Jurnal Teknik dan Science
JTS Vol 5 No. 2 June 2026 | ISSN: 2828-7002 (Print), ISSN: 2828-6871 (online), Hal 73-87

Optimasi Hyperparameter Random Forest menggunakan Bayesian Optimization Untuk Prediksi Kelulusan Mahasiswa

Ana Vivtia Setyawan[1], Indra Adi Permana[2]
Universitas Gunadarma

Article History
Received : 2026-03-07
Revised : 2026-04-23
Accepted : 2026-05-29
Published : 2026-07-19

Corresponding author*:
vivtia@staff.gunadarma.ac.id

Cite This Article: Ana Vivtia Setyawan, & Indra Adi Permana. (2026). Optimasi Hyperparameter Random Forest menggunakan Bayesian Optimization Untuk Prediksi Kelulusan Mahasiswa. Jurnal Teknik Dan Science, 5(2), 73–87.

DOI:
https://doi.org/10.56127/jts.v5i2.2946

Abstract: The Timely student graduation is one of the key indicators used to evaluate higher education performance. Early identification of students at risk of delayed graduation enables universities to implement appropriate academic interventions and improve student success rates. This study aims to develop a student graduation prediction model using the Random Forest algorithm and to optimize its hyperparameters through Bayesian Optimization. A synthetic dataset consisting of 750 student records was employed, incorporating demographic, socioeconomic, and academic variables. The research methodology included data preprocessing, one-hot encoding, stratified data splitting with an 80:20 ratio for training and testing sets, baseline Random Forest model development, Bayesian hyperparameter optimization using the Tree-structured Parzen Estimator (TPE) with 20 optimization trials, and model evaluation using accuracy, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and average precision. Experimental results showed that the baseline Random Forest achieved an accuracy of 70.67%, precision of 75.00%, recall of 71.43%, F1-score of 73.17%, and ROC-AUC of 82.74%. Bayesian Optimization identified the optimal hyperparameter configuration consisting of 350 trees, a maximum tree depth of six, a minimum split size of seven samples, and a minimum leaf size of four samples. Although the optimized model produced identical accuracy and F1-score values, it improved the ROC-AUC to 83.66% and the average precision to 87.68%, indicating better probability discrimination between the two classes. Feature importance analysis revealed that the number of failed courses, first-semester GPA, cumulative GPA after two semesters, attendance rate, and repeated courses were the most influential predictors. These findings demonstrate that Bayesian Optimization enhances the probabilistic ranking capability of Random Forest and can support the development of an early warning system for identifying students at risk of delayed graduation. However, further validation using real-world academic data is required before practical implementation.

Keywords: Bayesian Optimization; Early Warning System; Machine Learning; Random Forest; Student Graduation Prediction

PENDAHULUAN

Perguruan tinggi memiliki tanggung jawab untuk memastikan proses pembelajaran dapat diselesaikan mahasiswa sesuai dengan masa studi yang telah ditetapkan. Keberhasilan mahasiswa dalam menyelesaikan studi tidak hanya menjadi indikator capaian individu, tetapi juga mencerminkan efektivitas proses akademik, kualitas layanan pendidikan, serta kinerja institusi. Sebaliknya, keterlambatan kelulusan dan kegagalan menyelesaikan studi dapat menimbulkan berbagai konsekuensi, seperti meningkatnya beban biaya pendidikan, penggunaan sumber daya institusi yang kurang efisien, serta menurunnya tingkat retensi dan kelulusan perguruan tinggi. Oleh karena itu, identifikasi

dini terhadap mahasiswa yang berpotensi mengalami keterlambatan kelulusan diperlukan agar perguruan tinggi dapat memberikan intervensi akademik secara tepat waktu.

Perkembangan sistem informasi akademik memungkinkan perguruan tinggi menyimpan data mahasiswa dalam jumlah besar, meliputi data demografis, riwayat penerimaan, indeks prestasi semester, indeks prestasi kumulatif, jumlah satuan kredit semester yang telah ditempuh, kehadiran, status pembayaran, pengulangan mata kuliah, dan durasi studi. Data tersebut tidak hanya berfungsi sebagai dokumentasi administratif, tetapi juga dapat dimanfaatkan untuk menemukan pola yang berkaitan dengan keberhasilan akademik mahasiswa. Pemanfaatan data pendidikan melalui educational data mining dan machine learning memungkinkan institusi mengembangkan model prediktif yang dapat mendukung pengambilan keputusan berbasis data (Batool et al., 2023; Chaka, 2022).

Educational data mining merupakan bidang yang mengembangkan metode untuk mengeksplorasi data yang berasal dari lingkungan pendidikan. Salah satu penerapannya adalah memprediksi performa akademik, risiko putus studi, retensi, dan kelulusan mahasiswa. Kajian literatur menunjukkan bahwa prediksi performa mahasiswa merupakan salah satu topik yang paling banyak dikembangkan dalam educational data mining. Data akademik, seperti nilai mata kuliah, indeks prestasi, jumlah kredit yang diperoleh, dan riwayat pengambilan mata kuliah, sering ditemukan sebagai prediktor penting terhadap keberhasilan studi mahasiswa (Batool et al., 2023; Hellas et al., 2018; Yağcı, 2022).

