 ISSN:2302-8149
e-ISSN:2540-9719

Deteksi Tingkat Potensi Kelulusan Calon Mahasiswa menggunakan *Algoritma Random Forest*

Detection of Graduation Potential in Prospective Students using the Random Forest Algorithm

[1]Puguh Hasta Gunawan*, [2]Irving Vitra Paputungan
[1,2]Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia
[1,2]Jl. Kaliurang km 14.5, Sleman, Yogyakarta 55584
*e-mail: 23917022@students.uii.ac.id

(***received*:** 20 May 2025, ***revised*:** 11 June 2025, ***accepted*:** 13 June 2025)

Abstrak

Deteksi potensi kelulusan mahasiswa umumnya dilakukan dengan mengevaluasi berbagai faktor akademik dan non-akademik. Penelitian ini bertujuan untuk membangun model prediksi kelulusan mahasiswa sejak awal masa studi dengan memanfaatkan data akademik dari sekolah menengah atas, seperti nilai, kehadiran, waktu belajar, serta faktor demografis dan sosial, sehingga dapat diterapkan di lingkungan kampus untuk membantu institusi dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan. Dengan prediksi yang akurat, kampus diharapkan mampu merancang intervensi akademik yang lebih tepat sasaran, seperti pendampingan belajar, konseling, atau dukungan tambahan lainnya. Sebanyak 396 data siswa digunakan dalam penelitian ini dan diproses melalui tahap pra-pemrosesan, seperti penghapusan data tidak relevan dan pengkodean variabel kategorikal. Model dikembangkan menggunakan algoritma *Random Forest* dengan parameter *max_depth* = 15 dan *random_state* = 42. Evaluasi performa dilakukan dengan metrik akurasi, recall, F1-score, serta *ROC Curve*. Hasil menunjukkan akurasi model sebesar 89%, dengan performa deteksi kelas *Pass* memiliki *recall* 87% dan *F1-score* 91%, serta kelas *Fail* dengan *recall* 92% dan *F1-score* 84%. Selain itu, nilai *Area Under the Curve (AUC)* sebesar 0.94 menunjukkan kemampuan model yang sangat baik dalam membedakan antara mahasiswa yang berpotensi lulus dan tidak. Penelitian ini menegaskan bahwa model efektif dalam klasifikasi kelulusan mahasiswa berdasarkan data awal. Untuk pengembangan lebih lanjut, disarankan agar penelitian mencakup variabel tambahan seperti aspek psikologis, motivasi belajar, dan kondisi sosial ekonomi, serta melakukan tuning dengan menambahkan parameter lain pada model seperti *n_estimators*, *min_samples_split*, dan *max_features*, guna meningkatkan akurasi dan generalisasi model.

Kata kunci: *random forest*, *confusion matrix*, deteksi kelulusan

Abstract

Detecting students' graduation potential is commonly performed by evaluating various academic and non-academic factors. This study aims to develop a predictive model for student graduation from the beginning of their academic journey, utilizing high school academic data such as grades, attendance, study hours, as well as demographic and social factors. The goal is to enable universities to identify students who are at risk of delayed graduation. With accurate predictions, institutions are expected to design more targeted academic interventions, such as tutoring, counseling, or other forms of academic support. A total of 396 student records were used in this study and processed through a series of preprocessing steps, including the removal of irrelevant data and the encoding of categorical variables. The model was developed using the Random Forest algorithm with parameters set to max_depth = 15 and random_state = 42. Model performance was evaluated using accuracy, recall, F1-score, and the ROC curve. The results show that the model achieved an accuracy of 89%, with the Pass *class having a recall of 87% and an F1-score of 91%, and the* Fail *class showing a recall of 92% and an F1-score of 84%. Additionally, the Area Under the Curve (AUC) value of 0.94 indicates excellent model performance in distinguishing between students likely to graduate and those at risk of not graduating.*

 ISSN:2302-8149
e-ISSN:2540-9719

This study confirms that the model is effective in classifying graduation outcomes based on early academic data. For further development, it is recommended to include additional variables such as psychological factors, learning motivation, and socioeconomic conditions. Moreover, tuning the model by adding other parameters—such as n_estimators, min_samples_split, and max_features—is suggested to improve the model's accuracy and generalizability.

Keywords: *random forest, confusion matrix, graduation detection*

1 Pendahuluan

Kinerja akademik mahasiswa merupakan indikator strategis dalam menilai efektivitas sistem pendidikan tinggi, tidak hanya bagi mahasiswa secara individu, namun juga bagi institusi seperti perguruan tinggi [1]. Dalam beberapa tahun terakhir, penerapan pendekatan berbasis pembelajaran mesin telah banyak dilakukan untuk menganalisis data akademik mahasiswa guna memprediksi potensi kelulusan [2],[3]. Di sisi lain, data akademik sewaktu menempuh pendidikan menengah atas juga dapat memiliki pengaruh signifikan terhadap keberhasilan studi mahasiswa di jenjang perguruan tinggi, di antaranya meliputi nilai akademik, kehadiran, waktu belajar, dukungan keluarga, dan akses ke fasilitas pendidikan. Oleh karena itu, penting untuk memahami pola-pola yang mempengaruhi performa akademik di tingkat sekolah menengah agar potensi masalah dapat diidentifikasi sejak dini dan intervensi yang tepat dapat dilakukan. Evaluasi ini tidak hanya melibatkan hasil akademik berupa nilai, tetapi juga mencakup keterlibatan dalam kegiatan akademik, keterampilan karya ilmiah, aktivitas ekstrakurikuler, serta faktor non-akademik seperti latar belakang ekonomi dan sosial [1].

