SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 8 No 3 – Desember 2025 | https://s.id/sintechjournal
DOI: https://doi.org/10.31598
Publishing: Prahasta Publisher

Penerapan *Random Forest* Dan *Borderline* SMOTE Untuk Prediksi Risiko *Drop Out* Mahasiswa

Christian Bautista[1], Daniel Udjulawa[2]

[1,2]Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Jalan Rajawali 14, Palembang, Indonesia

e-mail: christianbautista_2226250002@mhs.mdp.ac.id[1], daniel@mdp.ac.id[2]

Received : November, 2025	Accepted : December, 2025	Published : December, 2025

Abstract

Education plays a key role in developing qualified, knowledgeable, and competitive human resources. However, one of the main challenges in higher education is the phenomenon of student dropout, which can impact academic quality and institutional accreditation. This study aims to predict the risk of student dropout using the Random Forest algorithm optimized through Grid Search and data balancing with Borderline-SMOTE to address class imbalance. The dataset used comes from the Open University Learning Analytics Dataset (OULAD), which includes demographic, academic, and online learning activity data of students. The research stages include data cleaning, feature transformation and normalization, application of 5-Fold Cross Validation, and determination of optimal parameters (n_estimators and max_depth) using Grid Search. The evaluation results show that both models with and without Borderline-SMOTE have similar performance, with accuracies of 79.1% and 78.3%, respectively. Therefore, data balancing does not significantly improve model performance. Feature importance analysis reveals that the score attribute and total VLE clicks are the most influential factors on the risk of dropout. This model is expected to be used as an early warning system for universities in identifying at-risk students early on.

Keywords: *drop out, prediction, random forest, borderline-SMOTE, machine learning*

Abstrak

Pendidikan berperan sebagai sarana utama dalam pengembangan sumber daya manusia yang berkualitas, berwawasan luas, dan kompetitif. Namun, salah satu tantangan utama di perguruan tinggi adalah fenomena drop out mahasiswa yang dapat memengaruhi kualitas akademik dan akreditasi institusi. Penelitian ini bertujuan untuk memprediksi risiko drop out mahasiswa menggunakan algoritma Random Forest yang dioptimalkan melalui Grid Search serta penyeimbangan data dengan Borderline-SMOTE untuk mengatasi ketidakseimbangan kelas. Dataset yang digunakan berasal dari Open University Learning Analytics Dataset (OULAD) yang mencakup data demografis, akademik, dan aktivitas pembelajaran daring mahasiswa. Tahapan penelitian meliputi pembersihan data, transformasi dan normalisasi fitur, penerapan 5-Fold Cross Validation, serta penentuan parameter optimal (n_estimators dan max_depth) menggunakan Grid Search. Hasil evaluasi menunjukkan bahwa baik model dengan maupun tanpa Borderline-SMOTE memiliki kinerja serupa, dengan akurasi masing-masing 79,1% dan 78,3%, sehingga penyeimbangan data tidak memberikan peningkatan signifikan terhadap performa model. Analisis feature importance mengungkap bahwa atribut score dan total klik VLE merupakan faktor paling berpengaruh terhadap risiko drop out. Model ini diharapkan dapat digunakan sebagai sistem peringatan dini bagi perguruan tinggi dalam mengidentifikasi mahasiswa berisiko sejak awal.

Kata Kunci: *drop out, prediksi, random forest, borderline-SMOTE, machine learning*

1. PENDAHULUAN

Saat ini, pendidikan sangat penting bagi masyarakat, terutama dalam membangun perekonomian, menjaga stabilitas sosial-politik, serta meningkatkan indeks pembangunan manusia [1]. Hal ini terlihat semakin banyaknya individu yang memilih melanjutkan pendidikan ke tingkat universitas atau sekolah tinggi menunjukkan pentingnya peran pendidikan tinggi di masyarakat [2]. Perguruan tinggi berperan sebagai lembaga yang menyelenggarkaan pendidikan akademik bagi mahasiswa dengan tujuan memberikan proses pembelajaran yang berkualitas, sehingga mampu menghasilkan sumber daya manusia yang berpengetahuan, cerdik, dan inovatif yang berkontribusi pada pembangunan negara [3].

Namun, dalam proses penyelenggaraan pendidikan tinggi, salah satu permasalahan yang cukup serius adalah fenomena *drop out* atau putus studi mahasiswa [4]. Situasi ini tidak hanya berdampak negatif bagi mahasiswa secara pribadi, tetapi juga menimbulkan konsekuensi bagi perguruan tinggi sebagai institusi penyelenggara pendidikan [5]. Mahasiswa *drop out* (DO) bahkan menjadi persoalan yang cukup mengganggu bagi perguruan tinggi di Indonesia karena dapat memengaruhi akreditasi maupun pemeringkatan klasterisasi universitas [6]. Berdasarkan Statistik Pendidikan Tinggi 2022, tercatat sebanyak 375.134 mahasiswa di Indonesia mengalami *drop out* pada tahun tersebut, atau sekitar 4,02% dari total mahasiswa aktif di semua jenjang perguruan tinggi [7].

Mahasiswa yang termasuk dalam kategori *drop out* terdiri atas mereka yang diberhentikan oleh pihak kampus, menghentikan kuliahnya di tengah jalan, maupun mengundurkan diri secara sukarela [8]. Fenomena ini menjadi tantangan besar bagi perguruan tinggi karena apabila tidak terkendali, dapat memberikan konsekuensi bagi institusi pendidikan tinggi [9].

