
Akurasi (Accuracy)	82,37%	79,61%	84,16%
Sensitivitas (Sensitivity / Recall)	73,94%	68,52%	77,46%
Spesifisitas (Specificity)	87,78%	86,74%	88,46%
Nilai AUC (Area Under Curve)	0,884	0,841	0,912

Berdasarkan Tabel 3, Gradient Boosted Trees (GBT) menghasilkan nilai accuracy, sensitivity, specificity, dan AUC tertinggi dibandingkan Random Forest dan AdaBoost pada skema pengujian yang digunakan. GBT memperoleh accuracy sebesar 84,16%, sensitivity sebesar 77,46%, specificity sebesar 88,46%, dan AUC sebesar 0,912. Random Forest menghasilkan accuracy sebesar 82,37%, sensitivity sebesar 73,94%, specificity sebesar 87,78%, dan AUC sebesar 0,884, sedangkan AdaBoost menghasilkan accuracy sebesar 79,61%, sensitivity sebesar 68,52%, specificity sebesar 86,74%, dan AUC sebesar 0,841.

Perbedaan nilai tersebut menunjukkan bahwa pada dataset dan konfigurasi eksperimen yang digunakan, GBT memiliki kinerja klasifikasi yang lebih baik dibandingkan Random Forest dan AdaBoost. Nilai AUC sebesar 0,912 juga menunjukkan kemampuan diskriminasi model yang baik dalam membedakan kelas Dropout dan Graduate pada data pengujian. Namun, perbedaan performa antar algoritma dalam penelitian ini tidak dapat dinyatakan sebagai perbedaan yang signifikan secara statistik karena penelitian tidak melakukan pengujian signifikansi statistik atau pengulangan eksperimen melalui cross-validation. Oleh karena itu, istilah “lebih baik” dalam penelitian ini merujuk pada nilai metrik evaluasi yang lebih tinggi pada skema pengujian yang digunakan, bukan pada bukti signifikansi statistik.

Dari sisi sensitivity, GBT mencapai nilai 77,46%, lebih tinggi dibandingkan Random Forest sebesar 73,94% dan AdaBoost sebesar 68,52%. Hasil ini menunjukkan bahwa GBT mampu mengidentifikasi proporsi kelas Dropout yang lebih besar pada data pengujian. Temuan tersebut mendukung penggunaan GBT sebagai kandidat model untuk pengembangan sistem peringatan dini [15], [7], [17], tetapi penerapan pada lingkungan perguruan tinggi tetap memerlukan validasi menggunakan data institusi yang relevan.

3.5 Analisis Dekonstruksi dan Kelemahan Algoritma Non-Unggul

Meskipun seluruh algoritma yang diuji berbasis pada metode ensemble learning, terdapat diferensiasi performa yang cukup signifikan antara Gradient Boosted Trees (84,16%) dengan Random Forest (82,37%) dan AdaBoost (79,61%). Dekonstruksi teoretis dan eksperimental terhadap kelemahan algoritma non-unggul ini penting dilakukan untuk memahami karakteristik interaksi data sosio-ekonomi kemahasiswaan dengan arsitektur komputasi multi-pohon.

Kelemahan paling mencolok terlihat pada algoritma AdaBoost yang menghasilkan tingkat akurasi terendah di bawah 80% dan sensitivitas terkecil (68,52%). Secara mekanis, AdaBoost bekerja dengan meningkatkan bobot instans secara agresif pada setiap data latih yang salah diklasifikasikan oleh weak learner sebelumnya. Namun, dataset sosio-ekonomi dari UCI Machine Learning Repository memiliki karakteristik noise yang tinggi dan variasi nilai kategori yang sangat ekstrem pada atribut latar belakang orang tua (Mother's/Father's Qualification and Occupation). Performa AdaBoost yang lebih rendah pada eksperimen ini menunjukkan bahwa model tersebut kurang sesuai dengan konfigurasi data dan parameter yang digunakan. Salah satu kemungkinan penyebabnya adalah sensitivitas mekanisme boosting terhadap kesalahan klasifikasi pada data latih, sehingga karakteristik data dan konfigurasi hyperparameter dapat memengaruhi kemampuan generalisasi model [12], [7]. Akibatnya, batas keputusan (decision boundary) yang dihasilkan menjadi terlalu kompleks dan gagal melakukan generalisasi dengan baik saat diuji pada data baru (testing pool).