Memahami pola-pola tersebut memungkinkan institusi pendidikan tinggi untuk melakukan deteksi dini terhadap potensi keterlambatan studi mahasiswa dan melaksanakan intervensi akademik secara lebih tepat sasaran. Dari sisi institusi pendidikan tinggi, urgensi deteksi kelulusan mahasiswa ini juga terkait erat dengan pencapaian Indikator Kinerja Utama (IKU), terutama IKU 1 (Lulusan mendapatkan pekerjaan yang layak) dan IKU 2 (Mahasiswa mendapatkan pengalaman di luar kampus) [4]. Capaian IKU sangat bergantung pada keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu. Apabila mahasiswa mengalami keterlambatan, mereka tidak hanya menghadapi risiko kerugian finansial akibat bertambahnya beban biaya studi, tetapi juga kehilangan momentum karier, kepercayaan diri, dan peluang bersaing di dunia kerja. Selain itu, keterlambatan studi turut berdampak negatif terhadap citra institusi, baik dari sisi akreditasi maupun peringkat.

Berdasarkan kajian literatur yang telah dilakukan, sejumlah penelitian terdahulu dalam deteksi kelulusan mahasiswa masih berfokus pada pendekatan berbasis data akademik semata. Beberapa di antaranya menggunakan data IPK, kehadiran, dan nilai mata kuliah sebagai variabel utama [2],[3],[5], serta memanfaatkan faktor seperti hafalan dan nilai ujian dalam konteks pendidikan berbasis agama [5]. Meskipun pendekatan ini menunjukkan performa yang cukup baik dari sisi akurasi, model-model tersebut umumnya belum menghubungkan riwayat akademik pada jenjang sekolah menengah atas dengan performa akademik mahasiswa di perguruan tinggi.

Namun, pendekatan yang masih terbatas pada variabel akademik belum mampu menggambarkan keseluruhan faktor yang mempengaruhi kelulusan. Faktor-faktor non-akademik seperti kondisi sosial ekonomi, motivasi belajar, keterlibatan dalam organisasi, serta kondisi psikologis siswa masih jarang diadopsi dalam model prediktif yang ada. Selain itu, penggunaan algoritma seperti *Decision Tree*, *k-Nearest Neighbors* (k-NN), dan *Support Vector Machine* (SVM) juga memiliki keterbatasan. Misalnya, *Decision Tree* rentan terhadap *overfitting*, *k-NN* sensitif terhadap skala data dan membutuhkan waktu klasifikasi tinggi, sedangkan *SVM* sangat bergantung pada pemilihan kernel dan tidak efisien untuk dataset besar [6],[7],[8]. Sebaliknya, *Random Forest* terbukti mampu mengatasi kelemahan tersebut melalui pendekatan *ensemble* yang menggabungkan banyak pohon keputusan.

Teknik ini dapat meningkatkan stabilitas, mengurangi *overfitting*, serta bekerja lebih baik dalam menangani data berdimensi tinggi tanpa memerlukan normalisasi atau *tuning* parameter yang kompleks [6],[8]. *Random Forest* juga toleran terhadap fitur yang tidak relevan dan menghasilkan performa yang lebih *robust* di berbagai skala data [7]. Studi empiris menunjukkan bahwa algoritma ini berhasil mendeteksi performa akademik siswa lebih akurat dibandingkan metode lain [9],[10], serta efektif diterapkan pada kasus nyata seperti di Program Studi Informatika Universitas Baturaja [14].

 ISSN:2302-8149
e-ISSN:2540-9719

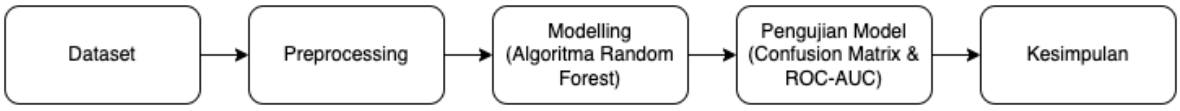
Berdasarkan latar belakang dan celah penelitian yang telah diidentifikasi, penelitian ini bertujuan untuk membangun model prediksi potensi kelulusan mahasiswa berbasis data akademik dan demografis dari jenjang sekolah menengah atas menggunakan algoritma *Random Forest Classifier* (RFC) serta melakukan evaluasi kinerja model berdasarkan metrik akurasi, recall, F1-score, dan AUC

2 Tinjauan Literatur

Algoritma *Random Forest* merupakan salah satu metode pembelajaran mesin berbasis *ensemble* yang telah banyak digunakan dalam berbagai bidang karena kemampuannya dalam menangani data kompleks, baik numerik maupun kategorikal, serta memberikan hasil klasifikasi yang stabil dan akurat. Keunggulan algoritma ini terletak pada kemampuannya mengurangi risiko *overfitting* dan meningkatkan akurasi prediksi melalui mekanisme agregasi dari banyak pohon keputusan. Dalam praktiknya, *Random Forest* telah diterapkan secara luas dengan hasil yang menjanjikan. Misalnya, pada sektor kesehatan, algoritma ini digunakan untuk memprediksi waktu tunggu pasien rawat jalan di RSJ Dr. Soeharto Heerdjan dengan akurasi mencapai 97,6%, membuktikan kemampuannya dalam mengelola data pelayanan kesehatan yang kompleks [16]. Di bidang pendidikan berbasis keagamaan, penerapannya untuk memprediksi kelulusan santri Tahfidz menunjukkan akurasi sangat tinggi sebesar 99,64%, dengan menekankan pentingnya variabel non-akademik seperti hafalan dan kehadiran [5]. Dalam sektor transportasi, algoritma ini digunakan dalam deteksi tarif penerbangan berbasis *AutoML* dengan nilai R^2 sebesar 85,87%, mengungguli metode *Logistic Regression* dan *Gradient Boosting* [17]. Di ranah pendidikan tinggi, *Random Forest* digunakan untuk memprediksi kelulusan mahasiswa di Indonesia dan berhasil mencapai akurasi 87,5%, di mana IPK dan kehadiran menjadi variabel utama [3]. Aplikasi lainnya meliputi klasifikasi kualitas air berbasis web dengan akurasi 78% [18], serta prediksi harga properti dengan MAE 2,48 dan RMSE 2,89, di mana jarak ke pusat kota menjadi determinan utama [19]. Sementara itu, algoritma *Naïve Bayes* juga digunakan dalam domain pendidikan dan mampu mencapai akurasi sebesar 77,97% dalam mendeteksi kegagalan studi siswa [20], namun performanya masih berada di bawah akurasi *Random Forest*, terutama dalam konteks prediksi pendidikan yang kompleks. Berdasarkan temuan-temuan tersebut, dapat disimpulkan bahwa *Random Forest* memiliki potensi unggul dalam membangun model deteksi kelulusan yang lebih akurat, khususnya dengan konfigurasi parameter seperti max_depth = 15 dan random_state = 42, serta menjadi alternatif yang lebih efektif dibandingkan metode klasifikasi tradisional yang memiliki keterbatasan dalam skalabilitas dan generalisasi.

