Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 5, Issue: 2, Month: Juni, Year: 2026
Page: 602 - 608, Available online at: https://rcf-indonesia.org/home/

ANALISIS FAKTOR DOMINAN PREDIKSI DROPOUT MAHASISWA MENGGUNAKAN RANDOM FOREST, XGBOOST, DAN XAI

Norma Devi Kurniasari[1], Trian Basofi Rohman[2], Ari Widianto[3], Anis Shobikah[4], Gaguk Triono[5]
[1,2,3,4]Institut Teknologi Insan Cendekia Mandiri,
[5]Universitas Maarif Hasyim Latif
Corresponding Author: [1] normadevi@iticm.ac.id

Article Info

Article history:

Received: Jun 15, 2026
Revised: Jun 17, 2026
Accepted: Jun 24, 2026
Published: Jun 26, 2026

Keywords:

Dropout Mahasiswa
Random Forest
XGBoost
XAI
SHAP

ABSTRAK

Masalah mahasiswa yang putus kuliah (*dropout*) menjadi tantangan signifikan bagi perguruan tinggi karena memengaruhi reputasi institusi dan efisiensi pemanfaatan sumber daya kampus. Untuk mengatasi fenomena tersebut, diperlukan sistem prediksi dini guna mengidentifikasi mahasiswa yang berisiko sejak awal perkuliahan. Penelitian ini bertujuan membandingkan kinerja dua algoritma machine learning, yaitu *Random Forest* dan *XGBoost*, sekaligus memetakan faktor utama penyebab dropout menggunakan pendekatan *Explainable AI* (XAI) melalui metode SHAP (*SHapley Additive exPlanations*). Studi ini menggunakan dataset sekunder dengan total 4.424 record data mahasiswa yang mencakup variabel demografi, sosial-ekonomi, makroekonomi, dan akademik. Proses eksperimen dan pemodelan dilakukan menggunakan lingkungan *Google Colab*. Hasil pengujian menunjukkan bahwa model Random Forest menghasilkan tingkat akurasi optimal yang lebih tinggi yaitu sebesar 77,40%, dibandingkan model *XGBoost* yang menghasilkan akurasi sebesar 76,05%. Melalui analisis interpretasi SHAP, penelitian ini menemukan bahwa jumlah mata kuliah yang lulus di semester dua *(Curricular units 2nd sem (approved))* dan status kelancaran pembayaran biaya kuliah *(Tuition fees up to date)* merupakan faktor paling dominan yang memengaruhi keputusan prediksi. Hasil penelitian ini memberikan dasar empiris bagi pengelola perguruan tinggi untuk memprioritaskan kebijakan intervensi pada aspek akademik tahun kedua serta stabilitas finansial mahasiswa sebagai strategi menekan angka putus kuliah

1. INTRODUCTION

Masalah mahasiswa yang putus kuliah (dropout) sampai saat ini masih menjadi salah satu tantangan terbesar bagi perguruan tinggi. Masalah ini berdampak langsung pada operasional dan kondisi keuangan kampus, serta merugikan mahasiswa secara pribadi karena kehilangan kesempatan menyelesaikan studi mereka [1]. Selain itu, angka putus kuliah yang tinggi juga bisa menurunkan nilai akreditasi institusi karena buruknya tingkat retensi mahasiswa [2]. Oleh karena itu, pihak manajemen kampus tidak bisa lagi memakai cara lama yang pasif atau sekadar menunggu. Kampus sangat membutuhkan sistem peringatan dini yang bisa mendeteksi mahasiswa yang berisiko sejak tahun pertama agar bantuan atau intervensi bisa diberikan secepatnya. Dengan memanfaatkan data historis seperti latar belakang demografi, kondisi ekonomi, dan nilai akademik, penggunaan model algoritma cerdas dapat menjadi solusi preventif yang nyata untuk menjaga mahasiswa agar tetap kuliah sampai lulus.