Di sisi lain, Random Forest menunjukkan performa yang jauh lebih stabil (AUC 0,884) dibandingkan AdaBoost, namun tetap belum mampu mengimbangi keunggulan Gradient Boosted Trees. Hal ini terjadi karena Random Forest berbasis pada metode bagging (pembentukan pohon secara paralel dari sampel bootstrap) [14], [19]. Karakteristik bagging sangat bias terhadap atribut yang memiliki tingkat kardinalitas tinggi. Dalam dataset ini, atribut pekerjaan orang tua memiliki banyak nilai kategori unik, sehingga Random Forest cenderung memilih

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 230
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

atribut tersebut secara berulang pada simpul-simpul awalnya, meskipun berdasarkan uji Information Gain (Tabel 2), atribut likuiditas keuangan seperti Tuition fees up to date bernilai lebih baik.

Kelemahan struktural dari kedua algoritma di atas berhasil diatasi secara sempurna oleh Gradient Boosted Trees. GBT tidak memanipulasi bobot data secara acak seperti AdaBoost, melainkan mengoptimalkan bias model secara gradual dengan meminimalkan nilai pseudo-residuals menggunakan pendekatan gradient descent. Strategi penambahan pohon baru yang murni didasarkan pada arah gradien negatif dari fungsi kerugian klasifikasi membuat GBT sangat efisien dalam menangani data kategori sosio-ekonomi yang timpang tanpa mengalami bias terhadap atribut berkardinalitas tinggi. Rangkuman perbandingan karakteristik teknis hasil eksperimen ini disajikan secara komprehensif pada Tabel 4.

Tabel 4. Rangkuman Karakteristik dan Sifat Komputasi Algoritma Ensemble Berdasarkan Hasil Eksperimen

Parameter Karakteristik	Random Forest (RF)	AdaBoost (AD)	Gradient Boosted Trees (GBT)
Paradigma Ensemble	Parallel Bagging	Sequential Boosting	Sequential Gradient Boosting
Kerentanan Noise	Rendah (Robust)	Sangat Tinggi (Sensitive)	Rendah-Sedang (Regularized)
Metode Optimasi	Majority Voting	Exponential Weighting	Gradient Descent Optimization
Efisiensi Data Tipang	Sedang	Rendah	Sangat Tinggi
Risiko Overfitting	Sangat Rendah	Tinggi	Dikontrol oleh Early Stopping

Berdasarkan Tabel 4, masing-masing algoritma ensemble learning memiliki karakteristik komputasi yang berbeda. Random Forest menerapkan paradigma parallel bagging dengan mekanisme majority voting, sedangkan AdaBoost menggunakan sequential boosting melalui pendekatan exponential weighting yang memiliki sensitivitas lebih tinggi terhadap noise. Sementara itu, Gradient Boosted Trees menerapkan sequential gradient boosting dengan mekanisme optimasi gradient descent dan early stopping untuk mengendalikan risiko overfitting. Perbedaan karakteristik tersebut memberikan gambaran mengenai mekanisme kerja masing-masing algoritma dalam memproses data sosio-ekonomi mahasiswa. Dalam eksperimen ini, GBT menghasilkan performa klasifikasi tertinggi sebagaimana ditunjukkan pada Tabel 3, sehingga menjadi kandidat model yang paling potensial untuk dikembangkan lebih lanjut sebagai dasar sistem peringatan dini risiko dropout mahasiswa.

3.6 Implikasi Manajerial dan Desain Sistem Peringatan Dini

Keberhasilan pemodelan berbasis Gradient Boosted Trees di platform RapidMiner ini memberikan kontribusi teoretis dan praktis yang signifikan bagi tata kelola perguruan tinggi. Secara teoretis, hasil riset ini berhasil menjawab tantangan keterbatasan (research gap) dari penelitian-penelitian terdahulu yang sering kali mengabaikan faktor ekonomi dan bersandar penuh pada variabel nilai akademik mahasiswa. Riset ini membuktikan secara ilmiah bahwa integrasi data sosio-ekonomi keluarga yang dikombinasikan dengan kecerdasan komputasi multi-pohon (ensemble) mampu menghasilkan akurasi prediksi di atas 84%, yang menyiratkan bahwa pola kerentanan mahasiswa sejatinya sudah dapat dipetakan bahkan sebelum proses perkuliahan semester pertama dimulai.