3 Metode Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini sebagai berikut:



Gambar 1. Alur penelitian

Berikut penjelasan detail terkait gambar 1 diatas:

1 Dataset

a. Menggunakan pendekatan kuantitatif dengan metode survei dan analisis data sekunder. Desain penelitian yang dipilih adalah desain eksplanatori karena bertujuan untuk mengidentifikasi dan menjelaskan faktor-faktor yang mempengaruhi kelulusan mahasiswa.
b. dataset didapatkan melalui *uci edu student performance* "https://archive.ics.uci.edu/dataset/320/student+performance".

2 *Preprocessing*

Pada tahapan selanjutnya, dilakukan proses preprocessing data yang bertujuan untuk menyiapkan dataset agar dapat digunakan secara optimal oleh algoritma pembelajaran mesin. Proses ini mencakup beberapa langkah krusial, yaitu penghapusan kolom tidak relevan *(delete)*, pengkodean variabel kategorikal *(encode)*, dan pemisahan data *(split)*.

3 *Modelling*

 ISSN:2302-8149
e-ISSN:2540-9719

Menggunakan Algoritma *Random Forest* untuk mendeteksi kelulusan mahasiswa berdasarkan berbagai faktor yang mempengaruhi hasil dari data akademik mereka. *Random Forest* adalah metode *ensemble learning* yang menggabungkan beberapa *decision tree* untuk menghasilkan deteksi yang lebih akurat dan kuat terhadap *overfitting*. Setiap *decision tree* dalam *random forest* dilatih menggunakan subset acak dari data dan fitur, dan hasil dari semua tree digabungkan menggunakan teknik voting (untuk klasifikasi) atau averaging (untuk regresi).

4. Pengujian Model Menggunakan *Confusion Matrix* dan *ROC-AUC*

Confusion matrix digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi model terhadap data aktual. Untuk kasus klasifikasi biner, *confusion matrix* berbentuk tabel 2x2 yang terdiri dari empat komponen utama:

1. *True Positive* (TP): Data yang benar diklasifikasikan sebagai positif.
2. *True Negative* (TN): Data yang benar diklasifikasikan sebagai negatif.
3. *False Positive* (FP): Data negatif yang salah diklasifikasikan sebagai positif.
4. *False Negative* (FN): Data positif yang salah diklasifikasikan sebagai negatif.

Receiver Operating Characteristic (ROC) Curve digunakan untuk mengevaluasi kemampuan model dalam membedakan antara dua kelas. Kurva ini memvisualisasikan hubungan antara *True Positive Rate (TPR)* dan *False Positive Rate (FPR)* pada berbagai nilai ambang klasifikasi. Luas area di bawah kurva tersebut, dikenal sebagai *Area Under the Curve (AUC)*, digunakan sebagai indikator numerik performa model. Nilai AUC yang mendekati 1 menunjukkan kemampuan klasifikasi yang sangat baik, sedangkan nilai mendekati 0,5 mengindikasikan performa yang setara dengan tebakan acak.

4 Hasil dan Pembahasan

Pada bagian ini, akan dipaparkan hasil dari proses klasifikasi *performance* siswa menggunakan algoritma *Random Forest*, yang bertujuan untuk mendeteksi potensi kelulusan berdasarkan data akademik dan faktor pendukung lainnya. Proses analisis dimulai dengan tahapan *preprocessing* data, pembagian data menjadi subset pelatihan model dan pengujian, hingga evaluasi performa model menggunakan metrik akurasi, *precision*, *recall*, *and f1-score.*

3.1. Dataset

Dataset yang digunakan dalam penelitian ini adalah sebanyak 495 data dengan jumlah data gender untuk perempuan 187 dan untuk laki-laki 208, Berikut hasil contoh dataset yang digunakan

Tabel 1. Contoh dataset

SE	FA	PS	ME	FE	TT	ST	FAI	SC	HI	IN	RO	FE	HE	AB	G1	G2	G3
F	GT3	A	4	4	2	2	0	yes	yes	no	no	3	3	6	5	6	6
F	GT3	T	1	1	1	2	0	no	yes	yes	no	3	3	4	5	5	6
F	LE3	T	1	1	1	2	3	yes	yes	yes	no	3	3	10	7	8	10
F	GT3	T	4	2	1	3	0	no	yes	yes	yes	2	5	2	15	14	15
F	GT3	T	3	3	1	2	0	no	yes	no	no	3	5	4	6	10	10

Berikut adalah penjelasan pada tabel 1:

1. (SE) Sex : Jenis kelamin siswa: F untuk perempuan, M untuk laki-laki.
2. (FA) famsize: Ukuran keluarga: LE3 untuk keluarga kecil (3 anggota atau kurang), GT3 untuk keluarga besar (lebih dari 3 anggota).

