**SKANIKA: Sistem Komputer dan Teknik Informatika**
Volume 9, Nomor 1, Januari 2026, Halaman 76-86 E- ISSN:2721-4788

# Prediksi Kelulusan Mahasiswa Menggunakan Algoritma XGBoost

**Syaiful Imron[1*], Arbiati Faizah[2], Sugianto[3]**
[1,2,3]Fakultas Teknologi Informasi, Sistem dan Teknologi Informasi, Institut Teknologi dan Bisnis PGRI Dewantara, Jombang, Indonesia
E-mail: [1*]imron@itebisdewantara.ac.id, [2]arbiati.faizah@itebisdewantara.ac.id,
[3]sugianto@itebisdewantara.ac.id
(* : corresponding author)

## Abstrak

Waktu kelulusan mahasiswa seringkali sulit diprediksi sejak dini, hal ini menjadi masalah utama yang dihadapi institusi. Evaluasi yang masih bersifat manual sering kali gagal mengidentifikasi mahasiswa bermasalah, sehingga menyebabkan ketidaktepatan waktu kelulusan yang merugikan mahasiswa dan institusi. Hal ini sangat krusial karena durasi studi dan ketepatan waktu kelulusan merupakan kriteria penting dalam penilaian akreditasi dan mutu institusi. Sebagai solusi inovatif, penelitian ini mengembangkan model prediksi kelulusan menggunakan algoritma XGBoost dan *Random Forest* dengan menerapkan teknik optimasi *hyperparameter* melalui *Grid Search Cross Validation*. Hasil penelitian menunjukkan dengan parameter *default*, *Random Forest* lebih unggul dari pada XGBoost. Namun setelah melalui penyetelan *hyperparameter*, didapatkan XGBoost mendapatkan hasil akurasi yang lebih baik dibandingkan *Random Forest* dengan kenaikan akurasi yang signifikan dari 88,15% menjadi 92,66% (*precision* 91,87%, *recall* 91,67%, dan *F1-score* 91,38%). Hal ini menegaskan bahwa penyetelan *hyperparameter* yang tepat adalah kunci strategis untuk memaksimalkan efektivitas model klasifikasi. Dengan demikian, model ini dapat menjadi alat bantu bagi institusi dalam memantau dan mengintervensi potensi keterlambatan studi mahasiswa sejak dini.

**Kata kunci**: kelulusan mahasiswa, klasifikasi, random forest, XGBoost

## *Abstract*

*Student graduation times are often difficult to predict early, a major challenge facing institutions. Manual evaluations often fail to identify problematic students, leading to inaccurate graduation times that are detrimental to both students and institutions. This is crucial because study duration and timely graduation are important criteria in assessing institutional accreditation and quality. As an innovative solution, this study developed a graduation prediction model using the XGBoost and Random Forest algorithm, applying hyperparameter optimization techniques through Grid Search Cross Validation. The results showed that with default parameters, Random forest was superior to XGBoost. However, after hyperparameter tuning, XGBoost achieved better accuracy than Random Forest with a significant increase in accuracy, from 88.15% to 92.66% (precision 91.87%, recall 91.67%, and F1-score 91.38%). This confirms that appropriate hyperparameter tuning is a strategic key to maximizing the effectiveness of classification models. Thus, this model can be a tool for institutions to monitor and intervene early on in potential student delays.*

***Keywords**: classification, random forest, student graduation, XGBoost*

## 1. PENDAHULUAN

Pendidikan formal memuncak di perguruan tinggi [1], yang kini fokus pada pencapaian mutu unggul. Kualitas program studi di Indonesia dievaluasi melalui akreditasi oleh badan akreditasi nasional perguruan tinggi (BAN-PT). Akreditasi ini berfungsi sebagai mekanisme penjaminan mutu eksternal, yang merupakan komponen krusial dari sistem penjaminan mutu pendidikan tinggi. Proses ini berjalan melalui penilaian interaktif kriteria-kriteria dalam standar pendidikan tinggi [2].