Secara praktis, model prediktif GBT dengan nilai AUC sebesar 0,912 menunjukkan potensi untuk dijadikan dasar pengembangan modul Early Warning System (EWS) pada Sistem Informasi Akademik (SIAKAD) perguruan tinggi [11], [16], [19]. Pemanfaatan model tersebut dapat diarahkan untuk membantu Biro Administrasi Akademik dan Kemahasiswaan (BAAK) dalam melakukan pemantauan mahasiswa berdasarkan karakteristik yang teridentifikasi memiliki relevansi tinggi terhadap status dropout. Implementasi EWS pada lingkungan perguruan tinggi tetap memerlukan proses pengujian dan validasi menggunakan data internal institusi sebelum diterapkan secara operasional. Di sinilah letak transformasi fundamental dari sistem pengawasan yang awalnya bersifat reaktif menjadi proaktif-preventif. Mekanisme operasional EWS berbasis sosio-ekonomi ini dapat dijabarkan ke dalam tiga tahapan implementasi strategis:

a. Fase Input dan Skoring Awal (Awal Semester): Saat mahasiswa baru melakukan registrasi ulang dan mengisi formulir profil sosio-ekonomi (pendapatan, beasiswa, status daerah, pendidikan orang tua), data tersebut secara otomatis akan dialirkan ke dalam basis data sistem EWS. Model GBT yang tertanam dalam sistem akan melakukan kalkulasi skoring secara real-time terhadap profil tersebut.
b. Fase Kategorisasi Risiko: Mahasiswa yang memiliki kombinasi atribut rentan—seperti status non-beasiswa, orang tua berpendapatan rendah, serta memiliki riwayat penundaan pembayaran di sistem keuangan—secara otomatis akan ditandai (flagging) oleh sistem EWS ke dalam kategori "Risiko Tinggi Putus Kuliah". Daftar mahasiswa ini akan langsung tersinkronisasi di dasbor pemantauan milik pejabat BAAK dan dosen wali masing-masing.
c. Fase Intervensi Tepat Sasaran: Berbekal data prediktif tersebut, universitas dapat mengeksekusi tindakan preventif sebelum mahasiswa mengalami frustrasi akademik atau keuangan. BAAK dapat langsung merekomendasikan nama-nama mahasiswa di dasbor EWS tersebut untuk diprioritaskan menerima alokasi beasiswa internal kampus, bantuan dana stimulus, atau skema cicilan UKT yang lebih fleksibel. Di saat yang

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 231
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

sama, dosen wali dapat melakukan pemanggilan konseling secara personal untuk memberikan motivasi dan bimbingan psikologis, sehingga potensi mutasi atau drop out dapat dicegah sejak dini.

Penerapan sistem cerdas ini tidak hanya akan membantu universitas dalam mempertahankan tingkat retensi mahasiswa yang sehat, tetapi juga berdampak langsung pada optimalisasi pemanfaatan sumber dana bantuan finansial kampus agar benar-benar jatuh ke tangan mahasiswa yang paling membutuhkan. Secara makro, keberhasilan mitigasi ini akan menjaga stabilitas jumlah mahasiswa aktif per angkatan, yang merupakan salah satu variabel vital dalam penilaian borang akreditasi institusi oleh Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) maupun Lembaga Akreditasi Mandiri. Dengan demikian, integrasi ilmu penambangan data (data mining) dan kebijakan manajemen pendidikan tinggi ini terbukti mampu menjadi solusi strategis yang berkelanjutan dalam menekan angka putus kuliah di era modern.