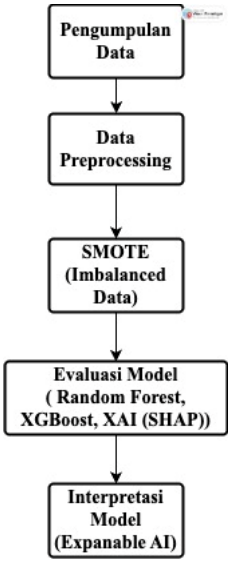
Penelitian sebelumnya sebenarnya sudah banyak yang mencoba membuat model prediksi untuk menekan angka dropout ini. Salah satu penelitian yang dilakukan oleh [3] membuktikan bahwa performa akademik atau nilai di awal semester kuliah merupakan indikator paling kuat untuk memprediksi apakah seorang mahasiswa akan terus lanjut atau berhenti kuliah. Melanjutkan fokus yang serupa, studi yang berjalan pada [4] menemukan bahwa faktor finansial orang tua dan ketepatan waktu membayar uang kuliah di awal semester juga punya pengaruh yang sangat besar pada keputusan mahasiswa untuk bertahan. Sementara itu, eksperimen yang tertuang dalam [5] mencoba menggunakan algoritma berbasis ensemble seperti Random Forest untuk mengatasi data dunia pendidikan yang biasanya tidak seimbang, dan terbukti mampu menghasilkan akurasi yang stabil.

Variabel yang diteliti pun terus berkembang dalam beberapa tahun terakhir. Contohnya seperti riset yang dipublikasikan oleh [6] yang memasukkan data keaktifan mahasiswa di platform pembelajaran digital (e-learning) untuk memperkuat hasil prediksi model. Di sisi lain, laporan ilmiah pada [7] lebih memilih algoritma XGBoost karena dinilai lebih cepat dalam mengoptimalkan pohon keputusan dan efisien saat memproses data. Meskipun akurasinya tinggi, temuan penting yang dipaparkan dalam [8] mengingatkan bahwa model-model kompleks ini sayangnya bersifat black-box atau sulit dipahami alur logikanya. Oleh karena itu, literatur tersebut menyarankan penggunaan Explainable AI (XAI) agar hasil prediksi model memiliki alasan logis yang bisa dipahami dengan mudah oleh para pengambil kebijakan di kampus.

Meskipun model-model dalam penelitian terdahulu punya akurasi yang tinggi, kebanyakan riset tersebut mengabaikan aspek kejelasan atau transparansi model. Bagi pihak kampus, sekadar tahu daftar nama mahasiswa yang berpotensi dropout tentu belum cukup tanpa tahu apa penyebab utamanya. Penelitian ini dibuat untuk mengisi celah tersebut dengan membandingkan kinerja algoritma Random Forest dan XGBoost, lalu membedah hasil prediksinya menggunakan pendekatan XAI dengan metode SHAP (SHapley Additive exPlanations). Berbeda dengan riset sebelumnya yang biasanya hanya fokus pada satu aspek, penelitian ini menganalisis variabel demografi, ekonomi, dan akademik secara bersamaan menggunakan 4.424 record data riil mahasiswa dari Kaggle. Selain itu, agar kampus tidak perlu membeli perangkat komputer yang mahal, seluruh proses eksperimen dan pembuatan aplikasi analisisnya dikerjakan menggunakan platform berbasis cloud Google Colab. Kombinasi metode ini diharapkan bisa memberikan masukan yang jelas bagi manajemen kampus dalam membuat kebijakan bantuan keuangan maupun pendampingan akademik yang tepat sasaran.

2. MATERIALS AND METHODS

Guna memberikan gambaran yang jelas mengenai metodologi data science yang diterapkan, proses eksperimen dijalankan secara sekuensial melalui beberapa fase krusial. Alur linear ini sengaja dirancang untuk memastikan penanganan masalah kualitas data dan ketidakseimbangan kelas dilakukan dengan tepat sebelum masuk ke tahap pengujian akhir. Kerangka kerja dan alur logis dari metode penelitian tersebut diilustrasikan pada Gambar 1.