 ISSN:2302-8149
e-ISSN:2540-9719

3. (PS) Pstatus: Status tempat tinggal orang tua: T untuk tinggal bersama, A untuk tinggal terpisah (misalnya, karena perceraian atau pekerjaan).
4. (ME)Medu: Tingkat pendidikan ibu: Skala 0-4 (0: tidak berpendidikan, 4: pendidikan tinggi seperti universitas).
5. (FE) Fedu: Tingkat pendidikan ayah: Skala 0-4 (0: tidak berpendidikan, 4: pendidikan tinggi seperti universitas).
6. (TT) *traveltime*: Waktu perjalanan ke sekolah: Skala 1-4 (1: kurang dari 15 menit, 4: lebih dari 1 jam).
7. (ST) *studytime*: Waktu belajar mingguan di luar sekolah: Skala 1-4 (1: kurang dari 2 jam, 4: lebih dari 10 jam).
8. (FAI) *failures*: Jumlah kegagalan akademik sebelumnya: Total jumlah mata pelajaran yang gagal (misalnya, 0 untuk tidak ada kegagalan).
9. (SC) *schoolsu*p: Apakah siswa menerima dukungan tambahan dari sekolah: yes untuk menerima, no untuk tidak.
10. (HI) *higher*: Apakah siswa berencana melanjutkan pendidikan ke jenjang lebih tinggi (universitas): yes untuk iya, no untuk tidak.
11. (IN) *internet*: Apakah siswa memiliki akses internet di rumah: yes untuk ya, no untuk tidak.
12. (RO) *romanti*c: Apakah siswa memiliki hubungan romantis: yes untuk ya, no untuk tidak.
13. (FE) *freetime*: Waktu luang siswa setelah sekolah: Skala 1-5 (1: sangat sedikit waktu luang, 5: banyak waktu luang).
14. (HE) *health*: Status kesehatan siswa: Skala 1-5 (1: sangat buruk, 5: sangat baik).
15. (AB) *absences*: Jumlah ketidakhadiran siswa di sekolah.
16. G1: Nilai akademik pada evaluasi periode pertama (G1) dalam skala 0-20.
17. G2: Nilai akademik pada evaluasi periode kedua (G2) dalam skala 0-20.
18. G3: Nilai akhir (G3) pada akhir periode akademik dalam skala 0-20 (biasanya digunakan untuk evaluasi performa siswa secara keseluruhan).

3.2. *Preprocessing*

Preprocessing pada penelitian ini melalui beberapa tahapan sebagai berikut:

1. *Delete* (Penghapusan Kolom Tidak Relevan)
   Pada tahap ini, kolom yang dianggap tidak memiliki kontribusi signifikan terhadap proses klasifikasi dihapus dari dataset. Contohnya, kolom seperti school_name, age, address, and G3 yang bersifat identifikasi atau administratif, tidak memiliki nilai prediktif dan berpotensi menimbulkan bias model, sehingga dieliminasi untuk menyederhanakan model.
2. *Encode* Categorical Variables (Pengkodean Variabel Kategorikal)
   Data kategorikal seperti *gender, famsize, Pstatus, schoolsup, internet, dan romantic* tidak bisa langsung diproses oleh algoritma Random Forest karena berupa teks. Oleh karena itu, dilakukan transformasi menggunakan metode *Label Encoding*, di mana setiap kategori diberi nilai numerik unik. Sebagai contoh, gender dengan nilai F dan M akan dikodekan menjadi 0 dan 1.
3. *Split Data into Features and Target* (Memisahkan Fitur dan Label)
4. Dataset kemudian dipisahkan menjadi dua bagian:
   a) Fitur (X): seluruh kolom yang digunakan sebagai input untuk prediksi, seperti *studytime, failures, absences*, G1, G2, dan hasil encoding kategorikal lainnya.
   b) Target (y): kolom label yang ingin diprediksi, yaitu G3 atau status kelulusan (Pass/Fail). Dalam penelitian ini, G3 diubah menjadi biner: jika nilai ≥ 10 dianggap "Pass" (1), dan jika < 10 dianggap "Fail" (0).
5. *Split* Data into *Train* and *Test Sets (*Membagi Data Latih dan Uji)
   Setelah data dipisahkan antara fitur dan target, data kemudian dibagi menjadi dua subset menggunakan teknik train-test split. Proporsi yang digunakan dalam penelitian ini adalah 80% untuk pelatihan dan 20% untuk pengujian. Pembagian ini dilakukan secara acak namun *reproducible* dengan menyetel parameter random_state = 42. Tujuannya adalah agar performa model dapat diuji pada data yang belum pernah dilihat sebelumnya, sehingga evaluasi menjadi objektif dan tidak overfit.