3.7 Pembahasan

Hasil penelitian ini menunjukkan bahwa pendekatan ensemble learning memiliki potensi yang baik dalam memprediksi status putus kuliah mahasiswa berdasarkan atribut sosio-ekonomi. Temuan tersebut sejalan dengan penelitian Lenin dan Chandrasekaran [12] yang menerapkan algoritma Random Forest, AdaBoost, dan Stochastic Gradient Boosting untuk memprediksi performa mahasiswa pada dataset pendidikan yang memiliki distribusi kelas tidak seimbang. Penelitian tersebut menunjukkan bahwa pendekatan ensemble mampu menghasilkan kinerja klasifikasi yang baik dibandingkan metode klasifikasi tunggal. Kesamaan kedua penelitian terletak pada pemanfaatan algoritma ensemble untuk mengidentifikasi pola dalam data mahasiswa. Namun, penelitian ini memiliki fokus yang lebih spesifik, yaitu membandingkan Random Forest, AdaBoost, dan Gradient Boosted Trees dalam mengklasifikasikan status Dropout dan Graduate dengan menggunakan atribut sosio-ekonomi sebagai prediktor utama.

Dari aspek pemilihan atribut, penelitian ini menemukan bahwa Tuition fees up to date memiliki nilai Information Gain tertinggi sebesar 0,2145, diikuti Scholarship holder sebesar 0,1582 dan Debtor sebesar 0,0984. Temuan ini memperlihatkan bahwa atribut yang berkaitan langsung dengan kondisi pembayaran dan dukungan finansial memiliki relevansi informasi lebih tinggi dibandingkan atribut demografis maupun latar belakang pendidikan dan pekerjaan orang tua. Hasil tersebut memiliki kesesuaian dengan penelitian Gutierrez-Pachas et al. [6] yang menganalisis faktor akademik, sosio-ekonomi, dan kesetaraan dalam prediksi dropout mahasiswa pada konteks perguruan tinggi di Amerika Latin. Meskipun penelitian tersebut menemukan bahwa variabel akademik memiliki peran yang lebih dominan dalam prediksi dropout dibandingkan sebagian variabel sosio-ekonomi dan kesetaraan, kedua penelitian sama-sama menunjukkan pentingnya mempertimbangkan karakteristik mahasiswa secara multidimensional. Perbedaan hasil tersebut dapat dipengaruhi oleh karakteristik dataset, cakupan variabel, dan konteks institusi yang digunakan. Dengan demikian, penelitian ini memberikan perspektif tambahan bahwa ketika pemodelan difokuskan pada atribut sosio-ekonomi, informasi mengenai status pembayaran biaya pendidikan dan kepemilikan beasiswa dapat menjadi atribut yang relevan dalam membedakan mahasiswa berstatus Dropout dan Graduate.

Temuan mengenai relevansi faktor finansial juga memiliki keterkaitan dengan penelitian Mustofa et al. [7] yang mengembangkan model prediksi risiko dropout berbasis kecerdasan buatan dan explainable artificial intelligence. Penelitian tersebut menggunakan dataset Predict Students Dropout and Academic Success dari UCI Machine Learning Repository dan mengidentifikasi sejumlah faktor yang berkaitan dengan risiko dropout, termasuk status utang, status beasiswa, serta pembayaran biaya pendidikan. Kesamaan sumber dataset dan ditemukannya atribut finansial sebagai faktor yang relevan memperkuat keterkaitan hasil penelitian ini dengan penelitian terdahulu. Meskipun demikian, terdapat perbedaan dalam pendekatan pengolahan data. Mustofa et al. menggabungkan kelas Enrolled dan Graduate menjadi kategori Non-Dropout, sedangkan penelitian ini mengeluarkan kelas Enrolled sehingga proses klasifikasi hanya membandingkan kelas Dropout dan Graduate. Selain itu, penelitian Mustofa et al. mengembangkan model hybrid Logistic Regression dan Artificial Neural Network serta menggunakan SHAP dan LIME untuk interpretasi model, sedangkan penelitian ini berfokus pada perbandingan tiga algoritma ensemble learning dan pembobotan atribut menggunakan Information Gain. Oleh karena itu, hasil kedua penelitian tidak dapat dibandingkan secara langsung berdasarkan nilai akurasi karena terdapat perbedaan formulasi kelas target dan konfigurasi eksperimen.