Akreditasi program studi dinilai berdasarkan sembilan kriteria, termasuk mahasiswa dan luaran tri dharma, yang mencakup IPK, durasi studi, dan ketepatan waktu kelulusan. Hal ini menuntut institusi fokus pada mutu lulusan dan ketepatan waktu studi. Mahasiswa dianggap lulus tepat waktu jika durasi studinya maksimum empat tahun (144 SKS minimal) [3].

Mendeteksi potensi keterlambatan kelulusan mahasiswa sejak dini merupakan tantangan, dan kegagalan dalam deteksi ini dapat berujung pada penyelesaian studi yang terlambat. Guna

mengatasi masalah tersebut, pendekatan berbasis *data mining* dapat diterapkan sebagai instrumen teknis untuk memproyeksikan keberhasilan kelulusan secara akurat [4]. *Data mining* adalah sebuah disiplin ilmu yang mengintegrasikan berbagai bidang seperti *machine learning*, pengenalan pola, statistika, dan visualisasi. *Data mining* mampu mengolah data dalam jumlah besar dan menghasilkan daya prediksi yang jauh lebih kuat dibandingkan model statistik konvensional [5].

Pemanfaatan data mining dalam sektor pendidikan terus berkembang pesat. Penelitian sebelumnya menunjukkan bahwa fitur akademik seperti indeks prestasi kumulatif (IPK), nilai asesmen, dan tingkat kehadiran merupakan faktor yang paling berpengaruh terhadap performa akademik mahasiswa [6]. Selain itu, tinjauan literatur sistematis menegaskan bahwa penggunaan *educational data mining* dan analisis prediktif sangat efektif untuk mengatasi masalah retensi mahasiswa, terutama dalam mengidentifikasi faktor kunci keberhasilan serta memitigasi risiko putus kuliah melalui berbagai algoritma *machine learning* [7].

Salah satu teknik yang paling sering dimanfaatkan dalam *data mining* adalah klasifikasi. Klasifikasi sendiri berfungsi sebagai proses analisis data untuk menciptakan model prediktif yang dapat mengelompokkan data ke dalam kategori-kategori spesifik [8]. Untuk klasifikasi, *eXtreme Gradient Boosting* (XGBoost) adalah metode *ensemble* berbasis *gradient boosting* yang efektif, terkenal akan fitur anti-*overfitting* dan performa superiornya. XGBoost sangat diunggulkan karena kemampuannya memproses data tabular berfitur melimpah (*high dimensional*) [9].

Tinjauan literatur terkait menjadi rujukan utama dalam riset ini, penelitian [10] mendemonstrasikan bahwa dengan penyetelan *hyperparameter*, model XGBoost mampu menghasilkan tingkat akurasi 80,039% saat digunakan untuk klasifikasi kelayakan kredit. Penelitian mendeteksi *spam* email dengan menggunakan XGBoost didapatkan akurasi sebesar 95,3% [11]. Penelitian [12] melaporkan bahwa XGBoost menghasilkan akurasi sebesar 93.10% untuk klasifikasi kegagalan pembayaran kredit nasabah bank.

Penerapan algoritma XGBoost telah banyak diteliti, sebagaimana ditunjukkan oleh referensi [10, 11, 12]. Meskipun demikian, penelitian yang memanfaatkan *eXtreme Gradient Boosting* (XGBoost), untuk memprediksi kelulusan mahasiswa tepat waktu masih sedikit. Hal ini menjadi research gap yang penting untuk di isi. Urgensi penelitian ini terletak pada pentingnya deteksi dini kelulusan mahasiswa untuk menjaga mutu institusi. Jika masalah ini dibiarkan dan institusi hanya mengandalkan evaluasi manual, dampaknya akan fatal. Institusi akan mengalami kegagalan dalam mendeteksi mahasiswa yang bermasalah. Akibatnya, banyak mahasiswa yang lulus tidak tepat waktu yang mengakibatkan mutu dan akreditasi intitusi menurun, serta terjadi pemborosan waktu dan biaya bagi mahasiswa.