Gambar 1. Metode Penelitian

2.1. Pengumpulan Data

Proses pengumpulan data dalam riset ini memanfaatkan dataset sekunder dari repositori Kaggle yang berjudul "Higher Education Predictors of Student Retention", sebuah dataset publikasi riil yang awalnya dibangun oleh Martins dkk. untuk mengevaluasi faktor retensi akademik [9]. Dataset ini memuat 4.424 record data mahasiswa yang tersebar di berbagai institusi pendidikan tinggi dengan 36 atribut prediktor, yang terbagi secara terstruktur ke dalam data demografi, sosial-ekonomi, makroekonomi, serta performa akademik mahasiswa pada akhir semester pertama dan kedua. Target klasifikasi dari dataset ini merekam status akhir kelangsungan studi mahasiswa yang terbagi menjadi tiga kelas, yaitu Dropout (putus studi), Enrolled (mahasiswa aktif), dan Graduate (lulus).

2.2. Data Preprocessing

Kualitas hasil prediksi sangat bergantung pada kebersihan data sebelum masuk ke tahap pelatihan model. Prapemrosesan data yang sistematis pada data tabular sangat penting untuk mengatasi data yang tidak lengkap, tidak konsisten, mengandung noise, maupun nilai yang hilang sehingga dapat meningkatkan kinerja algoritma klasifikasi [10]. Dalam tahapan ini, dilakukan pembersihan nama fitur menggunakan *str.strip()* untuk menghapus spasi tersembunyi, transformasi label kategorikal ke numerik melalui *LabelEncoder*, serta pengisian data kosong menggunakan *SimpleImputer* dengan strategi median. Penggunaan median dipilih karena kemampuannya dalam mempertahankan representasi data tanpa terlalu dipengaruhi oleh nilai ekstrem (*outliers*).

2.3. Penanganan Data Tidak Seimbang (SMOTE - *Imbalanced Data*)

Ketidakseimbangan kelas sering kali menjadi hambatan dalam riset pendidikan, di mana jumlah mahasiswa yang lulus biasanya jauh lebih banyak daripada mahasiswa yang aktif atau putus kuliah. Penggunaan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) terbukti mampu meningkatkan representasi kelas minoritas melalui pembentukan sampel sintetis, sehingga sensitivitas model terhadap kelas minoritas dapat meningkat sekaligus mengurangi kecenderungan overfitting yang umum terjadi pada teknik *random oversampling* [11].

2.4. Evaluasi Model (*Model Evaluation*)

Evaluasi model dilakukan untuk mengukur sejauh mana algoritma cerdas mampu mengenali pola perilaku mahasiswa berdasarkan data historis. Penggunaan metrik evaluasi yang komprehensif seperti accuracy, precision, recall, dan F1-score, serta penerapan teknik validasi silang (*cross validation*), merupakan praktik yang direkomendasikan untuk memperoleh estimasi performa model yang lebih andal dan mengurangi bias evaluasi[12] .

2.5. Interpretasi Model (Explainable AI - XAI)

Transparansi model menjadi kebutuhan penting agar hasil prediksi machine learning dapat diterima dan dimanfaatkan secara efektif oleh pihak manajemen kampus. Pendekatan *Explainable Artificial Intelligence* (XAI) melalui metode SHAP (*SHapley Additive exPlanations*) mampu menjelaskan kontribusi setiap variabel terhadap hasil prediksi sehingga model yang bersifat *black-box* dapat diinterpretasikan menjadi informasi yang lebih transparan dan mudah dipahami oleh pengambil keputusan [13].