Berdasarkan aspek performa algoritma, Gradient Boosted Trees (GBT) menghasilkan kinerja terbaik dalam penelitian ini dengan accuracy sebesar 84,16%, sensitivity sebesar 77,46%, specificity sebesar 88,46%, dan AUC sebesar 0,912. Hasil tersebut menunjukkan bahwa GBT memiliki kemampuan diskriminasi yang baik dalam membedakan mahasiswa berstatus Dropout dan Graduate pada dataset yang digunakan. Temuan ini sejalan dengan penelitian Niyogisubizo et al. [15] yang menerapkan pendekatan two-layer ensemble machine learning dengan stacked generalization untuk memprediksi dropout mahasiswa. Kedua penelitian menunjukkan bahwa pendekatan ensemble dapat dimanfaatkan untuk meningkatkan kemampuan pemodelan prediktif dalam konteks pendidikan tinggi. Akan tetapi, arsitektur model yang digunakan berbeda. Penelitian Niyogisubizo et al. menerapkan kombinasi model dalam pendekatan stacked generalization, sedangkan penelitian ini membandingkan performa masing-masing algoritma ensemble secara terpisah untuk menentukan model dengan hasil evaluasi terbaik. Dengan demikian, kontribusi penelitian ini terletak pada penyediaan bukti empiris mengenai kinerja GBT dibandingkan Random Forest dan AdaBoost pada penggunaan atribut sosio-ekonomi.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 232
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

Hasil penelitian ini juga dapat dibandingkan dengan penelitian Vaarma dan Li [21] yang melakukan prediksi dropout mahasiswa pada perguruan tinggi di Finlandia menggunakan data transkrip akademik, demografis, dan aktivitas Learning Management System (LMS). Penelitian tersebut menunjukkan bahwa akumulasi kredit, mata kuliah yang tidak lulus, serta aktivitas mahasiswa pada Moodle merupakan prediktor penting dalam mengidentifikasi risiko dropout. Selain itu, tingkat kepentingan atribut dapat berubah sesuai dengan waktu pengamatan selama masa studi. Berbeda dengan penelitian tersebut, penelitian ini secara khusus menitikberatkan penggunaan atribut sosio-ekonomi tanpa mengintegrasikan indikator akademik dan aktivitas LMS. Perbedaan tersebut menunjukkan bahwa identifikasi risiko dropout dapat dilakukan melalui berbagai kelompok atribut, baik akademik maupun non-akademik. Penggabungan kedua kelompok informasi tersebut berpotensi menjadi arah pengembangan penelitian berikutnya agar sistem prediksi mampu memberikan gambaran risiko mahasiswa secara lebih komprehensif.

Dalam konteks pengembangan sistem peringatan dini, hasil penelitian ini memiliki keterkaitan dengan penelitian Lee dan Chung [11] yang mengembangkan sistem peringatan dini berbasis machine learning untuk meningkatkan kinerja prediksi dropout mahasiswa. Penelitian tersebut menegaskan pentingnya pemanfaatan model prediktif untuk mendukung identifikasi mahasiswa yang berpotensi mengalami dropout sebelum permasalahan berkembang lebih lanjut. Sejalan dengan hal tersebut, hasil penelitian ini menempatkan GBT sebagai kandidat model yang dapat dipertimbangkan dalam pengembangan Early Warning System (EWS) pada Sistem Informasi Akademik (SIAKAD). Namun, model yang dihasilkan dalam penelitian ini masih berada pada tahap eksperimen dan belum diuji secara langsung pada data operasional perguruan tinggi. Oleh karena itu, penerapan lebih lanjut memerlukan validasi menggunakan data internal institusi, pengujian secara berkala, serta keterlibatan pihak akademik dalam menindaklanjuti hasil prediksi.

Secara keseluruhan, perbandingan dengan penelitian terdahulu memperlihatkan bahwa penelitian ini memiliki kesamaan dalam pemanfaatan machine learning untuk mendukung prediksi dropout mahasiswa, tetapi menawarkan fokus yang lebih spesifik pada penggunaan atribut sosio-ekonomi dan komparasi tiga algoritma ensemble learning. Temuan mengenai dominasi relevansi atribut pembayaran biaya pendidikan, kepemilikan beasiswa, dan status utang memberikan informasi tambahan bagi pengelola perguruan tinggi dalam memahami karakteristik non-akademik yang berkaitan dengan status keberlanjutan studi mahasiswa. Sementara itu, capaian GBT dengan AUC sebesar 0,912 menunjukkan potensi model tersebut untuk dikembangkan lebih lanjut. Meskipun demikian, hasil ini perlu diinterpretasikan sesuai batasan dataset dan konfigurasi eksperimen yang digunakan. Pengujian dengan stratified cross-validation, data longitudinal, serta dataset dari perguruan tinggi yang berbeda masih diperlukan untuk menilai konsistensi dan kemampuan generalisasi model sebelum diterapkan sebagai sistem pendukung keputusan secara operasional.