Salah satu upaya untuk mengurangi risiko mahasiswa berhenti kuliah adalah dengan melihat data keaktifan mahasiswa serta memperkirakan potensi terjadinya ketidakaktifan dalam proses terjadinya ketidakaktifan dalam proses belajar [10]. Analisis terhadap kondisi tersebut akan membantu mengungkap faktor-faktor yang menjadi penyebab utama *drop out* [11]. Oleh karena itu, diperlukan sebuah metode prediksi yang mampu memberikan hasil mengenai mahasiswa yang berpotensi mengalami *drop out*, sehingga perguruan tinggi dapat mengambil langkah pencegahan sejak dini [5].

Algoritma *Random Forest* digunakan sebagai metode prediksi dalam penelitian ini. Untuk menghasilkan prediksi yang lebih akurat dan konsisten, algoritma ini digunakan dalam metode pembelajaran kelompok yang menggabungkan berbagai pohon keputusan [12]. Selain itu, *Random Forest* juga efektif dalam mengolah *dataset* yang bersifat kompleks maupun tidak seimbang [13].

Kekacauan kelas sering terjadi pada data besar. Keseluruhan kelas terjadi ketika kelas tertentu memiliki jumlah sampel yang jauh lebih besar daripada kelas lainnya [14]. Kondisi ini dapat menyebabkan model prediksi cenderung mengabaikan kelas minoritas, sehingga akurasi dalam mengenali kasus yang jarang terjadi menjadi rendah [15]. Untuk menyelesaikan masalah ini, penelitian ini menggunakan teknik *Borderline-Synthetic Minority Over-sampling Technique* (SMOTE). *Borderline*-SMOTE adalah pengembangan SMOTE sebelumnya yang digunakan untuk mengatasi situasi di mana data kelas minoritas terlalu dekat dengan data kelas mayoritas. Teknik ini membuat data sintetis dengan lebih hati-hati agar tidak terjadi tumpang tindih antara kedua kelas [16].

Hasil penelitian terdahulu mendukung efektivitas kedua metode tersebut. Penelitian yang dilakukan oleh Harkamsyah Andrianof, Aggy Pramana Gusman, dan Okta Andrica Putra pada tahun 2025 menggunakan data akademik seperti nilai mata kuliah inti, tingkat kehadiran, dan Indeks Prestasi Kumulatif (IPK) untuk memprediksi kelulusan siswa. Dalam penelitian tersebut, algoritma *Random Forest* diterapkan melalui beberapa tahapan, mulai dari *preprocessing* data, implementasi model, hingga evaluasi kinerja. Hasil menunjukkan bahwa *Random Forest* mampu memberikan kinerja yang cukup baik, dengan tingkat akurasi mencapai 87,5%, presisi sebesar 86,3%, dan *recall* sebesar 85,9% [17].

Pada tahun 2023, penelitian yang dilakukan oleh Ery Permana Yudha, Eko Purwanto, dan Joni

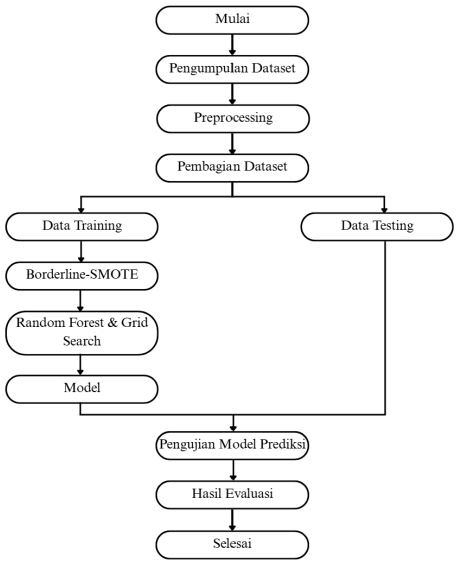
Maulindar fokus pada sistem deteksi penyakit hepatitis dengan memanfaatkan metode *Fuzzy K-NN* dan *Ensemble Learning*, serta penyeimbangan data menggunakan teknik *Borderline*-SMOTE. Tahapan penelitian meliputi preprocessing dataset dengan cara mengisi data yang hilang (*missing value*) dan menyeimbangkan distribusi kelas. Hasil pengujian menunjukkan bahwa pendekatan tersebut mampu mencapai rata-rata akurasi sebesar 94,87% [18].

Berdasarkan uraian tersebut, Kombinasi *Random Forest* dan *Borderline*-SMOTE diterapkan untuk memprediksi mahasiswa yang berpotensi mengalami *drop out*. *Random Forest* dipilih karena mampu menghasilkan prediksi yang akurat dan stabil, tidak sensitif terhadap *noise*, serta menyediakan *feature importance* yang bermanfaat untuk analisis kebijakan [19]. Di sisi lain, *Borderline*-SMOTE digunakan untuk mengatasi ketidakseimbangan data dengan membentuk distribusi sampel yang lebih konsisten pada area *borderline*, sehingga kinerja model menjadi lebih stabil dibandingkan tanpa penyeimbangan [20].

Dari sisi penelitian, kombinasi *Random Forest* dan *Borderline*-SMOTE pada dataset OULAD masih jarang dievaluasi secara mendalam, terutama terkait stabilitas model serta relevansi fitur demografis, akademik, dan interaksi digital. GAP ini menjadi dasar penelitian dilakukan, ditambah urgensi tingginya angka *drop out* yang berdampak pada akreditasi dan kualitas institusi. Oleh karena itu, penelitian ini bertujuan menerapkan algoritma *Random Forest* yang dikombinasikan dengan *Borderline*-SMOTE untuk membangun model prediksi yang lebih informatif dan stabil, sehingga dapat dimanfaatkan sebagai alat pendukung dalam membantu perguruan tinggi mengidentifikasi mahasiswa yang berpotensi mengalami putus studi sejak awal.