Prediksi kelulusan mahasiswa dapat dirumuskan sebagai permasalahan klasifikasi dengan keluaran berupa kategori tertentu, misalnya lulus tepat waktu dan tidak lulus tepat waktu. Pada konteks yang lebih luas, status mahasiswa juga dapat dibedakan menjadi lulus, masih aktif, dan putus studi. Realinho et al. (2022) mengembangkan dataset pendidikan tinggi yang memformulasikan prediksi keberhasilan akademik sebagai klasifikasi tiga kelas, yaitu graduate, enrolled, dan dropout. Dataset tersebut memuat karakteristik mahasiswa pada saat pendaftaran serta hasil akademik pada semester pertama dan kedua. Penelitian tersebut menunjukkan bahwa data yang tersimpan pada sistem pendidikan dapat digunakan untuk mendukung identifikasi mahasiswa yang berisiko mengalami kegagalan akademik.

Berbagai algoritma klasifikasi telah diterapkan dalam prediksi performa dan kelulusan mahasiswa, seperti Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Gradient Boosting, dan Random Forest. Di antara metode tersebut, Random Forest memiliki sejumlah keunggulan karena mampu mengolah data dengan jumlah variabel yang relatif besar, menangani hubungan nonlinier, mengurangi risiko overfitting melalui pembentukan banyak pohon keputusan, dan menghasilkan ukuran kepentingan variabel. Random Forest membentuk sejumlah pohon keputusan berdasarkan sampel dan subset fitur yang dipilih secara acak, kemudian menentukan hasil klasifikasi berdasarkan agregasi prediksi setiap pohon (Breiman, 2001).

Pemanfaatan Random Forest dalam analisis pendidikan telah menunjukkan hasil yang menjanjikan. Hardman et al. (2013) menggunakan Random Forest untuk memprediksi perkembangan mahasiswa dalam pendidikan tinggi dan menunjukkan bahwa metode tersebut dapat digunakan untuk mengidentifikasi faktor-faktor yang berhubungan dengan keberlanjutan studi. Beaulac dan Rosenthal (2019) juga menggunakan Random Forest untuk memprediksi keberhasilan mahasiswa dalam memperoleh gelar berdasarkan mata kuliah yang diselesaikan pada semester awal. Selain menghasilkan prediksi, Random Forest dapat menyediakan informasi mengenai variabel yang paling berpengaruh terhadap status akademik mahasiswa. Penelitian lainnya menemukan bahwa Random Forest dapat digunakan untuk memperkirakan nilai akhir dan performa mata kuliah mahasiswa berdasarkan data transkrip akademik (Nachouki et al., 2023).

Meskipun Random Forest memiliki kemampuan klasifikasi yang baik, performanya sangat dipengaruhi oleh pemilihan nilai hyperparameter. Beberapa hyperparameter penting dalam Random Forest mencakup jumlah pohon (n_estimators), kedalaman maksimum setiap pohon (max_depth), jumlah minimum data untuk membentuk percabangan (min_samples_split), jumlah minimum data pada daun (min_samples_leaf), dan jumlah fitur yang dipertimbangkan dalam setiap pemisahan (max_features). Kombinasi hyperparameter yang tidak tepat dapat menghasilkan model yang terlalu sederhana, mengalami underfitting, atau terlalu kompleks sehingga kurang mampu melakukan generalisasi terhadap data baru. Probst et al. (2019) menunjukkan bahwa pemilihan hyperparameter merupakan bagian penting dalam pembentukan model Random Forest dan dapat memberikan pengaruh nyata terhadap performa prediksi.

Penentuan hyperparameter sering dilakukan menggunakan pengaturan bawaan, pencarian manual, grid search, atau random search. Pengaturan bawaan belum tentu memberikan hasil terbaik karena karakteristik setiap dataset berbeda. Pencarian manual sangat bergantung pada pengalaman peneliti, sedangkan grid search mengevaluasi seluruh kombinasi parameter yang telah ditentukan sehingga memerlukan waktu komputasi yang besar ketika ruang pencarian semakin luas. Bergstra dan Bengio (2012) menunjukkan bahwa random search dapat lebih efisien daripada grid search, karena tidak seluruh hyperparameter memiliki pengaruh yang sama terhadap kinerja model. Namun, random search tetap melakukan pemilihan kombinasi secara acak tanpa memanfaatkan informasi dari hasil evaluasi sebelumnya.

Salah satu pendekatan yang dapat digunakan untuk meningkatkan efisiensi pencarian hyperparameter adalah Bayesian Optimization. Metode ini memperlakukan proses evaluasi model sebagai fungsi objektif yang mahal untuk dihitung. Bayesian Optimization membangun surrogate model berdasarkan kombinasi hyperparameter yang telah dievaluasi, kemudian menggunakan acquisition function untuk menentukan kombinasi berikutnya yang paling menjanjikan. Proses tersebut menyeimbangkan eksplorasi terhadap wilayah hyperparameter yang belum banyak diuji dan eksploitasi terhadap wilayah yang telah menunjukkan performa baik. Dengan mekanisme ini, Bayesian Optimization berpotensi memperoleh konfigurasi hyperparameter yang optimal dengan jumlah evaluasi lebih sedikit dibandingkan pencarian kombinasi secara menyeluruh (Shahriari et al., 2016; Snoek et al., 2012).

Bayesian Optimization telah digunakan secara luas untuk optimasi hyperparameter berbagai algoritma machine learning. Snoek et al. (2012) menunjukkan bahwa pendekatan tersebut mampu mencapai atau melampaui hasil optimasi yang dilakukan secara manual pada sejumlah permasalahan pembelajaran mesin. Feurer et al. (2015) juga menjelaskan bahwa optimasi hyperparameter secara otomatis dapat mengurangi ketergantungan terhadap proses eksperimental manual sekaligus meningkatkan kualitas model. Selain itu, sistem optimasi berbasis Bayesian seperti SMAC dapat menangani ruang pencarian yang memuat hyperparameter numerik, kategoris, dan bersyarat, sehingga sesuai untuk algoritma Random Forest yang memiliki berbagai jenis parameter (Hutter et al., 2011; Lindauer et al., 2022).