3. RESULTS AND DISCUSSION

3.1. Pengumpulan Data

Dataset dalam penelitian ini diambil dari situs *kaggle* yang tersusun atas 36 atribut prediktor yang secara sistematis dikelompokkan ke dalam beberapa klaster fitur utama. Pembagian kelompok ini bertujuan untuk memudahkan pemetaan dampak variabel dari lingkungan eksternal maupun internal mahasiswa terhadap kelangsungan akademiknya. Secara rinci, klasifikasi variabel, nama atribut asli dalam dataset, beserta tipe data yang diolah disajikan pada Tabel 1.

Tabel 1. Identifikasi Atribut dan Tipe Data

Class of Attribute	Attribute	Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete

Class of Attribute	Attribute	Type
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Target	Target	Categorical

3.2. Data Preprocessing

Tahap prapemrosesan data merupakan fase krusial dalam pipa rekayasa data (*data pipeline*) untuk mentransformasikan data mentah agar memenuhi standar kualitas pemodelan machine learning melalui penanganan kualitas data (*cleaning*), kelengkapan data (*imputation*), hingga penanganan ketidakseimbangan kelas (*imbalance handling*). Langkah pertama dimulai dengan pembersihan fitur (*feature cleaning*) menggunakan fungsi str.strip untuk menghapus karakter spasi tersembunyi pada nama kolom guna mencegah kegagalan eksekusi indeks fungsi selama komputasi berlangsung. Langkah kedua adalah menerapkan LabelEncoder untuk mentransformasikan label kategorikal tekstual pada variabel dependen (*Target*) menjadi representasi angka diskret (0, 1, dan 2) agar dapat diproses secara matematis oleh algoritma XGBoost dan Random Forest. Langkah ketiga berfokus pada penanganan kelengkapan data melalui imputasi nilai kosong (*missing values*) menggunakan fungsi SimpleImputer dengan strategi nilai tengah (*median*), yang sengaja dipilih karena sifatnya yang jauh lebih kebal terhadap pengaruh nilai ekstrem (*outliers*) dibandingkan dengan nilai rata-rata (*mean*). Langkah keempat adalah melakukan partisi data via fungsi *train_test_split* dengan proporsi 80% dialokasikan sebagai data latih (*training set*) dan 20% sebagai data uji (*testing set*), di mana parameter stratify=y diaktifkan secara ketat untuk mengunci persentase sebaran kelas target agar tetap sama rata antara porsi pelatihan dan pengujian. Langkah terakhir dijalankan menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan kelas target; berdasarkan log output data, kelas "*Enrolled*" yang awalnya hanya memiliki 635 sampel disetarakan secara sintetis menjadi 1.767 sampel agar seimbang dengan kelas mayoritas "*Graduate*", di mana proses SMOTE ini dieksekusi secara anti-kebocoran data (anti-data leakage) karena hanya diterapkan secara eksklusif pada sirkuit data latih tanpa menyentuh data uji yang murni.

3.3. SMOTE (*Imbalanced Data*)

Penerapan algoritma *Synthetic Minority Over-sampling Technique* (SMOTE) dalam penelitian ini berhasil menyelesaikan masalah ketimpangan sebaran kelas (*imbalanced data*) yang menjadi salah satu hambatan utama pada data mentah. Berdasarkan log output data pelatihan awal, ditemukan kesenjangan jumlah sampel yang sangat signifikan antara kelas mayoritas dan minoritas, di mana kelas *"Graduate"* mendominasi dengan 1.767 sampel, sedangkan kelas *"Enrolled"* berada di posisi paling bawah dengan hanya memiliki 635 sampel. Studi terdahulu menunjukkan bahwa penerapan Synthetic Minority Over-sampling Technique (SMOTE) efektif dalam meningkatkan representasi kelas minoritas dan meningkatkan kinerja model klasifikasi pada dataset yang tidak seimbang [11]. Melalui operasi algoritma SMOTE, sampel pada kelas minoritas (*Enrolled*) tersebut diduplikasi secara sintetis dengan cara mengalkulasi dan memetakan titik data baru di sekitar tetangga terdekatnya (*k-nearest neighbors*), sehingga seluruh kelas target akhirnya mencapai jumlah sebaran yang seimbang, yaitu masing-masing tepat 1.767 sampel. Secara metodologis, keberhasilan proses penyeimbangan kelas ini memberikan dampak besar bagi model *Random Forest* dan XGBoost agar tidak mengalami bias prediksi yang cenderung hanya condong ke kelas mayoritas. Selain itu, pengeksekusian SMOTE yang sengaja dilakukan secara eksklusif hanya pada sirkuit data latih (*training set*) setelah proses pembagian data makro selesai berhasil menjamin bahwa data uji (*testing set*) tetap murni, sehingga seluruh rangkaian eksperimen ini sepenuhnya terbebas dari risiko kebocoran informasi data (*data leakage*).