2. METODE PENELITIAN

Pada Gambar 1 menampilkan tahapan metode penelitian yang digunakan dalam memprediksi risiko *drop out* mahasiswa, mulai dari pengolahan data, penerapan *Borderline*-SMOTE, hingga pembangunan model prediksi menggunakan algoritma *Random Forest*.



Gambar 1. Skema Penelitian

2.1 Pengumpulan *Dataset*

Penelitian ini menggunakan data dari *Open University Learning Analytics Dataset* (OULAD), yaitu dataset terbuka yang merepresentasikan proses pembelajaran *daring* di The Open University, Inggris. *Dataset* tersebut berisi beragam informasi terkait mahasiswa yang terdiri dari data studentInfo, studentRegistration, assessment, studentAssessment, vle, studentVle, dan courseVle. Secara keseluruhan, dataset ini mencatat sekitar 32.593 mahasiswa dari berbagai modul dan periode studi.

Dari seluruh data tersebut, penelitian ini hanya menggunaan studentInfo, studentAssessment, assessment, studentVle, dan vle. Data tersebut berisi fitur utama, yaitu demografis, akademik, dan perilaku di *Virtual Learning Environment* (VLE) [21]. Tabel data studentInfo digunakan karena memuat data demografi dan status akhir mahasiswa, assessment dan studentAssessment menyediakan informasi performa akademik, sedangkan studentVle dan vle digunakan untuk merepresentasikan aktivitas dan keterlibatan mahasiswa di lingkungan belajar daring [22].

2.2 *Preprocessing*

Preprocessing dilakukan untuk memastikan bahwa data dalam kondisi yang tepat dan sesuai untuk proses pemodelan. Langkah ini meliputi beberapa tahapan, di antaranya:

2.2.1 Pembersihan Data

Langkah awal adalah pembersihan data, bertujuan untuk meningkatkan kualitas dataset melalui berbagai proses perbaikan. Pada tahap ini, nilai yang hilang diidentifikasi dan diselesaikan, data yang tidak konsisten dihapus, dan duplikat data ditemukan dan dihapus [23].

2.2.2 Transformasi Data

Tahap transformasi data dilakukan dengan menyesuaikan format dan tipe variabel agar sesuai untuk proses pemodelan [24]. Pada tahap ini, kategori karakteristik diubah menjadi nilai numerik. Misalnya, variabel jenis kelamin dikodekan dengan nilai 1 untuk laki-laki dan 0 untuk perempuan, dan tingkat ekonomi bagian menjadi tiga kategori, serta variabel umur yang dikelompokkan ke dalam tiga rentang usia. Proses ini bertujuan agar seluruh data memiliki format yang dapat dikenali dan diolah oleh algoritma *machine learning*.

2.2.3 Normalisasi Data

Rentang nilai yang terlalu besar antar variabel dapat memengaruhi kinerja model dan menurunkan tingkat akurasi, sehingga diterapkan proses normalisasi untuk menyetarakan skala data [25]. Sebagai contoh, atribut jumlah klik pada sistem VLE dinormalisasi ke dalam rentang 0–100 agar tidak mendominasi fitur lainnya dalam proses pembelajaran model.

2.3 Pembagian *Dataset*

Penelitian ini menerapkan metode *K-Fold Cross Validation* dengan lima *fold* dalam proses pembagian data. Melalui metode ini, *dataset* dibagi menjadi lima bagian dengan ukuran yang seimbang, di mana setiap *fold* mewakili sekitar 20% dari total data. Metode ini memungkinkan setiap kolom digunakan sebagai data uji satu kali dan empat kali sebagai data latih. Ini menjadikan proses evaluasi model lebih objektif dan menyeluruh.

2.4 *Borderline*-SMOTE

Borderline-SMOTE merupakan pengembangan dari metode SMOTE yang dirancang untuk menangani ketidakseimbangan data ketika sampel dari kelas minoritas berada terlalu dekat dengan kelas mayoritas. Teknik ini menghasilkan data sintetis secara lebih selektif guna meminimalkan kemungkinan tumpang tindih antara kedua kelas [16].

Metode ini diawali dengan mengidentifikasi himpunan DANGER, yaitu kumpulan data dari kelas minoritas yang berada di sekitar kelas mayoritas dan berisiko tinggi salah diklasifikasikan. Setelah himpunan tersebut terdeteksi, algoritma SMOTE diterapkan untuk menghasilkan data sintetis baru pada area perbatasan antar kelas. Pendekatan ini memperkaya representasi kelas minoritas sekaligus mengurangi risiko *overfitting*, sehingga model yang dibangun memiliki kemampuan prediksi yang lebih optimal [26].

2.5 Model

2.5.1 Algoritma *Random Forest*

Random Forest adalah algoritma *ensemble learning* yang menggunakan beberapa pohon keputusan untuk menghasilkan hasil prediksi yang lebih stabil dan akurat. Algoritma ini dianggap sebagai salah satu pendekatan yang potensial karena mampu menangani data yang kompleks sekaligus membangun model prediktif dengan tingkat akurasi yang tinggi [17].

Random Forest melakukan dua proses utama, yakni membangun beberapa pohon keputusan secara acak dan menggabungkan hasilnya untuk memperoleh prediksi akhir. Proses ini dikenal sebagai *feature bagging*, yang bertujuan untuk memastikan setiap pohon memiliki keragaman struktur sehingga tidak terlalu mirip satu sama lain. Dengan demikian, ketika terdapat fitur yang sangat dominan dalam prediksi, algoritma ini mencegah fitur tersebut muncul berulang kali di banyak pohon yang dapat menyebabkan korelasi antar pohon terlalu tinggi [27].