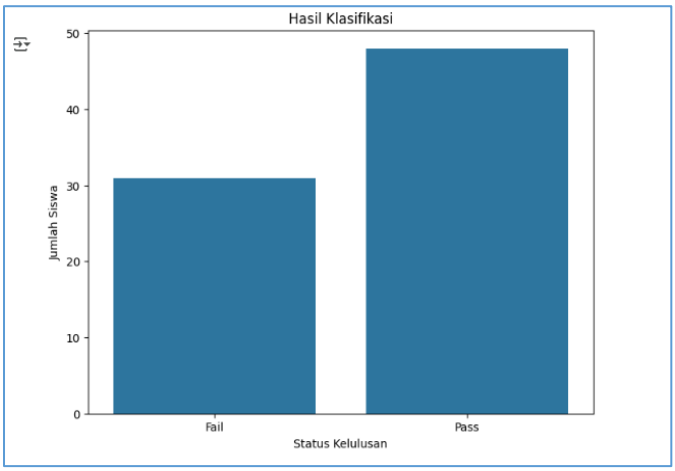
 ISSN:2302-8149
e-ISSN:2540-9719

3.3. Algoritma Random Forest

Pada tabel 2 data tersebut merupakan contoh hasil evaluasi model *Random Forest* dalam mendeteksi status kelulusan siswa berdasarkan data akademik dan non akademik. Model deteksi yang digunakan menunjukkan hasil yang sangat baik, di mana seluruh data mahasiswa yang diuji berhasil dideteksi sesuai dengan label aktualnya. Calon Mahasiswa dengan Student ID 225 dan 307 yang memiliki status aktual *Fail*, dideteksi dengan benar sebagai *Fail* beserta nilai *probabilitasnya*, sehingga model mampu mendeteksi kasus gagal dengan tepat. Sementara itu, mahasiswa dengan Student ID lainnya (305, 237, 318, 182, 84, 354,259) yang memiliki status aktual *Pass*, juga dideteksi dengan benar sebagai *Pass* dan nilai *probabilitas.*

Tabel 2. Contoh hasil klasifikasi

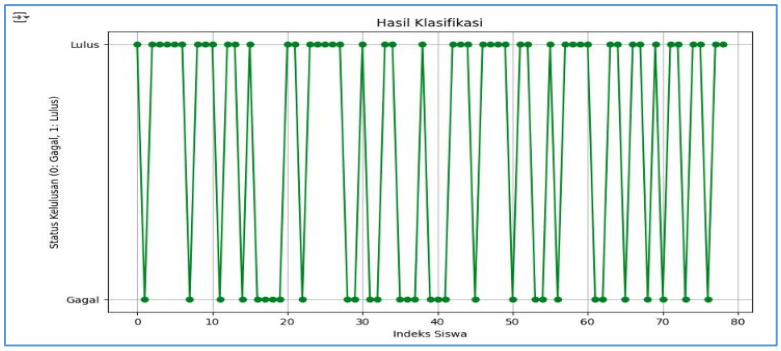
Student ID	Actual	Predicted
305	*Pass*	*Pass (95%)*
225	*Fail*	*Fail(24%)*
237	*Pass*	*Pass (90%*
318	*Pass*	*Pass(93%)*
182	*Pass*	*Pass(96%)*
84	*Pass*	*Pass (75%)*
354	*Pass*	*Pass(94%)*
259	*Fail*	*Fail(43%)*



Gambar 2. Distribusi klasifikasi

Gambar 2 di atas merupakan diagram batang yang menggambarkan distribusi status kelulusan siswa berdasarkan hasil klasifikasi. Perbandingan jumlah siswa berdasarkan hasil klasifikasi yang dilakukan. Terdapat dua kelompok, yaitu kelompok yang dideteksi *Fail* dan *Pass.* Dari grafik, dapat dilihat bahwa jumlah siswa yang diklasifikasikan sebagai *Pass* lebih banyak, yaitu mendekati 50 siswa, sedangkan jumlah siswa yang diklasifikasikan sebagai *Fail* berada di angka sekitar 30 siswa. Ini menunjukkan bahwa model deteksi memberikan hasil bahwa lebih banyak siswa diperkirakan akan lulus (*Pass*), sementara sebagian lainnya diperkirakan tidak lulus (*Fail*). Pola ini menggambarkan bahwa dalam dataset yang digunakan, proporsi siswa yang dideteksi berhasil lebih dominan, yang bisa jadi mencerminkan kondisi data aktual atau memang kecenderungan model yang lebih optimis terhadap kemungkinan kelulusan siswa.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 3. Hasil klasifikasi

Gambar 3 diatas merupakan diagram batang yang menggambarkan distribusi status kelulusan siswa berdasarkan hasil klasifikasi. Visualisasi ini menunjukkan hasil klasifikasi status kelulusan siswa berdasarkan indeks siswa. Sumbu Y menunjukkan dua kategori, yaitu Gagal (0) dan Lulus (1), sementara sumbu X menunjukkan urutan atau indeks siswa. Pola grafik menunjukkan bahwa deteksi model menghasilkan distribusi yang bervariasi di mana terdapat kombinasi antara siswa yang dideteksi Lulus dan Gagal di seluruh rentang indeks siswa. Meskipun sebagian besar siswa diklasifikasikan sebagai Lulus, terlihat adanya beberapa siswa yang dideteksi Gagal, yang tersebar di berbagai indeks. Ini menunjukkan bahwa model mampu membedakan antara siswa yang berpotensi lulus dan gagal, meskipun terdapat fluktuasi deteksi pada beberapa indeks yang menunjukkan adanya perbedaan karakteristik antar siswa yang mempengaruhi keputusan model. Pola ini mengindikasikan bahwa tidak terjadi deteksi satu arah saja, melainkan model secara dinamis memetakan kelulusan siswa sesuai data yang dimilikinya. Visualisasi ini juga membantu untuk mengidentifikasi siswa mana yang perlu diperhatikan lebih lanjut, terutama yang dideteksi Gagal, agar dapat diberikan intervensi lebih awal.