Penelitian tentang prediksi performa akademik telah banyak membandingkan algoritma klasifikasi, tetapi sebagian penelitian masih menggunakan konfigurasi bawaan atau prosedur optimasi sederhana. Kondisi tersebut menyebabkan performa Random Forest yang dilaporkan belum tentu mencerminkan kemampuan terbaik algoritma pada dataset yang digunakan. Selain itu, penelitian prediksi mahasiswa sering kali lebih berfokus pada perbandingan nilai akurasi beberapa algoritma, sedangkan pembahasan mengenai efisiensi dan pengaruh optimasi hyperparameter belum dilakukan secara mendalam. Villar dan de

Santana (2024), misalnya, menunjukkan pentingnya evaluasi model klasifikasi pada data pendidikan yang tidak seimbang, sementara penelitian prediksi kelulusan dan retensi berskala besar memperlihatkan bahwa kualitas model sangat bergantung pada pemilihan fitur, karakteristik data, dan konfigurasi algoritma yang digunakan (Okoye et al., 2024).

Berdasarkan permasalahan tersebut, diperlukan penelitian yang mengintegrasikan algoritma Random Forest dengan Bayesian Optimization untuk memprediksi kelulusan mahasiswa. Bayesian Optimization digunakan untuk mencari kombinasi hyperparameter Random Forest yang memberikan performa terbaik berdasarkan fungsi objektif dan prosedur validasi yang telah ditentukan. Kinerja model hasil optimasi selanjutnya perlu dibandingkan dengan Random Forest menggunakan hyperparameter bawaan agar peningkatan yang dihasilkan dapat diukur secara objektif.

Evaluasi model tidak cukup dilakukan hanya menggunakan akurasi, terutama apabila jumlah mahasiswa pada setiap kelas tidak seimbang. Oleh karena itu, evaluasi perlu melibatkan metrik precision, recall, F1-score, dan area under the receiver operating characteristic curve atau AUC. Penggunaan beberapa metrik memungkinkan peneliti mengevaluasi kemampuan model dalam mengidentifikasi mahasiswa yang berpotensi tidak lulus tepat waktu, bukan hanya menghitung jumlah prediksi yang benar secara keseluruhan. Validasi silang juga diperlukan untuk mengurangi ketergantungan hasil terhadap satu pembagian data pelatihan dan pengujian.

Penelitian berjudul "Optimasi Hyperparameter Random Forest Menggunakan Bayesian Optimization untuk Prediksi Kelulusan Mahasiswa" diharapkan menghasilkan model klasifikasi yang lebih akurat, efisien, dan mampu melakukan generalisasi terhadap data mahasiswa yang belum pernah diproses sebelumnya. Model yang dikembangkan dapat menjadi dasar sistem peringatan dini bagi program studi dan perguruan tinggi dalam mengidentifikasi mahasiswa yang membutuhkan pendampingan akademik. Informasi mengenai variabel yang berpengaruh terhadap prediksi juga dapat dimanfaatkan untuk merancang intervensi, seperti bimbingan akademik, pengaturan beban studi, pengulangan mata kuliah, dan evaluasi perkembangan mahasiswa. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi metodologis dalam optimasi algoritma klasifikasi, tetapi juga memberikan manfaat praktis bagi pengelolaan pendidikan tinggi berbasis data.

METODE PENELITIAN

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional. Pendekatan kuantitatif digunakan karena penelitian memanfaatkan data akademik mahasiswa dalam bentuk numerik dan kategorikal untuk membangun model prediksi kelulusan. Metode eksperimen komputasional dilakukan dengan mengembangkan model klasifikasi menggunakan algoritma Random Forest, kemudian mengoptimalkan hyperparameter model menggunakan Bayesian Optimization.

Penelitian ini bertujuan untuk mengetahui kemampuan algoritma Random Forest dalam memprediksi kelulusan mahasiswa serta menganalisis pengaruh optimasi hyperparameter terhadap peningkatan kinerja model. Model Random Forest yang telah dioptimasi dibandingkan dengan Random Forest menggunakan pengaturan hyperparameter bawaan atau default.

Eksperimen dilakukan dengan menggunakan dua skenario pemodelan. Skenario pertama menggunakan Random Forest tanpa optimasi hyperparameter sebagai model dasar atau baseline. Skenario kedua menggunakan Random Forest dengan hyperparameter terbaik yang diperoleh melalui Bayesian Optimization. Kedua model dievaluasi

menggunakan data pengujian yang sama agar hasil perbandingan dapat dilakukan secara objektif.

Sumber dan Jenis Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari sistem informasi akademik perguruan tinggi. Data tersebut berisi informasi mengenai karakteristik dan riwayat akademik mahasiswa yang telah memiliki status kelulusan.

Unit analisis dalam penelitian ini adalah mahasiswa program sarjana. Data mahasiswa yang digunakan harus memiliki informasi akademik yang lengkap dan status kelulusan yang dapat ditentukan. Mahasiswa yang masih aktif dan belum mencapai batas waktu kelulusan tidak digunakan sebagai data penelitian, kecuali statusnya telah ditentukan berdasarkan periode pengamatan tertentu.

Data yang digunakan dapat meliputi data demografis dan data akademik mahasiswa, seperti jenis kelamin, usia masuk, jalur penerimaan, asal sekolah, indeks prestasi semester, indeks prestasi kumulatif, jumlah satuan kredit semester yang telah ditempuh, jumlah mata kuliah tidak lulus, jumlah pengulangan mata kuliah, kehadiran, status pembayaran, status cuti, dan status kelulusan mahasiswa.