2.5.2 *Grid Search*

Penentuan parameter pada algoritma *Random Forest* dilakukan melalui proses pencarian kombinasi parameter terbaik secara sistematis menggunakan pendekatan *Grid Search*. Proses ini bertujuan untuk menemukan konfigurasi parameter yang menghasilkan kinerja model paling optimal [28].

Dua parameter utama yang diuji, yaitu jumlah pohon (*n_estimators*) dan batas kedalaman maksimum pohon (*max_depth*). Pada penelitian ini, *n_estimators* diuji pada rentang [100, 200], sedangkan *max_depth* diuji pada kombinasi [5, 10, 20]. Rentang ini dipilih untuk menjaga keseimbangan antara kinerja dan efisiensi

komputasi, sehingga pencarian parameter melalui *Grid Search* dapat meningkatkan performa dan stabilitas model [29]. Metode *K-Fold Cross Validation* dengan lima lipatan digunakan untuk memutar setiap kombinasi parameter. Metode ini membagi data menjadi lima bagian yang seimbang, empat di antaranya digunakan sebagai data latih dan satu lagi digunakan secara bergantian sebagai data uji.

Berdasarkan rata-rata hasil evaluasi tertinggi dari lima *fold*, sehingga parameter yang dipilih terbukti memberikan kinerja yang paling konsisten. Setelah parameter optimal diperoleh, model akhir dilatih menggunakan seluruh data pelatihan dan diuji terhadap data pengujian untuk mengukur performa akhir model [30][31].

2.6 Evaluasi Model

Dalam penelitian ini, *Confusion Matrix* digunakan sebagai dasar evaluasi kinerja model untuk menghitung metrik akurasi, presisi, *recall*, dan *F1-score*.

Akurasi merupakan metrik yang menggambarkan tingkat ketepatan model dalam melakukan prediksi secara keseluruhan [32]. Berikut rumus menghitung akurasi menggunakan persamaan (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

Presisi merupakan perbandingan antara jumlah data yang diprediksikan positif secara keseluruhan dan jumlah prediksi positif yang benar (*True Positive*) [32]. Berikut rumus menghitung presisi menggunakan persamaan (2).

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

Recall merupakan ukuran yang menunjukkan seberapa besar proporsi data positif yang berhasil diidentifikasi dengan benar oleh model, yaitu rasio antara *True Positive* (TP) dan seluruh data yang sebenarnya termasuk kelas positif [32]. Berikut rumus menghitung *recall* menggunakan persamaan (3).

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (3)$$

F1-Score digunakan untuk menilai kinerja model secara keseluruhan dan merupakan rata-rata dari nilai presisi dan *recall* [32]. Berikut rumus menghitung *F1-Score* menggunakan persamaan (4).

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \times 100\% \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil *Preprocessing*

Proses *preprocessing* dilakukan pada setiap tabel dalam *dataset* OULAD untuk memastikan data memiliki kualitas dan konsistensi yang baik sebelum tahap pemodelan. Tahapan ini meliputi penghapusan data kosong, perbaikan format yang tidak valid, normalisasi nilai numerik, serta transformasi atribut kategorikal menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*.

Dataset yang digunakan mencakup informasi demografis mahasiswa (studentInfo), hasil tugas (studentAssessment dan assessment), dan interaksi mahasiswa dalam pembelajaran daring (studentVle dan vle). Seluruh data tersebut digabungkan berdasarkan atribut *id_student*, sehingga menghasilkan satu *dataset* terintegrasi yang merepresentasikan profil akademik dan perilaku pembelajaran daring setiap mahasiswa.

Setelah pembersihan dan penggabungan, diperoleh *dataset* akhir dengan sebelas atribut prediktor, yaitu *gender, imd_band*, *age_band*, *disability ,num_of_prev_attempts*, *student_credits, score*, *jumlah_telat_mengumpulkan*, *total_klik_vle*, dan *dropout* sebagai atribut target yang akan diprediksi. *Dataset* akhir terdiri dari 24.682 data mahasiswa, yang kemudian digunakan sebagai data dasar dalam tahap pemodelan prediksi risiko *drop out*. Tabel 1 menampilkan hasil ringkasan dari *preprocessing* pada *dataset* OULAD.

Tabel 1: Hasil *Preprocessing Dataset*

id_student	imd_band	age_band	gender	disability	num_of_prev_attempts	student_credits	score	jumlah_telat_kumpul	total_klik_vle	dropout
3733	10	3	1	0	0	60	0	0	0	1
6516	9	3	1	0	0	60	61.8	0	74.7	0
8462	4	3	1	0	1	60	87	0	0	1
11391	10	3	1	0	0	240	82	0	25.9	0
23629	3	1	0	0	2	60	82.5	0	3.6	1

3.2 Pembagian *Dataset*

Setelah data siap digunakan, *dataset* tersebut dibagi menggunakan 5-*Fold Cross Validation* untuk memastikan setiap *fold* memiliki proporsi kelas yang seimbang.

3.3 Hasil Model

Pada tahap pemodelan, dua pengujian dilakukan, yaitu *Random Forest* pada data asli dan *Random Forest* dengan penerapan *Borderline*-SMOTE. *Borderline*-SMOTE diterapkan hanya pada data latih untuk menyeimbangkan distribusi kelas, sedangkan data uji tetap menggunakan data asli. Kedua pengujian menggunakan strategi optimasi yang sama, yaitu *Grid Search* dengan ruang pencarian parameter tetap pada kombinasi *n_estimators* = [100, 200] dan *max_depth* = [5, 10, 20].