3.4. Confusion Matrix

```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.92      0.84        26
           1       0.96      0.87      0.91        53

    accuracy                           0.89        79
   macro avg       0.87      0.90      0.88        79
weighted avg       0.90      0.89      0.89        79


Confusion Matrix:
              Predicted: Fail  Predicted: Pass
Actual: Fail               24                2
Actual: Pass                7               46
```

Gambar 4. Hasil Confusion Matrix

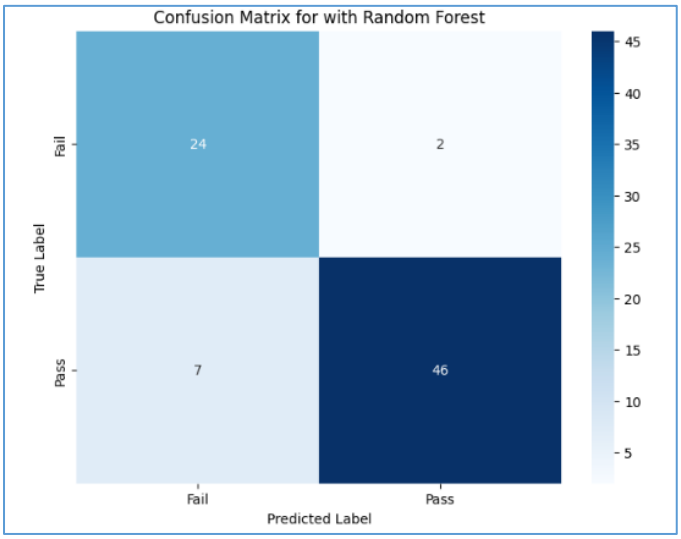
Gambar 4 menunjukkan hasil evaluasi performa model *Random Forest* dengan parameter *max_depth = 15* dan *random_state = 42*, yang menghasilkan akurasi keseluruhan sebesar 89%. Berdasarkan *confusion matrix* menunjukkan kinerja yang sangat baik dalam mengklasifikasikan data ke dalam dua kategori, yaitu lulus (*Pass)* dan tidak lulus *(Fail).* Untuk kelas 0 (*Fail), precision* sebesar 0.77 menunjukkan bahwa dari seluruh mahasiswa yang diprediksi tidak lulus, sebanyak 77% memang benar-benar tidak lulus, sementara sisanya merupakan *false positive. Recall* pada kelas ini sebesar 0.92, yang berarti 92% dari mahasiswa yang benar-benar tidak lulus berhasil terdeteksi oleh model, dengan hanya 8% yang tidak teridentifikasi (*false negative*). *F1-score* sebesar 0.84 pada kelas *Fail* menunjukkan keseimbangan yang cukup baik antara *precision* dan *recall*, yang menandakan bahwa model cukup sensitif terhadap risiko kelulusan tanpa mengorbankan terlalu banyak ketepatan.

Sementara itu, pada kelas 1 *(Pass*), model menunjukkan *precision* yang sangat tinggi yaitu 0.96, yang mengindikasikan bahwa hampir semua mahasiswa yang diprediksi lulus memang benar-benar lulus. *Recall* pada kelas ini sebesar 0.87, menunjukkan bahwa sebagian besar mahasiswa yang lulus berhasil diidentifikasi dengan benar, meskipun masih terdapat 13% yang salah diklasifikasikan sebagai tidak lulus. *F1-score* sebesar 0.91 pada kelas Pass menunjukkan bahwa performa model

 ISSN:2302-8149
e-ISSN:2540-9719

dalam mendeteksi mahasiswa yang lulus cukup stabil dan akurat. Dari 26 mahasiswa yang sebenarnya tidak lulus, sebanyak 24 berhasil diprediksi dengan benar, sedangkan 2 salah diklasifikasikan sebagai lulus. Sebaliknya, dari 53 mahasiswa yang benar-benar lulus, sebanyak 46 berhasil diklasifikasikan dengan tepat, sementara 7 salah diklasifikasikan sebagai tidak lulus.

Secara keseluruhan, model menunjukkan kinerja yang seimbang dalam menangani kedua kelas, dengan nilai *macro average* sebesar 0.88 dan *weighted average* sebesar 0.89. Nilai ini mencerminkan bahwa performa model tidak hanya kuat secara umum, tetapi juga konsisten di antara kelas dengan jumlah sampel yang berbeda. Model cenderung lebih konservatif dalam mendeteksi kelas *Fail*, dengan *recall* yang tinggi, yang sangat penting untuk menghindari kegagalan dalam mengidentifikasi mahasiswa yang membutuhkan intervensi dini. Di sisi lain, meskipun *precision* pada kelas *Pass* sangat tinggi, *recall* yang sedikit lebih rendah menunjukkan perlunya peningkatan agar tidak terjadi kesalahan dalam menyatakan mahasiswa gagal padahal sebenarnya lulus. Penelitian ini menunjukkan bahwa model *Random Forest* cukup dapat diandalkan dalam memprediksi kelulusan mahasiswa, namun tetap ada ruang untuk optimasi, khususnya dalam meningkatkan sensitivitas terhadap kelas *Pass* untuk menekan jumlah *false negative*.



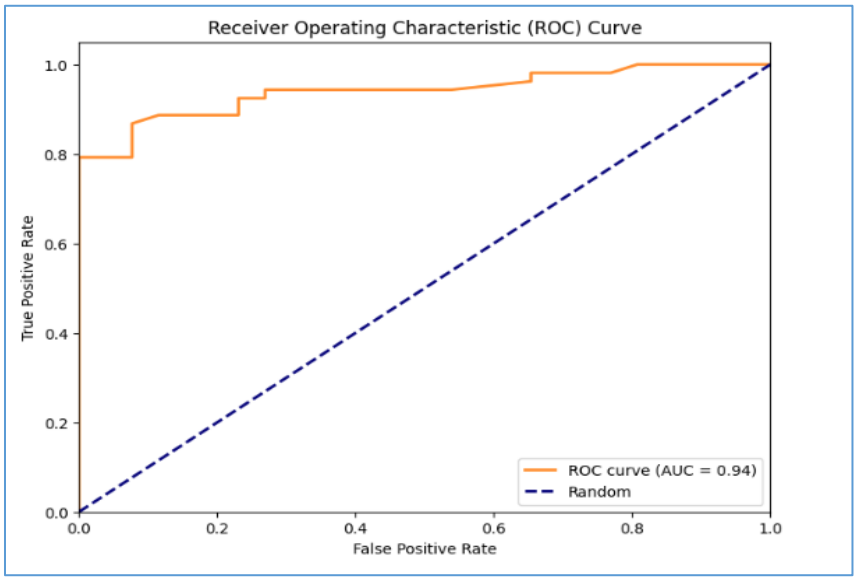
Gambar 5. Diagram Hasil Confusion Matrix