Untuk mengatasi resiko tersebut, Penelitian ini bertujuan untuk mengembangkan model peramalan (prediktif) yang dapat memprakirakan apakah mahasiswa akan lulus tepat waktu. Prediksi tersebut akan didasarkan berdasarkan detail akademik, demografi, dan sosial-ekonomi setiap mahasiswa. Dengan memanfaatkan data tersebut, hasil penelitian ini diharapkan dapat menjadi dasar bagi inisiatif institusi dalam menyusun intervensi berbasis data. Secara praktis, langkah ini bertujuan untuk mengidentifikasi dan menanggulangi secara dini kemungkinan mahasiswa yang berisiko terlambat lulus atau gagal dalam studinya.

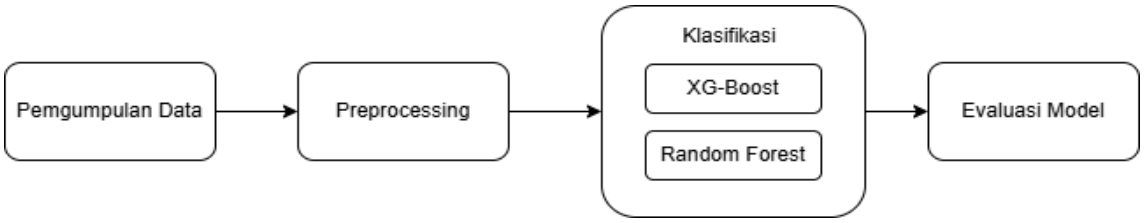
Algoritma XGBoost dipilih karena kemampuannya yang sangat akurat dalam mengolah data akademik yang kompleks dan memiliki fitur pencegah kesalahan prediksi (*overfitting*). Selain algoritma XGBoost yang menjadi fokus utama, penelitian ini juga menerapkan *Random Forest* sebagai algoritma pembanding untuk mengevaluasi efektivitas dan akurasi model yang diusulkan. *Random Forest* dikenal sangat andal dalam menangani data tabular karena kemampuannya membangun banyak pohon keputusan dan menggabungkannya untuk mendapatkan prediksi yang lebih stabil dan akurat [13] .

Kontribusi utama dari penelitian ini terletak pada optimasi model prediksi melalui peyetelan *hyperparameter* pada algoritma XGBoost dan Random Forest. Berbeda dengan penelitian sebelumnya yang sering kali menggunakan parameter standar (*default*), penelitian ini melakukan pencarian parameter optimal secara sistematis untuk meningkatkan presisi prediksi kelulusan. Dengan menyesuaikan parameter kunci seperti *learning rate*, *max depth*, dan *n_estimators* pada

XGBoost, serta jumlah pohon dan fitur pada *Random Forest*. Pendekatan ini menghasilkan model yang lebih akurat dalam memprediksi kelulusan mahasiswa dibandingkan model biasa. Dengan demikian, Penelitian ini memberikan cara baru bagi institusi untuk membuat sistem peringatan dini yang lebih akurat dalam memprediksi kelulusan mahasiswa.

## 2. METODE PENELITIAN

Tujuan utama penelitian ini adalah menerapkan algoritma XGBoost untuk memprediksi kelulusan mahasiswa tepat waktu. Metodologi yang digunakan dibagi menjadi empat fase krusial: (1) Mengumpulkan Data, (2) Melakukan *Preprocessing* dan Memilih Fitur, (3) Mengimplementasikan Metode Klasifikasi, dan (4) Mengevaluasi Kinerja Model. Secara keseluruhan, alur penelitian ini disajikan dalam Gambar 1. Metodologi Penelitian.