3.3.1 Tanpa *Borderline*-SMOTE

Pengujian awal dilakukan pada *dataset* yang belum diseimbangkan untuk menilai performa dasar algoritma. Berdasarkan hasil *preprocessing*, terdapat 12.598 mahasiswa *drop out* dan 12.084 mahasiswa *non-drop out*, sehingga model cenderung lebih mudah mengenali pola dari kelas mayoritas.

Hasil terbaik diperoleh pada kombinasi parameter *n_estimators* = 200 dan *max_depth* = 10. Nilai akurasi, presisi, *recall*, dan *F1-score* dari setiap *fold* disajikan pada Tabel 2 berikut.

Tabel 2: Hasil Evaluasi Tanpa *Borderline*-SMOTE

Fold	Akurasi	Presisi	*Recall*	*F1-Score*
1	0.787	0.843	0.715	0.774
2	0.793	0.847	0.727	0.782
3	0.787	0.838	0.723	0.776
4	0.797	0.853	0.728	0.786
5	0.788	0.841	0.721	0.776
Rata-Rata	0.791	0.845	0.723	0.779

Secara keseluruhan, model tanpa Borderline-SMOTE mencapai akurasi 79,1%, presisi 84,5%, *recall* 72,3%, dan *F1-score* 77,9%. Hasil ini menunjukkan kinerja yang baik dengan keseimbangan antara ketepatan prediksi dan kemampuan mendeteksi mahasiswa *drop out*.

Kinerja tertinggi diperoleh pada *fold* ke-4 dengan akurasi 79,7%, sedangkan nilai terendah muncul pada *fold* ke-3 sebesar 78,7%. Rentang akurasi antar-*fold* yang sempit (±1%) menandakan konsistensi model pada berbagai pembagian data. Meskipun demikian, nilai *recall* yang hanya berkisar 0,71–0,73 memperlihatkan bahwa model masih kurang optimal dalam mengenali seluruh mahasiswa berisiko *drop out* akibat ketidakseimbangan kelas.

3.3.2 *Borderline*-SMOTE

Selanjutnya diterapkan *Borderline*-SMOTE untuk menyeimbangkan jumlah data pada kedua kelas. Teknik ini menghasilkan sampel sintetis di area perbatasan antar kelas sehingga total data menjadi 25.196 mahasiswa dengan distribusi seimbang dan meningkatkan kemampuan model dalam mendeteksi kelas minoritas yang sebelumnya sulit teridentifikasi.

Model diuji kembali menggunakan kombinasi parameter terbaik (*n_estimators* = 200, *max_depth* = 20), dan hasilnya disajikan pada Tabel 3.

Tabel 3: Hasil Evaluasi *Borderline*-SMOTE

Fold	Akurasi	Presisi	*Recall*	*F1-Score*
1	0.787	0.861	0.694	0.769
2	0.780	0.815	0.737	0.774
3	0.785	0.857	0.695	0.768
4	0.781	0.815	0.737	0.774
5	0.780	0.813	0.740	0.774
Rata-Rata	0.783	0.833	0.720	0.772

Rata-rata hasil menunjukkan bahwa model dengan *Borderline*-SMOTE memperoleh akurasi 78,3%, presisi 83,3%, *recall* 72,0%, dan *F1-score* 77,2%. Kinerja ini relatif stabil dan menunjukkan keseimbangan yang baik antara ketepatan prediksi serta kemampuan mendeteksi mahasiswa *drop out*.

Performa terbaik dicapai pada *fold* ke-1 dengan akurasi 78,7%, sedangkan *fold* lainnya menunjukkan nilai yang tidak jauh berbeda (sekitar 78,0%). Rentang akurasi antar-*fold* hanya ±0,7%, yang menandakan kestabilan model pada berbagai pembagian data. Nilai *recall* yang meningkat konsisten di kisaran 0,69–0,74 menegaskan bahwa penyeimbangan data melalui *Borderline*-SMOTE membantu model menjadi lebih sensitif terhadap kelas minoritas tanpa menurunkan stabilitas performa.

3.4 Analisis *Feature Importance*

Analisis *feature importance* dilakukan untuk mengetahui tingkat kontribusi masing-masing atribut terhadap kemampuan model dalam memprediksi risiko *drop out* mahasiswa. Nilai kepentingan setiap fitur diperoleh dari rata-rata bobot feature importance yang dihasilkan oleh algoritma Random Forest pada seluruh fold pengujian, baik sebelum maupun sesudah penerapan Borderline-SMOTE.

Tabel 4: Nilai *Feature Importance*

Fitur	**Tanpa *Borderline*-SMOTE**	***Borderline*-SMOTE**
score	0.689	0.644
total_klik_vle	0.200	0.191
imd_band	0.032	0.061
student_credits	0.032	0.045
num_of_prev_attempts	0.019	0.019
gender	0.010	0.014
age_band	0.009	0.014
disability	0.007	0.010
jumlah_telat_mengumpulkan	0	0

Berdasarkan Tabel 4, fitur *score* tetap menjadi faktor paling dominan pada kedua model, meskipun nilainya sedikit menurun dari 0,689 menjadi 0,644 setelah penerapan *Borderline*-SMOTE. Penurunan ini terjadi karena penyeimbangan data membuat model mengeksplorasi pola keputusan yang lebih beragam pada kelas minoritas, sehingga ketergantungan terhadap satu fitur dominan berkurang. Dengan kata lain, *Borderline*-SMOTE memperluas representasi data di sekitar area batas antar kelas, sehingga kontribusi prediktif fitur lain menjadi lebih terlihat.