3.4. Evaluasi Model (*Model Evaluation*)

Berdasarkan hasil eksperimen setelah melalui tahap optimalisasi parameter, pengujian performa komparatif antara dua algoritma *ensemble learning* menunjukkan bahwa model **Random Forest** memiliki keunggulan performa makro yang lebih stabil dibandingkan dengan **XGBoost** dalam memprediksi status retensi mahasiswa Seperti pada gambar 2 berikut ini.

```
--- HASIL PERFORMA MODEL (SETELAH OPTIMALISASI) ---
Akurasi Optimal Random Forest : 77.40%
Akurasi Optimal XGBoost       : 76.05%

Detail Report Random Forest:
              precision    recall  f1-score   support

     Dropout       0.83      0.73      0.78       284
    Enrolled       0.51      0.52      0.52       159
    Graduate       0.83      0.90      0.86       442

    accuracy                           0.77       885
   macro avg       0.73      0.71      0.72       885
weighted avg       0.78      0.77      0.77       885


Detail Report XGBoost:
              precision    recall  f1-score   support

     Dropout       0.80      0.74      0.77       284
    Enrolled       0.48      0.44      0.46       159
    Graduate       0.82      0.89      0.86       442

    accuracy                           0.76       885
   macro avg       0.70      0.69      0.69       885
weighted avg       0.75      0.76      0.76       885
```

Gambar 2. Hasil Evaluasi Model

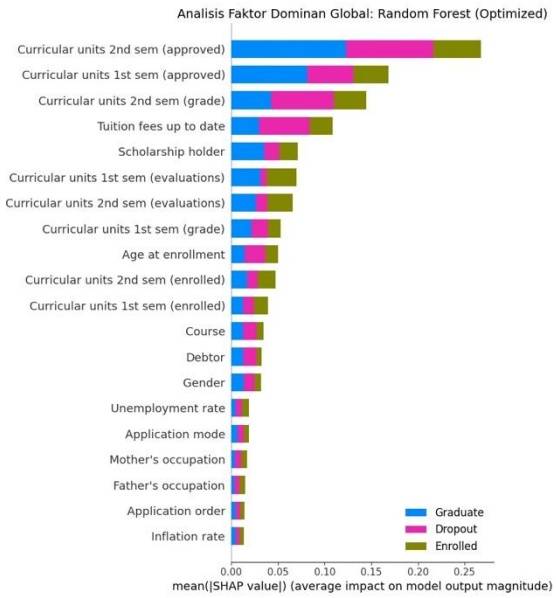
Evaluasi metrik akurasi global mencatat bahwa Random Forest berhasil mencapai tingkat akurasi optimal sebesar **77,40%**, unggul tipis sebesar 1,35% di atas XGBoost yang menghasilkan akurasi sebesar **76,05%**. Dari total 885 sampel data uji murni (*support*), sebaran data evaluasi terdistribusi secara objektif pada tiga kelas target, yaitu kelas *Dropout* sebanyak 284 sampel, kelas *Enrolled* sebanyak 159 sampel, dan kelas *Graduate* sebanyak 442 sampel.