Variabel Penelitian

Variabel penelitian terdiri atas variabel prediktor dan variabel target. Variabel prediktor merupakan atribut yang digunakan oleh model untuk memprediksi kelulusan mahasiswa. Variabel target merupakan kelas yang menjadi hasil prediksi.

Variabel target dalam penelitian ini dibagi menjadi dua kelas, yaitu:

1. **Lulus tepat waktu**, yaitu mahasiswa yang menyelesaikan masa studi maksimal delapan semester.
2. **Tidak lulus tepat waktu**, yaitu mahasiswa yang menyelesaikan studi lebih dari delapan semester, belum lulus sampai batas waktu pengamatan, atau mengalami putus studi.

Variabel prediktor yang dapat digunakan disajikan pada Tabel 1.

Tabel 1. Variabel penelitian

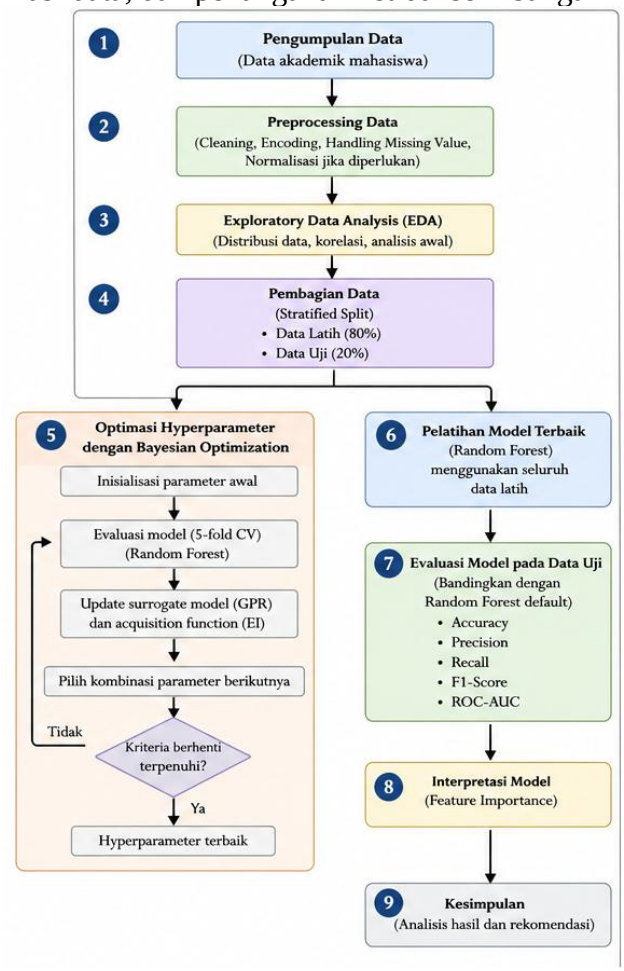
No.	Variabel	Jenis Data	Keterangan
1	Jenis kelamin	Kategorikal	Laki-laki atau perempuan
2	Usia saat masuk	Numerik	Usia mahasiswa saat diterima
3	Jalur masuk	Kategorikal	Reguler, prestasi, transfer, atau lainnya
4	Asal sekolah	Kategorikal	SMA, SMK, MA, atau lainnya
5	IP semester 1	Numerik	Indeks prestasi semester pertama
6	IP semester 2	Numerik	Indeks prestasi semester kedua
7	IPK	Numerik	Indeks prestasi kumulatif
8	SKS ditempuh	Numerik	Jumlah SKS yang telah ditempuh
9	Mata kuliah tidak lulus	Numerik	Jumlah mata kuliah yang tidak lulus
10	Pengulangan mata kuliah	Numerik	Jumlah mata kuliah yang diulang
11	Persentase kehadiran	Numerik	Tingkat kehadiran mahasiswa
12	Status pembayaran	Kategorikal	Lancar atau memiliki tunggakan
13	Status cuti	Kategorikal	Pernah atau tidak pernah cuti

No.	Variabel	Jenis Data	Keterangan
14	Status kelulusan	Kategorikal	Lulus tepat waktu atau tidak tepat waktu

Pemilihan variabel akhir disesuaikan dengan ketersediaan data pada sistem informasi akademik. Variabel yang dapat menyebabkan kebocoran data, seperti tanggal kelulusan atau lama studi akhir, tidak digunakan sebagai prediktor apabila model ditujukan untuk prediksi dini.

Tahapan Prapemrosesan Data

Prapemrosesan data dilakukan untuk meningkatkan kualitas dataset sebelum digunakan dalam proses pemodelan. Tahapan prapemrosesan meliputi pembersihan data, penanganan nilai hilang, transformasi data, dan penanganan ketidakseimbangan kelas.



Gambar 1. Alur Penelitian

Optimasi Hyperparameter Menggunakan Bayesian Optimization

Bayesian Optimization digunakan untuk mencari kombinasi hyperparameter Random Forest yang menghasilkan nilai performa terbaik. Metode ini tidak mengevaluasi seluruh kemungkinan kombinasi parameter secara langsung, tetapi menggunakan hasil evaluasi sebelumnya untuk menentukan kombinasi berikutnya yang paling menjanjikan.

Optimasi dilakukan dengan membangun surrogate model yang memperkirakan hubungan antara kombinasi hyperparameter dan nilai fungsi objektif. Acquisition function

kemudian digunakan untuk menentukan kombinasi hyperparameter yang akan dievaluasi pada iterasi berikutnya.

Fungsi objektif yang digunakan adalah rata-rata F1-score dari hasil stratified 5-fold cross-validation. F1-score digunakan karena dapat mempertimbangkan keseimbangan antara precision dan recall.