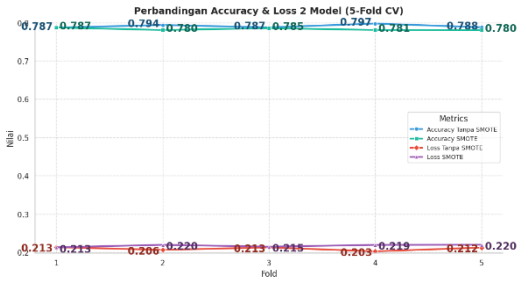
Fitur *total_klik_vle* menempati posisi kedua dengan nilai sekitar 0,19–0,20, mengindikasikan bahwa tingkat aktivitas mahasiswa di *platform* pembelajaran daring juga memiliki pengaruh kuat terhadap keberlanjutan studi. Korelasi antara rendahnya aktivitas daring dan meningkatnya risiko *drop out* memperkuat temuan bahwa perilaku interaksi digital memiliki peranan yang hampir setara dengan performa akademik.

Sementara itu, beberapa fitur seperti *imd_band* dan *student_credits* menunjukkan peningkatan bobot setelah penyeimbangan data. Hal ini mengindikasikan bahwa model yang telah diseimbangkan menjadi lebih peka terhadap faktor non-akademik seperti kondisi sosial-ekonomi dan beban kredit studi. Fitur *num_of_prev_attempts, gender, age_band*, dan *disability* juga memiliki kontribusi yang lebih kecil, tetapi tetap memberikan informasi tambahan terhadap profil mahasiswa berisiko.

Secara keseluruhan, hasil ini menegaskan bahwa performa akademik dan partisipasi belajar daring merupakan penentu utama risiko *drop out*, sementara *Borderline*-SMOTE membantu memperkuat sensitivitas model terhadap faktor-faktor pendukung lainnya tanpa mengubah urutan dominansi antar fitur.

3.5 Pembahasan

Hasil pengujian menunjukkan bahwa baik model tanpa maupun dengan penerapan *Borderline*-SMOTE memiliki performa yang relatif stabil. Model tanpa SMOTE menghasilkan akurasi 79,1 %, presisi 84,5 %, *recall* 72,3 %, dan *F1-score* 77,9 %. Sementara model yang telah diseimbangkan menggunakan *Borderline*-SMOTE mencatat akurasi 78,3 %, presisi 83,3 %, *recall* 72,0 %, dan *F1-score* 77,2 %. Selisih performa antar model relatif kecil, sehingga strategi penyeimbangan tidak memberikan peningkatan yang signifikan terhadap kualitas prediksi.



Gambar 2. Grafik Perbandingan Akurasi dan *Loss*

Berdasarkan Gambar 2, model tanpa *Borderline*-SMOTE memiliki rentang akurasi sekitar ±1%, lebih tinggi dibandingkan model dengan *Borderline*-SMOTE yang hanya ±0,7%. Hal ini menunjukkan bahwa meskipun akurasi total model setelah penyeimbangan data sedikit lebih rendah, tingkat kestabilannya lebih baik di setiap *fold*. Konsistensi ini penting, terutama karena data aktual yang digunakan dalam konteks pendidikan cenderung heterogen dan tidak selalu sesuai dengan pola di data pelatihan.

Presisi model tanpa *Borderline*-SMOTE (84,5%) lebih tinggi daripada model dengan *Borderline*-SMOTE (83,3%). Hal ini menunjukkan bahwa model tanpa penyeimbangan lebih tepat dalam mengidentifikasi mahasiswa yang benar-benar drop out, sementara model dengan SMOTE cenderung melakukan prediksi yang lebih inklusif terhadap kelas berisiko. Jadi, model dengan *Borderline*-SMOTE menjadi lebih peduli terhadap potensi risiko, meskipun sesekali salah memprediksi.

Nilai *recall* pada kedua model hampir sama, berada di sekitar 72%. Model dengan *Borderline*-SMOTE memperlihatkan stabilitas *recall* yang lebih baik antar-*fold*, menandakan kemampuan yang lebih konsisten dalam mengenali mahasiswa berisiko *drop out*. Ini menunjukkan bahwa penyeimbangan data membantu mengurangi ketidakstabilan prediksi pada berbagai pembagian data, meskipun tidak meningkatkan nilai *recall* secara signifikan.

Nilai F1-*score* kedua model relatif berdekatan, yaitu 77,9% (tanpa *Borderline*-SMOTE) dan 77,2% (dengan *Borderline*-SMOTE). Perbedaan yang sangat kecil ini menunjukkan bahwa keseimbangan antara presisi dan *recall* tetap terjaga pada kedua pengujian. *Borderline*-SMOTE tidak meningkatkan performa secara keseluruhan, namun membantu menjaga konsistensi prediksi terutama pada kelas minoritas.

Kinerja model secara umum dipengaruhi oleh dominasi fitur *score* dan *total_klik_vle* sebagai prediktor utama. Tingkat pencapaian akademik dan intensitas aktivitas belajar daring terbukti menjadi indikator kuat terhadap risiko *drop out*. Mahasiswa dengan nilai rendah dan aktivitas VLE terbatas cenderung memiliki risiko lebih tinggi dibandingkan mahasiswa yang aktif dan memiliki capaian akademik yang stabil.