Gambar 1. Metodologi Penelitian

### 2.1. Pengumpulan Data

Penelitian ini akan memanfaatkan sebanyak 1140 data mahasiswa program studi Manajemen di ITEBIS PGRI Dewantara Jombang dari angkatan 2017 hingga 2021. Sumber data mahasiswa ini berasal dari Sistem Informasi Manajemen (SIM) institusi. Dataset yang digunakan terdiri dari atribut-atribut berikut sebagai variabel *input* antara lain usia, jenis kelamin, status mahasiswa, Indeks Prestasi (IP) dari Semester 1 hingga 4, IP Kumulatif (IPK), dan jumlah SKS total. Sementara itu, Status Kelulusan akan berperan sebagai variabel *target* (kelas). Rincian dataset disajikan dalam Tabel 1 berikut.

Tabel 1. Rincian Dataset

| ***Attribute*** | **Nama** | **Tipe Data** | **Kategori** |
|---|---|---|---|
| $X_1$ | Usia | Numerik | - |
| $X_2$ | Jenis Kelamin | Kategorikal | 0. Laki-laki<br>1. Perempuan |
| $X_3$ | Status Mahasiswa | Kategorikal | 0. Bekerja<br>1. Tidak Bekerja |
| $X_4$ | Indeks Prestasi Semester 1 (IPS1) | Numerik | - |
| $X_5$ | Indeks Prestasi Semester 1 (IPS1) | Numerik | - |
| $X_6$ | Indeks Prestasi Semester 2 (IPS1) | Numerik | - |
| $X_7$ | Indeks Prestasi Semester 3 (IPS1) | Numerik | - |
| $X_8$ | Indeks Prestasi Semester 4 (IPS1) | Numerik | - |
| $X_9$ | Indeks Prestasi Komulatif (IPK) | Numerik | - |
| $X_{10}$ | SKS Total | Numerik | |
| $Y$ | Status Kelulusan | Kategorikal | 0. Tepat Waktu<br>1. Terlambat |

### 2.2. *Preprocessing*

Tahap *preprocessing* melibatkan serangkaian tahapan untuk mempersiapkan data. Langkah awal adalah pada pembersihan data (*data cleaning*), pada tahap ini meliputi menghilangkan data anomali (*outlier*) dan perbaikan atau pengisian data kosong [14]. Penghapusan data anomali ini sangat penting karena kehadirannya berpotensi mengganggu atau menurunkan tingkat akurasi dari model prediktif yang akan dikembangkan. Setelah data anomali ditangani, tahap selanjutnya dilakukan pemeriksaan terhadap data kosong (*missing values*) dengan memastikan setiap entri data telah lengkap. Hal cukup krusial karena apabila terjadi data kosong (missing values) bisa memicu prediksi model yang tidak akurat [15].

Setelah kelengkapan data diverifikasi, tahap berikutnya adalah mengeliminasi entri data yang memiliki duplikasi untuk menjamin keunikan setiap baris informasi. Tahap akhir dari *preprocessing* ini berfokus pada transformasi data. Agar data dapat diolah secara efektif oleh algoritma XGBoost maupun *Random Forest*. Seluruh variabel kategorikal yang ada harus diubah formatnya menjadi data numerik (angka). Konversi ini diimplementasikan menggunakan teknik *label encoding*, yang merupakan langkah penting agar semua fitur siap digunakan dalam pelatihan model [16].

### 2.3. XGBoost

Algoritma *eXtreme Gradient Boosting* (XGBoost), yang merupakan pengembangan dari konsep *Gradient Boosting*, pertama kali diperkenalkan pada tahun 2014 oleh Dr. Tianqi Chen dari Universitas Washington [17]. XGBoost merupakan algoritma *machine learning* yang berakar pada struktur pohon keputusan (*decision tree*). Secara spesifik, algoritma ini dikembangkan untuk mengoptimalkan prosedur *boosting* sembari menjamin efisiensi komputasi dan menghasilkan tingkat akurasi prediksi yang superior [18, 19]. XGBoost merupakan wujud implementasi mutakhir dari konsep *gradient boosting* tingkat lanjut. Desainnya memungkinkan penanganan yang efektif terhadap dataset berskala besar, menangani masalah fitur yang tidak seimbang (*imbalanced features*), dan mengelola berbagai macam format data masukan dengan sangat baik.