Secara lebih mendalam, pengamatan pada *classification report* memperlihatkan karakteristik penanganan klasifikasi multikelas yang serupa pada kedua algoritma. Kedua model menunjukkan tingkat presisi (*precision*) dan keandalan penarikan kembali (*recall*) yang sangat tinggi pada kelas mayoritas yaitu *Graduate*, di mana Random Forest mencetak skor *F1-score* sebesar 0,86 (presisi 0,83 dan *recall* 0,90) yang setara dengan capaian *F1-score* XGBoost sebesar 0,86 (presisi 0,82 dan *recall* 0,89). Kemampuan mengidentifikasi kelas kritis *Dropout* juga memperlihatkan hasil yang solid; Random Forest memperoleh *F1-score* 0,78, sedikit mengungguli XGBoost yang memperoleh skor 0,77.

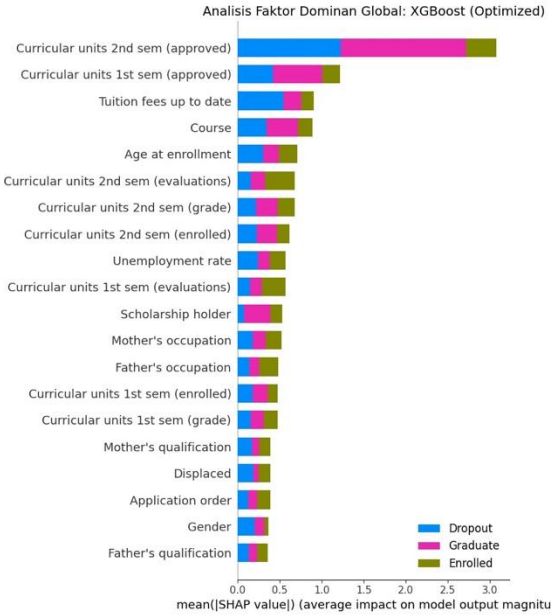
Namun, tantangan klasifikasi terbesar bagi kedua algoritma terletak pada kelas *Enrolled*, yang secara statistik merupakan kelas paling minoritas sebelum dilakukan manipulasi sintetis pada data latih. Model Random Forest menghasilkan skor *F1-score* sebesar 0,52 (presisi 0,51 dan *recall* 0,52), sedangkan XGBoost mengalami penurunan performa yang lebih landai dengan *F1-score* sebesar 0,46 (presisi 0,48 dan *recall* 0,44). Rendahnya performa pada kelas *Enrolled* ini terjadi karena karakteristik data mahasiswa yang berada pada fase transisi (masih aktif kuliah) secara natural memiliki irisan fitur yang sangat bias dan menyerupai pola data mahasiswa yang akan *Graduate* maupun yang berisiko *Dropout*. Kendati demikian, secara akumulatif, bobot rata-rata menyeluruh (*weighted average*) menegaskan keunggulan Random Forest dengan skor *F1-score* 0,77 dibandingkan XGBoost yang meraih skor 0,76, sehingga menetapkan model Random Forest sebagai arsitektur terbaik untuk diintegrasikan dengan fase interpretasi model menggunakan *Explainable AI* (XAI).