Hyperparameter yang dioptimasi disajikan pada Tabel 2.

Tabel 2. Variabel Hyperparameter

Hyperparameter	Rentang Nilai
n_estimators	100–1.000
max_depth	3–50 atau None
min_samples_split	2–20
min_samples_leaf	1–10
max_features	sqrt, log2, atau 0,3–1,0
criterion	gini atau entropy
bootstrap	True atau False
class_weight	None atau balanced

Proses Bayesian Optimization dilakukan melalui tahapan berikut:

1. Menentukan ruang pencarian hyperparameter.
2. Menentukan kombinasi parameter awal.
3. Melatih Random Forest menggunakan kombinasi parameter yang dipilih.
4. Mengevaluasi model menggunakan stratified 5-fold cross-validation.
5. Menghitung nilai fungsi objektif.
6. Memperbarui surrogate model berdasarkan hasil evaluasi.
7. Memilih kombinasi parameter berikutnya menggunakan acquisition function.
8. Mengulangi proses sampai jumlah iterasi maksimum tercapai.
9. Menentukan kombinasi hyperparameter terbaik.

HASIL DAN PEMBAHASAN

Karakteristik Dataset

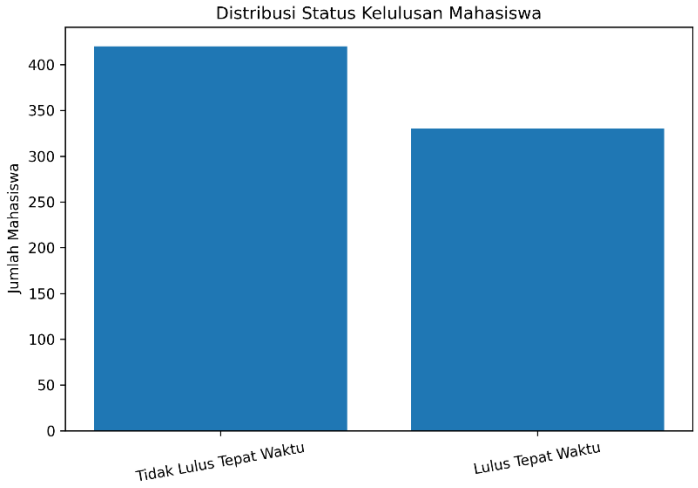
Dataset yang digunakan merupakan dataset sintetis sebanyak 750 mahasiswa dengan 18 variabel. Setelah variabel identitas, status kelulusan, dan target dikeluarkan, sebanyak 15 variabel digunakan sebagai prediktor dalam proses klasifikasi.

Berdasarkan distribusi kelas, terdapat 330 mahasiswa atau 44,00% yang lulus tepat waktu dan 420 mahasiswa atau 56,00% yang tidak lulus tepat waktu.

Tabel 3. Distribusi status kelulusan mahasiswa

Status kelulusan	Jumlah	Persentase
Lulus tepat waktu	330	44,00%
Tidak lulus tepat waktu	420	56,00%
Total	**750**	**100,00%**

Distribusi kelas menunjukkan adanya perbedaan jumlah antarkategori, tetapi ketidakseimbangannya belum tergolong ekstrem. Kondisi tersebut tetap perlu diperhatikan agar evaluasi model tidak hanya menggunakan accuracy, tetapi juga precision, recall, F1-score, dan ROC-AUC.



Gambar 2. Distribusi status kelulusan mahasiswa

Berdasarkan Gambar 2, jumlah mahasiswa yang tidak lulus tepat waktu lebih besar dibandingkan mahasiswa yang lulus tepat waktu. Oleh karena itu, pembagian data dilakukan menggunakan stratified sampling agar proporsi kedua kelas tetap terjaga.

Hasil Prapemrosesan Data

Prapemrosesan dilakukan dengan menghapus variabel ID mahasiswa, status kelulusan, dan target dari data prediktor. Variabel numerik diproses menggunakan imputasi median, sedangkan variabel kategorikal diproses menggunakan imputasi modus dan one-hot encoding.

Dataset kemudian dibagi menjadi 80% data pelatihan dan 20% data pengujian. Sebanyak 600 data digunakan untuk pelatihan dan 150 data digunakan untuk pengujian.

Tabel 3. Pembagian dataset

Bagian data	Jumlah	Persentase
Data pelatihan	600	80,00%
Data pengujian	150	20,00%
Total	**750**	**100,00%**

Kelas positif ditetapkan sebagai mahasiswa yang berisiko tidak lulus tepat waktu. Dengan demikian, recall menunjukkan kemampuan model dalam mendeteksi mahasiswa yang membutuhkan perhatian akademik.

Hasil Random Forest Default

Model Random Forest default digunakan sebagai model awal atau baseline. Hasil pengujian menunjukkan accuracy sebesar 0,7067, precision sebesar 0,7500, recall sebesar 0,7143, F1-score sebesar 0,7317, dan ROC-AUC sebesar 0,8274.