Gambar 5 menunjukkan hasil *confusion matrix* dari model *Random Forest*. Model ini menunjukkan kinerja yang baik dalam membedakan antara kelas *Fail* dan *Pass*. Dari 26 data aktual yang berlabel *Fail*, sebanyak 24 data berhasil diklasifikasikan dengan benar (*True Positive*/*TP*), sementara 2 data lainnya salah terklasifikasi sebagai *Pass* (*False Negative*/*FN*). Sementara itu, dari 53 data yang berlabel *Pass*, model berhasil mendeteksi 46 data secara tepat (*True Negative*/*TN*), namun terdapat 7 data yang salah diklasifikasikan sebagai *Fail* (*False Positive*/*FP*).

Hasil ini menunjukkan bahwa model lebih akurat dalam mengidentifikasi mahasiswa yang benar-benar berpotensi gagal (*Fail*), dengan jumlah kesalahan prediksi yang rendah ke kelas *Pass*. Sebaliknya, pada data *Pass*, meskipun sebagian besar berhasil diprediksi dengan benar, masih terdapat kesalahan klasifikasi yang menjadikan data *Pass* terdeteksi sebagai *Fail*. Pola ini mencerminkan bahwa model cenderung bersikap konservatif, yaitu lebih hati-hati dan cenderung menghindari kesalahan jenis *False Negative* pada kelas *Fail*. Strategi ini bisa menguntungkan dalam konteks tertentu, seperti pemantauan risiko akademik, di mana lebih baik memberikan perhatian lebih awal kepada mahasiswa yang berpotensi tidak lulus, meskipun ada risiko meningkatkan *False Positive*, yaitu salah menilai mahasiswa yang sebenarnya tidak bermasalah sebagai berisiko.

 ISSN:2302-8149
e-ISSN:2540-9719

3.5 ROC-AUC (Receiver Operating Characteristic -Area Under the Curve)



Gambar 6. Grafik hasil ROC-AUC

Berdasarkan gambar 6 *Receiver Operating Characteristic (ROC) Curve* yang ditampilkan, dapat disimpulkan bahwa model klasifikasi yang digunakan menunjukkan kinerja yang sangat baik. Hal ini terlihat dari nilai *Area Under the Curve (AUC)* sebesar *0.94*, yang menunjukkan bahwa model memiliki kemampuan sangat tinggi dalam membedakan antara kelas positif dan negatif. Nilai *AUC* yang mendekati *1* menandakan bahwa model mampu mengidentifikasi hampir seluruh kasus positif dengan tingkat kesalahan yang sangat rendah. Kurva *ROC* model berada jauh di atas garis diagonal (*baseline*), yang menunjukkan bahwa model bekerja jauh lebih baik dibandingkan dengan prediksi acak. Dengan kata lain, semakin ke kiri dan atas arah kurva *ROC*, semakin baik performa model, dan model ini memperlihatkan pola tersebut secara konsisten. Oleh karena itu, berdasarkan kurva *ROC* dan nilai *AUC* yang tinggi, dapat disimpulkan bahwa model berjalan dengan sangat baik dalam melakukan klasifikasi.

5 Kesimpulan

Model menunjukkan performa yang sangat baik dengan menggunakan algoritma *Random Forest* dan parameter *max_depth* = 15 serta *random_state* = 42, menghasilkan akurasi keseluruhan sebesar 89%. Deteksi terhadap kelas *Fail* sangat kuat dengan *recall* mencapai 92%, meskipun *precision*-nya berada di angka 77%, yang berarti masih ada sebagian mahasiswa yang sebenarnya *Pass* namun diklasifikasikan sebagai *Fail*. Sementara itu, untuk kelas *Pass*, *precision* model sangat tinggi yakni 96%, walaupun *recall*-nya lebih rendah di angka 87%. Pola ini tercermin pada *confusion matrix*, di mana dari 26 data *Fail*, hanya 2 yang salah diklasifikasikan sebagai *Pass*, dan dari 53 data *Pass*, terdapat 7 yang salah dideteksi sebagai *Fail*. Selain itu, nilai *Area Under the Curve (AUC)* sebesar *0.94* dari grafik *Receiver Operating Characteristic (ROC)* menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik, dengan keseimbangan tinggi antara sensitivitas dan spesifisitas. Temuan ini mengindikasikan bahwa model sangat sesuai digunakan dalam konteks akademik yang memprioritaskan deteksi dini terhadap mahasiswa berisiko tidak lulus. Namun demikian, kesalahan klasifikasi pada mahasiswa *Pass* tetap perlu diperhatikan agar tidak menimbulkan intervensi yang tidak diperlukan. Untuk pengembangan lebih lanjut, direkomendasikan agar penelitian mencakup variabel tambahan seperti aspek psikologis, motivasi belajar, dan kondisi sosial ekonomi, serta menerapkan teknik *cross-validation* dan eksplorasi parameter tambahan guna meningkatkan akurasi, generalisasi, dan kestabilan model dalam mendeteksi potensi kelulusan mahasiswa secara lebih menyeluruh.