3.4. Interpretasi Model (Expanable AI)

Analisis berbasis *Explainable AI* (XAI) menggunakan metode SHAP (*SHapley Additive exPlanations*) digunakan untuk mengidentifikasi kontribusi variabel prediktor pada model *Random Forest* dan XGBoost dalam memprediksi retensi mahasiswa. Berdasarkan kalkulasi nilai rata-rata absolut SHAP, pada gambar 3 dan 4 menunjukkan kedua model konsisten bahwa performa akademik pada tahun kedua perkuliahan merupakan indikator dengan tingkat prediktif tertinggi. Variabel Curricular units 2nd sem (approved) (jumlah mata kuliah yang lulus di semester 2) menempati hierarki tertinggi dengan magnitudo dampak paling signifikan, diikuti oleh Curricular units 1st sem (*approved*). Temuan ini sejalan dengan penelitian terkini oleh Tan dan Rahaman (2025) yang menyatakan bahwa akumulasi kelulusan satuan kredit semester pada tahun pertama merupakan prediktor paling valid dalam memetakan persistensi studi di institusi pendidikan tinggi [14]. Visualisasi SHAP mengonfirmasi bahwa rendahnya jumlah kelulusan mata kuliah di semester awal berkorelasi positif dengan peningkatan probabilitas mahasiswa untuk putus studi (*Dropout*). Temuan ini juga didukung oleh studi terkini yang menunjukkan bahwa variabel performa akademik memiliki kontribusi paling dominan dalam model prediksi dropout berbasis Explainable AI. Hasil tersebut mengindikasikan bahwa capaian akademik pada tahun pertama dapat dimanfaatkan sebagai indikator utama dalam pengembangan sistem peringatan dini untuk mengidentifikasi mahasiswa yang berisiko tinggi mengalami putus studi [15]. Kendati demikian, kedua algoritma menunjukkan perbedaan dalam mengekstrak fitur sekunder; *Random Forest* berhasil mengidentifikasi faktor stabilitas finansial sebagai variabel penting melalui parameter kelancaran pembayaran biaya kuliah (Tuition fees up to date) dan status penerima beasiswa (Scholarship holder) pada posisi lima besar. Sebaliknya, XGBoost memberikan bobot yang lebih besar pada fitur program studi (Course) dan jalur penerimaan (Application mode), serta menempatkan variabel latar belakang orang tua pada posisi dengan kontribusi terendah. Mengingat *Random Forest* menghasilkan akurasi global yang lebih tinggi yaitu sebesar 77,40% pada data pengujian, maka pola pemetaan fitur yang menempatkan kombinasi faktor akademik semester dua dan kendala pembayaran kuliah ini dinilai lebih representatif untuk digunakan oleh manajemen perguruan tinggi dalam merancang sistem peringatan dini (*early warning system*) sebelum mahasiswa menyelesaikan tahun pertama perkuliahan.



Gambar 3. Analisis Faktor Dominan Menggunakan Random Forest



Gambar 4. Analisis Faktor Dominan Menggunakan XGBoost

4. CONCLUSION

Berdasarkan hasil eksperimen, analisis komparatif, dan interpretasi model yang telah dilaksanakan dalam penelitian ini, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. Efektivitas Prapemrosesan Data dan SMOTE: Implementasi pipa data (*data pipeline*) yang terstruktur—mulai dari pembersihan fitur (*feature cleaning*), transformasi label target, hingga imputasi nilai kosong menggunakan strategi *median*—berhasil menjaga integritas data tanpa memicu bias sebaran variabel. Lebih lanjut, penggunaan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) terbukti efektif menyelesaikan masalah ketidakseimbangan kelas target secara signifikan pada sirkuit data latih, di mana kelas minoritas (*Enrolled*) berhasil disetarakan dari 635 sampel menjadi 1.767 sampel guna menghindari bias prediksi model terhadap kelas mayoritas.
2. Performa Model Terbaik: Berdasarkan hasil pengujian komparatif pada data uji murni (20% porsi data evaluasi), model Random Forest menghasilkan performa klasifikasi multikelas yang lebih unggul dengan akurasi global mencapai 77,40%. Performa ini mengungguli model XGBoost yang mencatatkan akurasi sebesar 76,05%. Kedua model menunjukkan akurasi dan keandalan yang tinggi dalam mengidentifikasi kelas *Graduate* dan *Dropout*, namun menghadapi tantangan klasifikasi yang serupa pada kelas transisi (*Enrolled*) akibat tingginya irisan karakteristik fitur pada fase tersebut.
3. Faktor Dominan Global (SHAP): Analisis berbasis *Explainable AI* (XAI) melalui metode SHAP berhasil membongkar mekanisme internal model dan menetapkan konsensus bahwa performa akademik pada tahun pertama perkuliahan merupakan indikator prediktif paling krusial. Variabel jumlah mata kuliah yang lulus di semester dua (Curricular units 2nd sem (approved)) menjadi faktor yang paling memengaruhi keputusan prediksi retensi mahasiswa, diikuti secara konsisten oleh jumlah kelulusan mata kuliah di semester pertama.
4. Implikasi Manajerial Perguruan Tinggi: Model terbaik (*Random Forest*) berhasil memetakan bahwa selain performa akademik, faktor stabilitas finansial—yang dipresentasikan melalui parameter kelancaran pembayaran biaya kuliah (Tuition fees up to date) dan status penerima beasiswa (Scholarship holder)—turut memegang peranan penting dalam struktur keputusan model. Kombinasi temuan ini memberikan landasan taktis bagi pengelola perguruan tinggi untuk merancang sistem peringatan dini (*early warning system*) yang berfokus memberikan intervensi bimbingan akademis intensif serta bantuan finansial tepat sebelum mahasiswa merampungkan tahun pertama perkuliahan guna menekan angka putus kuliah (*dropout*).