Tabel 4. Hasil Random Forest default

Metrik	Nilai
Accuracy	0,7067
Precision	0,7500
Recall	0,7143
F1-score	0,7317
ROC-AUC	0,8274

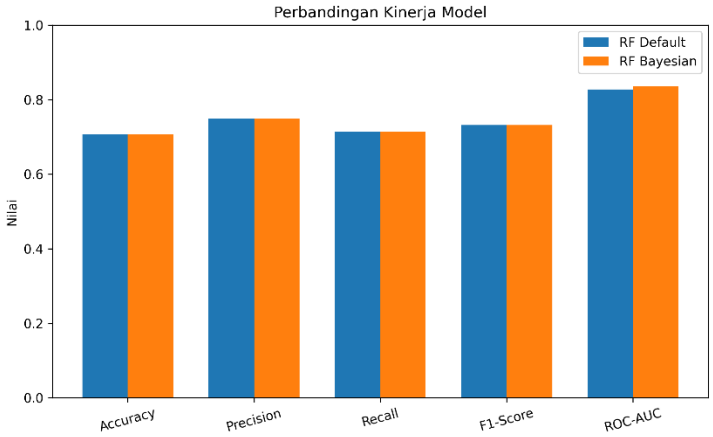
Average precision 0,8660

Nilai accuracy menunjukkan bahwa model mampu mengklasifikasikan dengan benar sekitar 70,67% data pengujian. Recall sebesar 71,43% menunjukkan bahwa sebagian besar mahasiswa berisiko telah berhasil dideteksi.

Confusion matrix Random Forest default menghasilkan 46 true negative, 20 false positive, 24 false negative, dan 60 true positive.

Tabel 5. Confusion matrix Random Forest default

Aktual/Prediksi	Tidak berisiko	Berisiko
Tidak berisiko	46	20
Berisiko	24	60



Gambar 3. Confusion matrix Random Forest default

Berdasarkan Gambar 3, model berhasil mendeteksi 60 dari 84 mahasiswa yang benar-benar berisiko tidak lulus tepat waktu. Namun, masih terdapat 24 mahasiswa berisiko yang tidak terdeteksi oleh model.

Hasil Bayesian Optimization

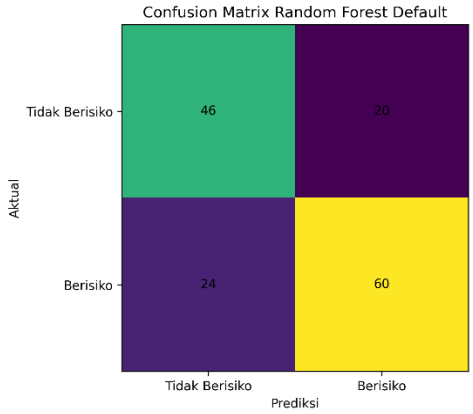
Bayesian Optimization dilakukan sebanyak 20 percobaan menggunakan 3-fold cross-validation. Fungsi objektif yang dimaksimalkan adalah F1-score.

Kombinasi hyperparameter terbaik menghasilkan F1-score validasi silang sebesar 0,7763.

Tabel 6. Hyperparameter terbaik

Hyperparameter	Nilai terbaik
n_estimators	350
max_depth	6
min_samples_split	7
min_samples_leaf	4
max_features	sqrt
criterion	entropy
bootstrap	False
class_weight	None

Hyperparameter tersebut menunjukkan bahwa model terbaik menggunakan 350 pohon dengan kedalaman maksimum enam. Pembatasan kedalaman dan jumlah minimum sampel pada percabangan membantu mengurangi kompleksitas model.



Gambar 4. Konvergensi Bayesian Optimization

Berdasarkan Gambar 4, nilai F1-score validasi berubah pada setiap iterasi hingga ditemukan kombinasi hyperparameter terbaik. Proses optimasi membutuhkan waktu sekitar 16,13 detik.

Hasil Random Forest dengan Bayesian Optimization

Model Random Forest kemudian dilatih kembali menggunakan hyperparameter terbaik. Hasil pengujian menunjukkan accuracy sebesar 0,7067, precision sebesar 0,7500, recall sebesar 0,7143, F1-score sebesar 0,7317, dan ROC-AUC sebesar 0,8366.

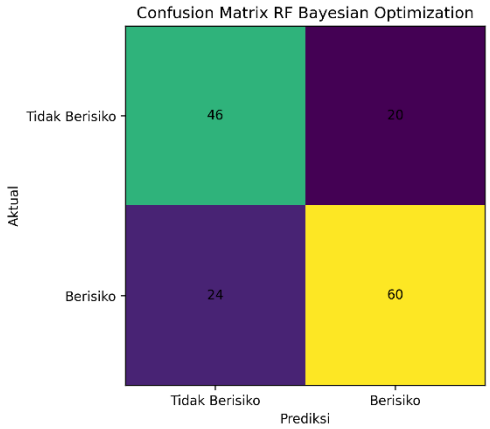
Tabel 7. Hasil Random Forest dengan Bayesian Optimization

Metrik	Nilai
Accuracy	0,7067
Precision	0,7500
Recall	0,7143
F1-score	0,7317
ROC-AUC	0,8366
Average precision	0,8768

Confusion matrix model hasil optimasi menghasilkan 46 true negative, 20 false positive, 24 false negative, dan 60 true positive.

Tabel 8. Confusion matrix model hasil optimasi

Aktual/Prediksi	Tidak berisiko	Berisiko
Tidak berisiko	46	20
Berisiko	24	60



Gambar 5. Confusion matrix Random Forest dengan Bayesian Optimization

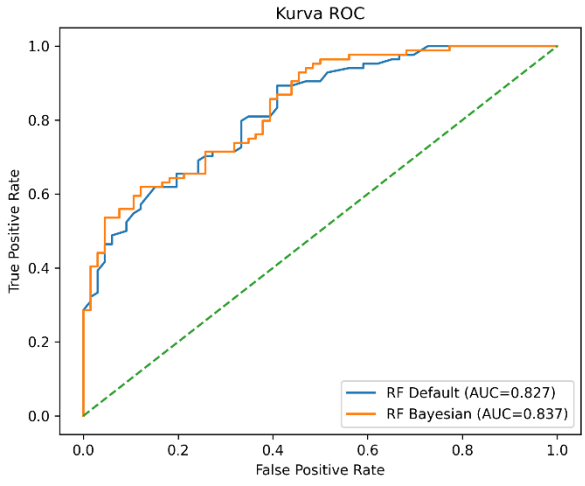
Berdasarkan Gambar 5, keputusan klasifikasi model hasil optimasi masih sama dengan model default pada ambang probabilitas 0,5. Namun, kualitas skor probabilitas mengalami peningkatan yang terlihat dari nilai ROC-AUC.