4. KESIMPULAN

Berdasarkan hasil eksperimen menggunakan algoritma Random Forest, AdaBoost, dan Gradient Boosted Trees (GBT), penelitian ini menunjukkan bahwa atribut finansial memiliki relevansi paling dominan dalam memprediksi status putus kuliah mahasiswa, dengan Tuition fees up to date memperoleh nilai Information Gain tertinggi sebesar 0,2145, diikuti Scholarship holder sebesar 0,1582 dan Debtor sebesar 0,0984. Dari ketiga algoritma yang diuji, GBT menunjukkan kinerja terbaik dengan accuracy 84,16%, sensitivity 77,46%, specificity 88,46%, dan AUC 0,912, sehingga memiliki kemampuan diskriminasi yang lebih baik dalam mengidentifikasi mahasiswa berisiko dropout pada dataset yang digunakan. Temuan ini menegaskan potensi pemanfaatan model GBT sebagai dasar pengembangan Early Warning System yang terintegrasi dengan Sistem Informasi Akademik (SIAKAD) untuk mendukung identifikasi risiko sejak awal semester, penentuan prioritas bantuan finansial, serta intervensi preventif melalui pendampingan akademik dan konseling mahasiswa. Meskipun demikian, hasil penelitian masih memiliki keterbatasan karena menggunakan dataset sekunder UCI, satu skema pembagian data 80:20, dan belum menerapkan cross-validation maupun validasi eksternal. Oleh karena itu, penelitian selanjutnya perlu menguji model menggunakan data aktual perguruan tinggi yang berbeda, menerapkan stratified cross-validation, serta mengevaluasi aspek fairness dan pengawasan manusia guna memastikan keandalan model sebelum diimplementasikan secara operasional.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi	: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Metodologi	: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Perangkat Lunak	: Sugeng Hendra Wijaya;
Validasi	: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Analisis Formal	: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Investigasi	: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Sumber Daya	: Sugeng Hendra Wijaya;

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 233
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

Kurasi Data : Sugeng Hendra Wijaya;
Penulisan Draf Awal : Sugeng Hendra Wijaya;
Penulisan, Peninjauan, dan Penyuntingan: Sugeng Hendra Wijaya dan Novi Andayaningtyas;
Visualisasi : Sugeng Hendra Wijaya.
Seluruh penulis telah membaca dan menyetujui versi akhir manuskrip.

5.2 Ketersediaan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang tersedia secara publik melalui UCI Machine Learning Repository dengan judul Predict Students' Dropout and Academic Success. Dataset dapat diakses melalui tautan https://doi.org/10.24432/C5MC89.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan khusus dari lembaga pendanaan publik, komersial, maupun nirlaba.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable). Penelitian ini menggunakan dataset sekunder yang tersedia secara publik melalui UCI Machine Learning Repository dan tidak melibatkan pengumpulan data primer secara langsung dari partisipan oleh penulis.

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable). Penelitian ini menggunakan dataset sekunder yang tersedia secara publik dan tidak melibatkan perekrutan atau interaksi langsung dengan partisipan penelitian.

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan teknologi AI generatif sebagai alat bantu dalam proses penyusunan dan penyempurnaan kejelasan bahasa manuskrip. Seluruh konten yang dihasilkan dengan bantuan AI telah ditinjau dan disunting oleh penulis. Para penulis bertanggung jawab penuh atas keakuratan, integritas, dan versi akhir manuskrip yang dipublikasikan.