Secara keseluruhan, kedua model dengan maupun tanpa Borderline-SMOTE menunjukkan performa yang sangat berdekatan pada seluruh metrik evaluasi. Perbedaan akurasi, presisi, recall, dan F1-score tergolong minimal, sehingga penerapan *Borderline*-SMOTE tidak memberikan peningkatan performa yang signifikan. Namun, teknik penyeimbangan ini tidak hanya menyeimbangkan distribusi kelas, tetapi juga memperkaya pola pada area batas antar kelas, sehingga model menjadi kurang sensitif terhadap variasi data latih dan menghasilkan stabilitas antar-*fold* yang lebih baik.

Dengan demikian, model tanpa *Borderline*-SMOTE sudah memadai untuk *dataset* dengan tingkat ketidakseimbangan yang rendah seperti pada penelitian ini. Sementara itu, *Borderline*-SMOTE dapat menjadi pilihan ketika model diterapkan pada *dataset* dengan ketimpangan kelas yang lebih tinggi, di mana peningkatan sensitivitas dan stabilitas prediksi menjadi lebih penting.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa *Random Forest* mampu memprediksi risiko *drop out* dengan performa yang baik, baik dengan maupun tanpa *Borderline*-SMOTE. Meskipun akurasi model tanpa *Borderline*-SMOTE sedikit lebih tinggi (79,1% dibanding 78,3%), penerapan *Borderline*-SMOTE memberikan keunggulan utama pada stabilitas antar-*fold*, karena teknik ini memperkaya representasi data pada area batas antar kelas. Hal ini membuat model lebih konsisten dan kurang sensitif terhadap variasi pembagian data. Fitur *score* dan *total_klik_vle* tetap menjadi prediktor dominan pada kedua skenario. Secara keseluruhan, *Borderline*-SMOTE menawarkan manfaat dari sisi

konsistensi dan pemerataan pola data, sehingga tetap relevan terutama untuk kondisi ketimpangan kelas yang lebih tinggi.

Sebagai tindak lanjut dari penelitian ini, ke depan model dapat dikembangkan dengan mempertimbangkan data perilaku mahasiswa yang lebih beragam untuk mendukung penerapan sistem peringatan sejak dini yang membantu institusi pendidikan melakukan intervensi lebih cepat terhadap mahasiswa yang berisiko *drop out*.

DAFTAR PUSTAKA

[1] N. T. H. Trinh, "Higher Education and Its Role for National Development. A Research Agenda with Bibliometric Analysis," *Interchange*, vol. 54, no. 2, pp. 125–143, 2023, doi: 10.1007/s10780-023-09493-9.

[2] H. Tanuwijaya and M. S. Erstiawan, "Peningkatan Pengetahuan Pendidikan Tinggi Bagi Peserta Didik SMA Barunawati Surabaya," *Kontribusi J. Penelit. dan Pengabdi. Kpd. Masy.*, vol. 4, no. 2, pp. 287–302, 2024, doi: 10.53624/kontribusi.v4i2.374.

[3] E. Mulyani, E. Ismantohadi, and K. Koriah, "Sistem Prediksi Potensi Drop Out Mahasiswa Menggunakan Rule Based System Pada Jurusan Teknik Informatika Politeknik Negeri Indramayu," *J. Inform.*, vol. 8, no. 1, pp. 19–25, 2020, doi: 10.36987/informatika.v8i1.1473.

[4] Moesarofah, "Mengapa mahasiswa putus kuliah sebelum lulus ?," *Pros. Semin. Nas. Progr. Pascasarj. Univ. PGRI Palembang*, pp. 52–55, 2021, [Online]. Available: https://jurnal.univpgri-palembang.ac.id/index.php/Prosidingpps/article/view/5472/4810

[5] A. S. Gustian and F. Mahardika, "Analisis Klasifikasi Risiko Dropout Mahasiswa Menggunakan Algoritma Decision Tree dan Random Forest," *Jupiter Publ. Ilmu Keteknikan Ind. Tek. Elektro dan Inform.*, vol. 3, no. 4, pp. 182–189, 2025, doi: 10.61132/jupiter.v3i4.980.

[6] T. A. Marzuqi, E. Kristiani, and Marcel, "Prediksi Mahasiswa Drop-Out Di Universitas XYZ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1345–1350, 2024, doi: 10.25126/jtiik.2024118689.

[7] R. B. Lubis, "Tingkat Drop Out Mahasiswa di Indonesia Kembali Turun pada 2022," GoodStats.id. Accessed: Sep. 27, 2025. [Online]. Available: https://goodstats.id/article/tingkat-drop-out-mahasiswa-di-indonesia-kembali-turun-pada-2022-4gr2P?

[8] R. Syahranita and S. Zaman, "Regresi Logistik Multinomial untuk Prediksi Kategori Kelulusan Mahasiswa," *J. Inform. Sunan Kalijaga)*, vol. 8, no. 2, pp. 102–111, 2023.

[9] A. W. Abdullah and A. Muhid, "Social support, academic satisfaction, and student drop out tendency," *Psikoislamika J. Psikol. dan Psikol. Islam*, vol. 18, no. 1, pp. 174–187, 2021, [Online]. Available: https://doi.org/10.18860/psi.v18i1.11546

[10] S. F. Puteri, B. S. Yulistiawan, and M. O. Pratama, "Identifikasi Dini Mahasiswa Berpotensi Drop-Out dengan Metode Machine Learning (Studi Kasus : Universitas Pembangunan Nasional ' Veteran ' Jakarta)," 2022.