Perbandingan Kinerja Model

Perbandingan kedua model menunjukkan bahwa Bayesian Optimization belum meningkatkan accuracy, precision, recall, dan F1-score. Namun, ROC-AUC meningkat dari 0,8274 menjadi 0,8366 dan average precision meningkat dari 0,8660 menjadi 0,8768.

Tabel 9. Perbandingan kinerja model

Metrik	RF default	RF Bayesian
Accuracy	0,7067	0,7067
Precision	0,7500	0,7500
Recall	0,7143	0,7143
F1-score	0,7317	0,7317
ROC-AUC	0,8274	0,8366
Average precision	0,8660	0,8768

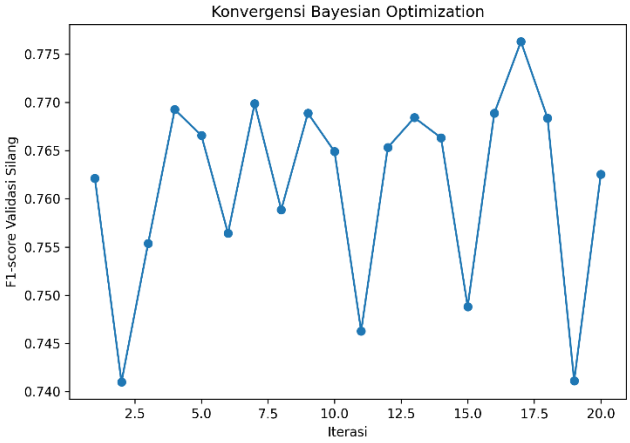


Gambar 6. Perbandingan kinerja Random Forest default dan Random Forest dengan Bayesian Optimization

Berdasarkan Gambar 6, kedua model memiliki nilai klasifikasi yang sama, tetapi model hasil optimasi memiliki kemampuan diskriminasi probabilitas yang sedikit lebih baik.

Analisis Kurva ROC

Kurva ROC digunakan untuk menilai kemampuan model dalam membedakan mahasiswa berisiko dan tidak berisiko pada berbagai nilai ambang klasifikasi.



Gambar 7. Kurva ROC Random Forest default dan Random Forest hasil optimasi

Berdasarkan Gambar 6, kurva model hasil optimasi berada sedikit di atas kurva model default. Nilai ROC-AUC sebesar 0,8366 menunjukkan bahwa model hasil optimasi memiliki kemampuan yang cukup baik dalam membedakan kedua kelas.

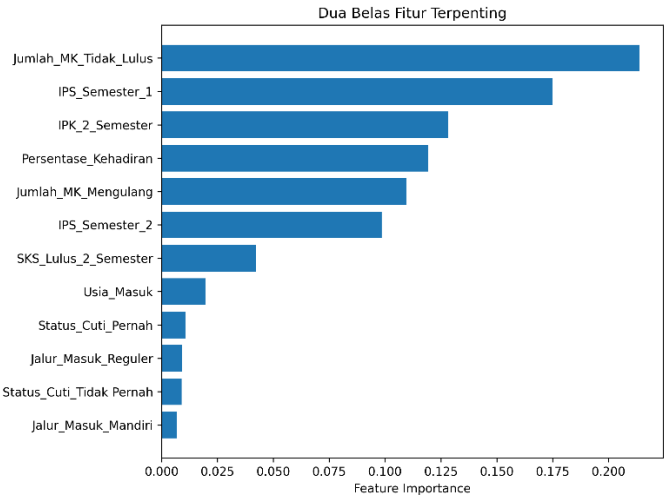
Peningkatan ROC-AUC tidak langsung meningkatkan accuracy karena accuracy dihitung menggunakan satu ambang tertentu, yaitu 0,5. Oleh karena itu, ambang klasifikasi dapat disesuaikan jika penelitian lebih memprioritaskan peningkatan recall.

Analisis Feature Importance

Feature importance digunakan untuk mengetahui variabel yang memberikan kontribusi terbesar terhadap hasil prediksi. Variabel paling penting adalah jumlah mata kuliah tidak lulus, IPS semester pertama, IPK dua semester, persentase kehadiran, dan jumlah mata kuliah mengulang.

Tabel 10. Fitur terpenting

Peringkat	Variabel	Importance
1	Jumlah mata kuliah tidak lulus	0,2138
2	IPS semester 1	0,1750
3	IPK dua semester	0,1283
4	Persentase kehadiran	0,1193
5	Jumlah mata kuliah mengulang	0,1095
6	IPS semester 2	0,0985
7	SKS lulus dua semester	0,0423
8	Usia masuk	0,0198



Gambar 8. Fitur terpenting dalam prediksi kelulusan mahasiswa

Berdasarkan Gambar 8, variabel akademik semester awal memiliki kontribusi lebih besar dibandingkan variabel demografis dan sosial ekonomi. Jumlah mata kuliah tidak lulus menjadi indikator paling dominan dalam memprediksi risiko keterlambatan kelulusan.

Pembahasan

Hasil penelitian menunjukkan bahwa Random Forest dapat digunakan untuk memprediksi risiko keterlambatan kelulusan mahasiswa. Model default menghasilkan accuracy sebesar 70,67%, F1-score sebesar 73,17%, dan ROC-AUC sebesar 82,74%.