Konstruksi model dalam XGBoost terjadi melalui mekanisme iteratif, di mana setiap pohon keputusan yang baru ditambahkan bertujuan untuk meminimalkan *residual error* atau kesalahan prediksi yang dihasilkan oleh model agregat sebelumnya. Algoritma ini mengadopsi prinsip pembelajaran *ensemble*, yang mensinergikan beberapa *weak learner* (pembelajar lemah) menjadi satu model yang kuat (*robust*) dan sangat kompeten [20]. XGBoost menunjukkan kapabilitas luar biasa dalam memodelkan hubungan *non-linier* di antara fitur-fitur dan menyediakan fleksibilitas yang diperlukan untuk beradaptasi dengan beragam skenario data. Lebih lanjut, untuk memitigasi risiko *overfitting*, XGBoost secara *inheren* menyertakan regularisasi sebagai bagian integral dari fungsi tujuannya.

Implementasi XGBoost diawali dengan inisialisasi model, yang mencakup penetapan *loss function* dan *hyperparameter* seperti jumlah pohon, kedalaman pohon maksimum, dan *parameter* regularisasi. Pertama membangun model dengan mengatur prediksi awal ($\hat{y}$) sebagai nilai target. Selanjutnya, pohon pertama dibangun untuk meminimalisir *loss function*. Proses ini berlangsung secara iteratif, dimana pohon baru ditambahkan pada setiap langkah untuk mengoreksi kesalahan prediksi dari pohon sebelumnya. Setiap penambahan pohon, juga disertai regularisasi untuk mencegah *overfitting*. Terakhir, prediksi akhir didapatkan melalui agregasi kontribusi dari seluruh pohon yang telah dikembangkan. Untuk perhitungan kolektifnya diformulasikan dalam persamaan 1.

$$\hat{y}_i = \sum_{K=1}^{K} f_K(x_i) \tag{1}$$

### 2.4. *Random Forest*

Pada tahun 2001, Breiman mengenalkan algoritma *Random Forest*, yaitu metode *ensemble learning* yang beroperasi dengan membangun sekumpulan pohon keputusan pada waktu pelatihan dan mengeluarkan kelas yang merupakan *mode* klasifikasi dari pohon-pohon individu tersebut. Keunggulan utama *Random Forest* terletak pada akurasi klasifikasinya yang tinggi serta mampu menangani *high dimentional data* dan *mixed type data* [21]. Algoritma ini dipilih karena ketahanannya terhadap *noise* dan kemampuannya untuk menangani hubungan *non-linear* antar fitur akademik [22].

Prioritas atau relevansi berbagai faktor dalam pemodelan Random Forest dapat diukur menggunakan koefisien *Gini*, dengan detail perhitungan rumus sebagai berikut:

$$G = p_{(s)} x\, p_{(m)} \tag{2}$$

Variabel G merepresentasikan indeks *Gini* yang berfungsi memetakan potensi kesalahan pengklasifikasian sampel dalam sebuah populasi. Melalui keterkaitan antara probabilitas seleksi $p_{(s)}$ dan probabilitas kesalahan $p_{(m)}$, kualitas himpunan dapat diukur. Nilai G yang rendah menandakan tingginya homogenitas atau kemurnian kelas, sedangkan nilai G yang tinggi menunjukkan bahwa himpunan tersebut cenderung tidak murni karena besarnya peluang terjadi salah klasifikasi [23].