[11] N. Y. L. Gaol, "Prediksi Mahasiswa Berpotensi Non Aktif Menggunakan Data Mining dalam Decision Tree dan Algoritma C4.5," *J. Inf. Teknol.*, vol. 2, pp. 23–29, 2020, doi: 10.37034/jidt.v2i1.22.

[12] M. Imani and A. Beikmohammadi, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE , ADASYN , and GNUS Under Varying Imbalance Levels," *Techonologies*, vol. 13, no. Ml, pp. 1–40, 2025.

[13] J. Dong and Q. Qian, "A Density-Based Random Forest for Imbalanced Data Classification," *Futur. Internet*, vol. 14, p. 90, 2022.

[14] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[15] S. Sidiq, Alfian, and N. S. Mabrur, "Pengembangan Model Prediksi Risiko Diabetes Menggunakan Pendekatan AdaBoost dan Teknik Oversampling SMOTE," *J. Ilm. Inform. dan Ilmu Komput.*, vol. 4, no. 1, pp. 13–23, 2025.

[16] R. Ridwan, E. H. Hermaliani, and M.

Ernawati, “Penerapan: Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian,” *Comput. Sci.*, vol. 4, no. 1, pp. 80–88, 2024, [Online]. Available: https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990

[17] H. Andrianof, A. P. Gusman, and O. A. Putra, “Implementasi Algoritma Random Forest untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus di Perguruan Tinggi Indonesia,” *J. Sains Inform. Terap. E-ISSN 2828-1659*, vol. 4, no. 1, pp. 24–28, 2024.

[18] E. P. Yudha, E. Purwanto, and J. Maulindar, “Diagnosis Penyakit Hepatitis Menggunakan Fuzzy K-NN dan Ensemble Learning,” pp. 677–682, 2023.

[19] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[20] N. H. Cahyana, Y. Fauziah, W. Wisnalmawati, A. S. Aribowo, and S. Saifullah, “The Evaluation of Effects of Oversampling and Word Embedding on Sentiment Analysis,” *J. Infotel*, vol. 17, no. 1, pp. 54–67, 2025, doi: 10.20895/infotel.v17i1.1077.

[21] V. Ren, E. Stella, C. Patruno, A. Capurso, G. Dimauro, and R. Maglietta, “applied sciences Learning Analytics : Analysis of Methods for Online Assessment,” *Appl. Sci.*, vol. 12, no. 18, pp. 1–10, 2022.

[22] K. Jawad, M. A. Shah, and M. Tahir, “Students’ Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing,” *Sustain.*, vol. 14, no. 22, 2022, doi: 10.3390/su142214795.

[23] M. Y. Putra and D. I. Putri, “Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI,” *J. Tekno Kompak*, vol. 16, no. 2, pp. 176–176, 2022.

[24] Cumel, D. Zamri, Rahmaddeni, and Syamsurizal, “Perbandingan Metode Data Mining untuk Prediksi Banjir dengan Algoritma Naïve Bayes dan KNN,” *SENTIMAS Semin. Nas. Penelit. dan …*, pp. 40–48, 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas/article/view/353%0Ahttps://journal.irpi.or.id/index.php/sentimas/article/download/353/132

[25] M. Sholeh, D. Andayati, and Y. Rachmawati, “Data Mining Model Klasifikasi Menggunakan K-Nearest Neighbor With Normalization For Diabetes Prediction,” *TeIka*, vol. 12, no. 1, pp. 77–87, 2022.

[26] I. Binanto, N. F. Sianipar, F. Dea, M. N. Primadani, and T. W. Kartikasari, “Klasifikasi Senyawa Keladi Tikus Menggunakan Algoritma Knn, Gaussian NaãVe Bayes Dengan Menerapkan Imbalance Data Borderline Smote,” *Pros. Sains Nas. dan Teknol.*, vol. 13, no. 1, pp. 377–383, 2023, doi: 10.36499/psnst.v13i1.9005.

[27] M. M. Rofi, F. A. Setiawan, and F. Riana, “Perbandingan Metode K-Nn Dan Random Forest Pada Klasifikasi Mahasiswa Berpotensi Dropout,” *INFOTECH J.*, vol. 10, no. 1, pp. 84–89, 2024, doi: 10.31949/infotech.v10i1.8856.

[28] A. M. Shetty, M. F. Aljunid, D. H. Manjaiah, and A. M. S. Shaik Afzal, “Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis,” *Lect. Notes Networks Syst.*, vol. 821, no. May, pp. 451–474, 2024, doi: 10.1007/978-981-99-7814-4_36.

[29] H. Yuliana, S. Basuki, M. R. Hidayat, and A. Charisma, “Hyperparameter Optimization of Random Forest Algorithm to Enhance Performance Metric Evaluation of 5G Coverage Prediction,” vol. 22, no. 1, pp. 75–90, 2024.

[30] D. El-Shahat, A. Tolba, M. Abouhawwash, and M. Abdel-Basset, “Machine learning and deep learning models based grid search cross validation for short-term solar irradiance forecasting,” *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00991-w.

[31] G. Saranya and A. Pravin, “Grid Search based Optimum Feature Selection by Tuning hyperparameters for Heart Disease Diagnosis in Machine learning,” *Open Biomed. Eng. J.*, vol. 17, no. 1, pp. 1–13, 2024, doi: 10.2174/18741207-v17-e230510-2022-ht28-4371-8.

[32] M. Maulidah, Windu Gata, Rizki Aulianita, and Cucu Ika Agustyaningrum, “Algoritma Klasifikasi Decision Tree Untuk Rekomendasi Buku Berdasarkan Kategori Buku,” *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 89–96, 2020, doi: 10.51903/e-bisnis.v13i2.251.