Bayesian Optimization belum meningkatkan hasil klasifikasi pada ambang 0,5, tetapi meningkatkan ROC-AUC menjadi 83,66%. Hal tersebut menunjukkan bahwa optimasi lebih berpengaruh terhadap kualitas skor probabilitas daripada keputusan kelas akhir.

Recall sebesar 71,43% menunjukkan bahwa model mampu mendeteksi sebagian besar mahasiswa berisiko. Namun, masih terdapat mahasiswa berisiko yang tidak terdeteksi. Untuk penerapan sistem peringatan dini, ambang probabilitas dapat diturunkan agar recall meningkat, meskipun dapat menyebabkan peningkatan false positive.

Feature importance menunjukkan bahwa variabel akademik semester awal menjadi faktor paling dominan. Jumlah mata kuliah tidak lulus, IPS semester pertama, IPK, kehadiran, dan mata kuliah mengulang dapat digunakan sebagai indikator awal untuk menentukan mahasiswa yang membutuhkan pendampingan.

Secara praktis, model dapat diterapkan dalam sistem informasi akademik untuk mengelompokkan mahasiswa ke dalam tingkat risiko rendah, sedang, dan tinggi. Mahasiswa dengan risiko tinggi dapat diberikan bimbingan akademik, evaluasi beban studi, program remedial, atau pemantauan kehadiran.

Meskipun demikian, penelitian ini menggunakan dataset sintetis sehingga hasilnya belum dapat dianggap mewakili kondisi perguruan tinggi tertentu. Penelitian lanjutan perlu menggunakan data akademik nyata agar model dapat diuji secara lebih valid dan dapat diterapkan dalam kondisi sebenarnya.

KESIMPULAN

Penelitian ini berhasil membangun model prediksi kelulusan mahasiswa menggunakan algoritma Random Forest dan Bayesian Optimization pada dataset sintetis yang terdiri atas 750 mahasiswa. Model Random Forest default menghasilkan accuracy sebesar 70,67%, precision sebesar 75,00%, recall sebesar 71,43%, F1-score sebesar 73,17%, dan ROC-AUC sebesar 82,74%.

Hasil Bayesian Optimization memperoleh kombinasi hyperparameter terbaik berupa 350 pohon, kedalaman maksimum enam, minimum tujuh sampel untuk percabangan, minimum empat sampel pada daun, penggunaan max_features sebesar sqrt, kriteria entropy, serta tanpa bootstrap. Model hasil optimasi menghasilkan accuracy, precision, recall, dan F1-score yang sama dengan model default, tetapi meningkatkan ROC-AUC menjadi 83,66% dan average precision menjadi 87,68%.

Hasil tersebut menunjukkan bahwa Bayesian Optimization belum meningkatkan keputusan klasifikasi pada ambang probabilitas 0,5, tetapi mampu meningkatkan kemampuan model dalam membedakan mahasiswa berisiko dan tidak berisiko berdasarkan skor probabilitas. Variabel yang memberikan kontribusi terbesar terhadap prediksi adalah jumlah mata kuliah tidak lulus, IPS semester pertama, IPK dua semester, persentase kehadiran, jumlah mata kuliah mengulang, dan IPS semester kedua.

Model yang dikembangkan berpotensi digunakan sebagai dasar sistem peringatan dini untuk membantu perguruan tinggi mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu. Namun, karena penelitian ini menggunakan dataset sintetis, hasil penelitian belum dapat digeneralisasikan pada kondisi perguruan tinggi tertentu. Penelitian selanjutnya disarankan menggunakan data akademik nyata dari beberapa angkatan dan program studi, melakukan optimasi ambang klasifikasi, serta membandingkan Bayesian Optimization dengan metode optimasi hyperparameter lainnya..

DAFTAR PUSTAKA

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28, 905–971. https://doi.org/10.1007/s10639-022-11152-y

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. Research in Higher Education, 60(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chaka, C. (2022). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews. Journal of e-Learning and Knowledge Society, 18(2), 1–14. https://doi.org/10.20368/1971-8829/1135578

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 28, pp. 2962–2970). Curran Associates.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting students' progression in higher education by using the random forest algorithm. Systems Research and Behavioral Science, 30(2), 194–203. https://doi.org/10.1002/sres.2130

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education (pp. 175–199). Association for Computing Machinery. https://doi.org/10.1145/3293881.329578_predicting

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In C. A. C. Coello (Ed.), Learning and intelligent optimization (pp. 507–523). Springer. https://doi.org/10.1007/978-3-642-25566-3_40

Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C., Ruhkopf, T., Sass, R., & Hutter, F. (2022). SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. Journal of Machine Learning Research, 23(54), 1–9.

Nachouki, M., Mohamed, E. A., Mehdi, R., & Abou Naaj, M. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. Trends in Neuroscience and Education, 33, 100214. https://doi.org/10.1016/j.tine.2023.100214

Okoye, K., Nganji, J. T., & Hosseini, S. (2024). A machine learning model and ensemble algorithm for predicting students' retention and graduation status in education. Decision Analytics Journal, 10, 100383. https://doi.org/10.1016/j.dajour.2023.100383

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3), Article e1301. https://doi.org/10.1002/widm.1301

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), Article 146. https://doi.org/10.3390/data7110146

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in neural information processing systems (Vol. 25, pp. 2951–2959). Curran Associates.

Villar, A., & de Santana, Á. L. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. Discover Artificial Intelligence, 4, Article 2. https://doi.org/10.1007/s44163-023-00079-z

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9, Article 11. https://doi.org/10.1186/s40561-022-00192-z