### 2.5. Evaluasi

Evaluasi model merupakan tahapan yang sangat penting (krusial) dalam siklus *machine learning*, yang berfungsi untuk mengukur kapabilitas model saat diterapkan pada data yang benar-benar baru, khususnya data uji [24]. Dalam penelitian ini, performa model dianalisis menggunakan empat metrik kinerja inti: *accuracy*, *recall*, *precision*, dan *F1-score*.

### 2.6. *Hyperparameter*

Penyetelan *hyperparameter* merupakan langkah krusial dalam meningkatkan performa model. Salah satu pendekatan efektif dalam penelitian adalah dengan memanfaatkan teknik *random search cross-validation*, yang mengeksplorasi beragam kombinasi *hyperparameter* secara acak. Dalam penelitian ini menggunakan *RandomizedSearchCV* [9]. *Hyperparameter* yang diterapkan dalam penelitian ini telah dirangkum dan dirinci dalam Tabel 2 dan Tabel 3.

Tabel 2. *Hyperparameter* XGBoost

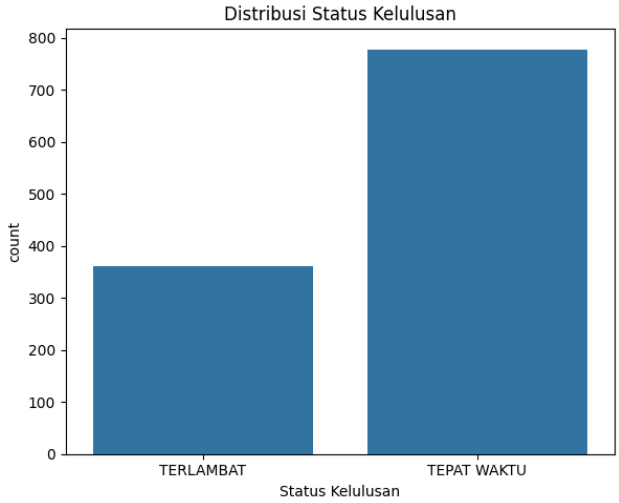
| Nama | Fungsi |
|---|---|
| *n_estimators* | Mengidentifikasi total pohon yang perlu dibuat |
| *learning_rate* | Menetapkan besaran bobot model oleh algoritma optimasi guna mengurangi error |
| *max_depth* | Tingkat kedalaman struktur pohon |
| *min_child_weight* | Menetapkan ambang batas minimum dari total bobot pada sebuah cabang pohon. |
| *subsample* | *Subset* data training yang dipilih untuk menyesuaikan *fit* dari setiap pohon keputusan dalam model. |
| *colsample_bylevel* | Rasio fitur yang diambil sampelnya (*subsampling*) untuk dipertimbangkan sebagai kandidat *split* pada setiap level pembentukan pohon. |
| *gamma* | Parameter yang mengatur intensitas pemangkasan yang diterapkan pada setiap tahap pemisahan (*splitting*) |

Tabel 3. *Hyperparameter Random Forest*

| Nama | Fungsi |
|---|---|
| *n_estimators* | Mengidentifikasi total pohon yang perlu dibuat |
| *max_features* | Batas tertinggi jumlah fitur acak yang dievaluasi oleh tiap simpul dalam proses penentuan percabangan |
| *max_depth* | Tingkat kedalaman struktur pohon |
| *min_samples_split* | Jumlah data minimal dalam tiap simpul untuk memicu terjadinya tahap *splitting* berikutnya |
| *min_samples_leaf* | Jumlah sampel minimal yang harus ada dalam setiap pohon setelah *splitting* |
| *criterion* | Fungsi statistik yang digunakan untuk mengukur kualitas pembelahan (*split*) |

## 3. HASIL DAN PEMBAHASAN

Dataset untuk memprediksi lulus kuliah tepat waktu bagi mahasiswa terdiri dari 1140 data dengan 11 kolom yang merangkum berbagai detail akademik, demografi, dan sosial-ekonomi setiap mahasiswa. Dataset ini bertujuan untuk menyelidiki faktor-faktor yang memengaruhi kelulusan mahasiswa. Label target dataset ini diklasifikasikan ke dalam dua kategori utama yaitu Tepat Waktu dan Terlambat. Tepat Waktu menandakan individu yang telah lulus kuliah dalam waktu maksimal 8 semester atau 4 Tahun. Terlambat merujuk pada individu yang kelulusannya melebihi 8 semester atau 4 Tahun. Gambar 1 mengilustrasikan distribusi label-label ini, yang menunjukkan distribusi yang tidak seimbang antara kedua kategori dalam dataset.