ACKNOWLEDGEMENTS

Penulis mengucapkan puji syukur kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Apresiasi tertinggi dan terima kasih kepada seluruh tim peneliti atas kolaborasi dan kerja kerasnya dalam merampungkan riset prediksi retensi mahasiswa berbasis Explainable AI (XAI) ini. Keberhasilan pencapaian hasil eksperimen yang objektif ini adalah buah dari dedikasi dan kerja sama terbaik kita semua.

REFERENCES

[1] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, “Predicting Student Dropout in Higher Education,” Jun. 2016, doi: 10.48550/arXiv.1606.06364.

[2] P. G. Kreysa, “Patterns of student departure and transfer behavior: a quantitative analysis of institutional retention,” *Front. Educ. (Lausanne).*, vol. Volume 11-2026, 2026, [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2026.1808302

[3] Á. Kocsis and G. Molnár, “Factors influencing academic performance and dropout rates in higher education,” *Oxf.*

Rev. Educ., vol. 51, no. 3, pp. 414–432, May 2025, doi: 10.1080/03054985.2024.2316616.

[4] S. Kim, E. Yoo, and S. Kim, "Why do students drop out? university dropout prediction and associated factor analysis using machine learning techniques," *arXiv preprint arXiv:2310.10987*, 2023.

[5] M. V Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-class phased prediction of academic performance and dropout in higher education," *Applied Sciences*, vol. 13, no. 8, p. 4702, 2023.

[6] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, p. 102474, 2024, doi: https://doi.org/10.1016/j.techsoc.2024.102474.

[7] A. Alhardi and S. Alan, *Predicting Student Dropout in Higher Education Using Machine Learning Techniques : A Predictive Model Using XGBoost Algorithm*. 2024.

[8] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 274–300, 2024, doi: 10.1007/s40593-023-00331-8.

[9] V. Realinho, J. Machado, L. Baptista, and M. V Martins, "Predicting student dropout and academic success," *Data (Basel).*, vol. 7, no. 11, p. 146, 2022.

[10] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance," *Data (Basel).*, vol. 6, no. 2, p. 11, 2021.

[11] A. Fernández, F. Herrera, and N. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/jair.1.11192.

[12] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information*, vol. 11, no. 4, p. 193, 2020.

[13] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[14] I. Alcauter, L. Martinez-Villaseñor, and H. Ponce, "Explaining factors of student attrition at higher education," *Computación y Sistemas*, vol. 27, no. 4, pp. 929–940, 2023.

[15] R. A. M. Al Hashmi, P. D. Zervopoulos, H. M. Elmehdi, and I. Ozturk, "Predicting Dropout in MENA STEM Higher Education Using Explainable AI: A Machine Learning Approach," *Emerging Science Journal*, vol. 9, no. Special Issue, pp. 268–286, 2025.