 ISSN:2302-8149
e-ISSN:2540-9719

Referensi

[1] Supriyanto, "Strategi Membangun Budaya Akademik Mahasiswa," *Ilmu Pendidikan: Jurnal Kajian Teori dan Praktik Kependidikan*, Vol. 6, No. 1, pp. 11–21, 2021. DOI: http://dx.doi.org/10.17977/um027v6i12021p011

[2] M. H. B. Roslan and C. J. Chen, "*Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021),*" *International Journal of Emerging Technologies in Learning*, Vol. 17, No. 5, pp. 147–179, 2022, doi: 10.3991/ijet.v17i05.27685.

[3] 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," *Jurnal Sains Informatika Terapan (JSIT) E-ISSN*, Vol. 4, No. 1, pp. 24–28, 2025, Accessed: May 16, 2025. [Online]. Available: https://rcf-indonesia.org/home/

[4] "Panduan Indikator Kinerja Utama (IKU) Perguruan Tinggi Tahun 2023", Direktorat Jenderal Pendidikan Tinggi," https://pddikti.kemdikbud.go.id.

[5] S. Sobari, A. I. Purnamasari, A. Bahtiar, and K. Kaslani, "Meningkatkan Model Prediksi Kelulusan Santri Tahfidz di Pondok Pesantren Al-Kautsar menggunakan *Algoritma Random Forest*," *Jurnal Informatika dan Teknik Elektro Terapan*, Vol. 13, No. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5704.

[6] L. Breiman, "*Random Forests*," Vol. 45, Kluwer Academic, 2001, pp. 5–32.https://doi.org/10.1023/A:1010933404324

[7] S. Ray, *A Quick Review of Machine Learning Algorithms*. 2019. doi: 10.1109/COMITCon.2019.8862451.

[8] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "*Text Classification Algorithms: A Survey*," 2019, *MDPI AG*. doi: 10.3390/info10040150.

[9] S. Kumar, F. Janan, and S. K. Ghosh, "*Prediction of Students' Performance using Random Forest Classifier*," in *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore*, Singapore, Mar. 2021, pp. 7089–7100. [Online]. Available: https://www.researchgate.net/publication/354925634

[10] C. Ma, "*Improving the Prediction of Student Performance by Integrating a Random Forest Classifier with Meta-Heuristic Optimization Algorithms*," *IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 6, pp. 1032–1044, 2024, [Online]. Available: www.ijacsa.thesai.org

[11] F. Orji and J. Vassileva, "*Using Machine Learning to Explore the Relation between Student Engagement and Student Performance*," in *Proceedings of the International Conference on Information Visualisation*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 480–485. doi: 10.1109/IV51561.2020.00083.

[12] Y. Chen and K. Jin, "*Educational Performance Prediction with Random Forest and Innovative Optimizers: A Data Mining Approach*," *IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 3, 2024, [Online]. Available: www.ijacsa.thesai.org

[13] S. M. F. D. S. Mustapha, "*Predictive Analysis of Students' Learning Performance using Data Mining Techniques: A Comparative Study of Feature Selection Methods*," *Applied System Innovation*, Vol. 6, No. 5, pp. 2–24, Oct. 2023, doi: 10.3390/asi6050086.

[14] J. Kuswanto, H. Lukmanul, A. Info, and K. Kunci, "Penerapan *Algoritma Random Forest* untuk memprediksi Performa Akademik Mahasiswa," *Decode (Jurnal Pendidikan Teknologi Informasi)*, Vol. 5, No. 1, pp. 262–270, 2025, doi: http://dx.doi.org/10.51454/decode.v5i1.1103ll.

[15] Y. Priantama, T. Azhima, and Y. Siswa, "*Optimasi Correlation-based Feature Selection untuk Perbaikan Akurasi Random Forest Classifier dalam Prediksi Performa Akademik Mahasiswa*," *Jurnal Informatika dan Komputer)*, Vol. 6, No. 2, pp. 251–260, 2022.

[16] R. P. Munggaran, M. Nurmalasari, H. Hosizah, and D. Krismawati, "Prediksi Waktu Tunggu Pelayanan Pasien Rawat Jalan dengan *Algoritma Random Forest*," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, Vol. 5, No. 1, pp. 35–40, Nov. 2024, doi: 10.57152/malcom.v5i1.1529.

 ISSN:2302-8149
e-ISSN:2540-9719

[17] R. Herdiana, "Prediksi Penetapan Tarif Penerbangan menggunakan *Auto-Ml* dengan *Algoritma Random Forest*," *Jurnal Ilmu Komputer Ruru*, Vol. 2, No. 1, pp. 17–23, 2025, doi: 10.69688/jikr.v2i1.10.

[18] R.S. Reza and M.A. Yusuf, "Penerapan *Algoritma Random Forest* untuk Klasifikasi Kualitas Air berbasis Web," *Jurnal Ilmu Komputer Dan Informatika*, Vol. 1, No. 3, pp. 79–88, Jan. 2025, Accessed: May 16, 2025. [Online]. Available: https://jurnal.globalscients.com/index.php/jiki

[19] A. Fauzi, N. Maulidah, R. Supriyadi, H. Nalatissifa, and S. Diantika, "Prediksi Harga Properti di Indonesia menggunakan *Algoritma Random Forest*," *Journal of Artificial Intelligence and Digital Business (RIGGS)*, Vol. 4, No. 1, pp. 43–49, 2025, doi: 10.31004/riggs.v4i1.367.

[20] Rumini, Norhikmah, "Prediksi Kegagalan Siswa dalam Data Mining dengan menggunakan *Metode Naive Bayes*," *Jurnal Mantik Penusa*, Vol. 3, No. 1, pp. 42–46, 2019, doi: 10.13140/RG.2.2.22726.42560.