Gambar 2. Perbandingan *Target*

Langkah awal pada fase pra-pemrosesan data (*preprocessing*) adalah menghapus variabel-variabel yang dianggap tidak relevan atau tidak diperlukan untuk analisis. Setelah itu, dilanjutkan dengan prosedur deteksi dan penanganan nilai hilang (*missing values*) serta data *outliers*. Sebelum memulai proses pemodelan (*modeling*), dataset harus terlebih dahulu dibagi dan dipisahkan menjadi dua *subset* yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Dalam penelitian ini, pembagian data dilakukan dengan proporsi 80% dari total jumlah dataset dialokasikan sebagai data latih. Sedangkan sisanya, 20% dari total dataset digunakan untuk data uji.

Tahapan pengujian ada 4 macam metode pengujian yang digunakan yaitu, XGBoost *default* parameter, XGBoost *hyperparamter*, *Random Forest default* parameter dan *Random Forest hyperparameter*. Optimasi ini bertujuan untuk meningkatkan kapabilitas dan performa model klasifikasi secara signifikan. Nilai *hyperparameter* terbaik dari XGBoost disajikan pada Tabel 4. Sedangkan nilai *hyperparameter* terbaik dari *Random Forest* disajikan pada Tabel 5.

Tabel 4. Nilai *Hyperparameter* XGBoost

| ***Hyperparameter*** | **Parameter** | **Parameter Terbaik** |
|---|---|---|
| *n_estimators* | 50, 100, 200, 300 | 300 |
| *learning_rate* | 0.01, 0.02, 0.1, 0.2 | 0.02 |
| *max_depth* | 3, 4, 5, 6, 7 | 3 |
| *min_child_weight* | 1, 2, 3, 4, 5 | 1 |
| *subsample* | 0.7, 0.8, 0.9 | 0.7 |
| *colsample_bylevel* | 0.1, 0.2, 0.25, 1 | 0.1 |
| *gamma* | 0, 0.1, 0.2, 1, 1.5, 2 | 0.1 |

Tabel 5. Nilai *Hyperparameter Random Forest*

| ***Hyperparameter*** | **Parameter** | **Parameter Terbaik** |
|---|---|---|
| *n_estimators* | 100, 200, 500 | 100 |
| *max_features* | *auto, sqrt, log2* | *sqrt* |
| *max_depth* | 4, 5, 6, 7, 8 | 5 |
| *min_samples_split* | 2, 5, 10 | 5 |
| *min_samples_leaf* | 1, 2, 4 | 1 |
| *criterion* | *gini, entropy* | *gini* |

Tabel 6 merupakan perbandingan antara pengujian parameter *default* dan pengujian dengan penyetelan *hyperparameter* dari algoritma XGBoost dan *Random Forest*.