REFERENCES

[1] F. Agrusti, G. Bonavolontà, and M. Mezzini, “University dropout prediction through educational data mining techniques: A systematic review,” J. E-Learning Knowl. Soc., vol. 15, no. 3, pp. 161–182, 2019, doi: 10.20368/1971-8829/1135017.
[2] M. Alban and D. Mauricio, “Predicting University Dropout trough Data Mining: A systematic Literature,” Indian J. Sci. Technol., vol. 12, no. 4, pp. 1–12, 2019, doi: 10.17485/ijst/2019/v12i4/139729.
[3] M. P. Colpo, T. Thompsen Primo, M. S. de Aguiar, and C. Cechinel, Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges, vol. 32. 2024. doi: 10.5753/rbie.2024.3559.
[4] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” Int. J. Educ. Technol. High. Educ., vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.
[5] M. Segura, J. Mello, and A. Hernández, “Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?,” Mathematics, vol. 10, no. 18, pp. 1–20, 2022, doi: 10.3390/math10183359.
[6] D. A. Gutierrez-Pachas, G. Garcia-Zanabria, E. Cuadros-Vargas, G. Camara-Chavez, and E. Gomez-Nieto, “Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context,” Educ. Sci., vol. 13, no. 2, 2023, doi: 10.3390/educsci13020154.
[7] S. Mustofa, Y. R. Emon, S. Bin Mamun, S. A. Akhy, and M. T. Ahad, “A novel AI-driven model for student dropout risk analysis with explainable AI insights,” Comput. Educ. Artif. Intell., vol. 8, no. August 2024, p. 100352, 2025, doi: 10.1016/j.caeai.2024.100352.
[8] W. M. Aldowah, H., Al-Samarraie, H., & Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” Telemat. Informatics, vol. 37, pp. 13–49, 2019, doi: 10.1016/j.tele.2019.01.007.
[9] S. Alturki, I. Hulpuș, and H. Stuckenschmidt, Predicting Academic Outcomes: A Survey from 2007 Till 2018, vol. 27, no. 1. Springer Netherlands, 2022. doi: 10.1007/s10758-020-09476-0.
[10] O. W. Adejo, “Predicting student academic performance using multi-model heterogeneous ensemble approach,” J. Appl. Res. High. Educ., vol. 00, no. 00, pp. 61–75, 2018, doi: 10.1108/JARHE-09-2017-0113.
[11] S. Lee and J. Y. Chung, “The machine learning-based dropout early warning system for improving the performance of dropout prediction,” Appl. Sci., vol. 9, no. 15, 2019, doi: 10.3390/app9153093.
[12] T. Lenin and N. Chandrasekaran, “Learning from Imbalanced Educational Data Using Ensemble Machine Learning Algorithms,” Webology, vol. 18, no. SpecialIssue1, pp. 183–195, 2021, doi: 10.14704/WEB/V18SI01/WEB18053.
[13] A. Jain, A. K. Dubey, S. Khan, A. Panwar, M. Alkhatib, and A. M. Alshahrani, “A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems,” Sci. Rep., vol. 15, no. 1, pp. 1–28, 2025, doi: 10.1038/s41598-025-97506-1.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 234
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 223−235
DOI 10.47065/josh.v8i1.11259
https://ejurnal.seminar-id.com/index.php/josh/

[14] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," Appl. Sci., vol. 11, no. 7, 2021, doi: 10.3390/app11073130.
[15] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," Comput. Educ. Artif. Intell., vol. 3, no. March, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.
[16] K. C. Minh Phan, Arno De Caigny, "A decision support framework to incorporate textual data for early student dropout prediction in higher education," Decis. Support Syst., vol. 168, 2023, doi: 10.1016/j.dss.2023.113940.
[17] E. Guamán-Hidalgo and L. Enciso, "Early Prediction of Student Dropout Through Machine Learning in Educational Data Mining: A Comparative Analysis of Algorithms and Data Balancing Techniques," Appl. Sci., vol. 16, no. 18, p. 8915, 2026, doi: 10.3390/app16188915.
[18] J. M. M. S. Martins, M. V., Pesquisa, D., & Real, "Predict Students Dropout and Academic Success." UCI Machine Learning Repository, 2022. doi: 10.24432/C5MC89.
[19] V. Dermol and V. Skrbinjek, "Predicting student dropout in higher education using Random Forest and survival analysis," Discov. Educ., vol. 5, no. 1, 2026, doi: 10.1007/s44217-026-01710-8.
[20] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," IEEE Access, vol. 9, pp. 33132–33143, 2021, doi: 10.1109/ACCESS.2021.3061368.
[21] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," Technol. Soc., vol. 76, no. February, p. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 235
This work is licensed under a Creative Commons Attribution 4.0 International License.