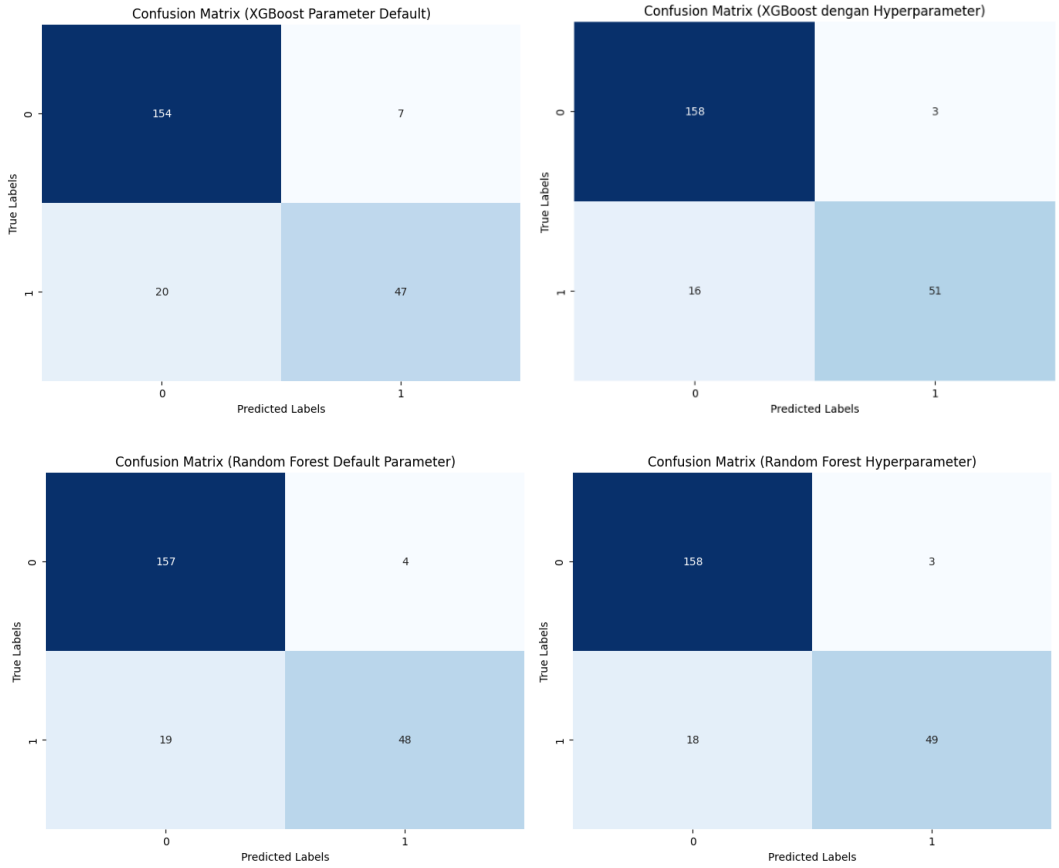
Tabel 6. Perbandingan Hasil ujicoba

| **Algoritma** | **Pengujian** | ***Precision*** | ***Recall*** | ***F1-Score*** | ***Accuracy*** |
|---|---|---|---|---|---|
| XGBoost | Parameter *Default* | 88,07% | 88,16% | 87,75% | 88,15% |

| Algoritma | Pengujian | *Precision* | *Recall* | *F1-Score* | *Accuracy* |
|---|---|---|---|---|---|
| Random Forest | Menggunakan Hyperparameter | 91,87% | 91,67% | 91,38% | 91,66% |
| | Parameter *Default* | 90,12% | 89,51% | 89,90% | 89,91% |
| | Menggunakan Hyperparameter | 91,08% | 90,79% | 90,41% | 90,78% |

Pada tahap ujicoba menggunakan parameter bawaan (*default*), algoritma *Random Forest* menghasilkan akurasi sebesar 89,91%, mengungguli XGBoost yang mencatatkan akurasi 88,15%. Namun, setelah melalui proses optimasi menggunakan penyetelan *hyperparameter*, terjadi peningkatan pada kedua model. Model XGBoost meningkat secara signifikan menjadi 91,66%, sementara *Random Forest* mencapai akurasi 90,78%. Perbandingan ini menunjukkan bahwa meskipun *Random Forest* lebih unggul pada kondisi parameter *default*, namun XGBoost memiliki ruang peningkatan (*gain*) yang lebih besar melalui penyetelan *hyperparameter* yang tepat.

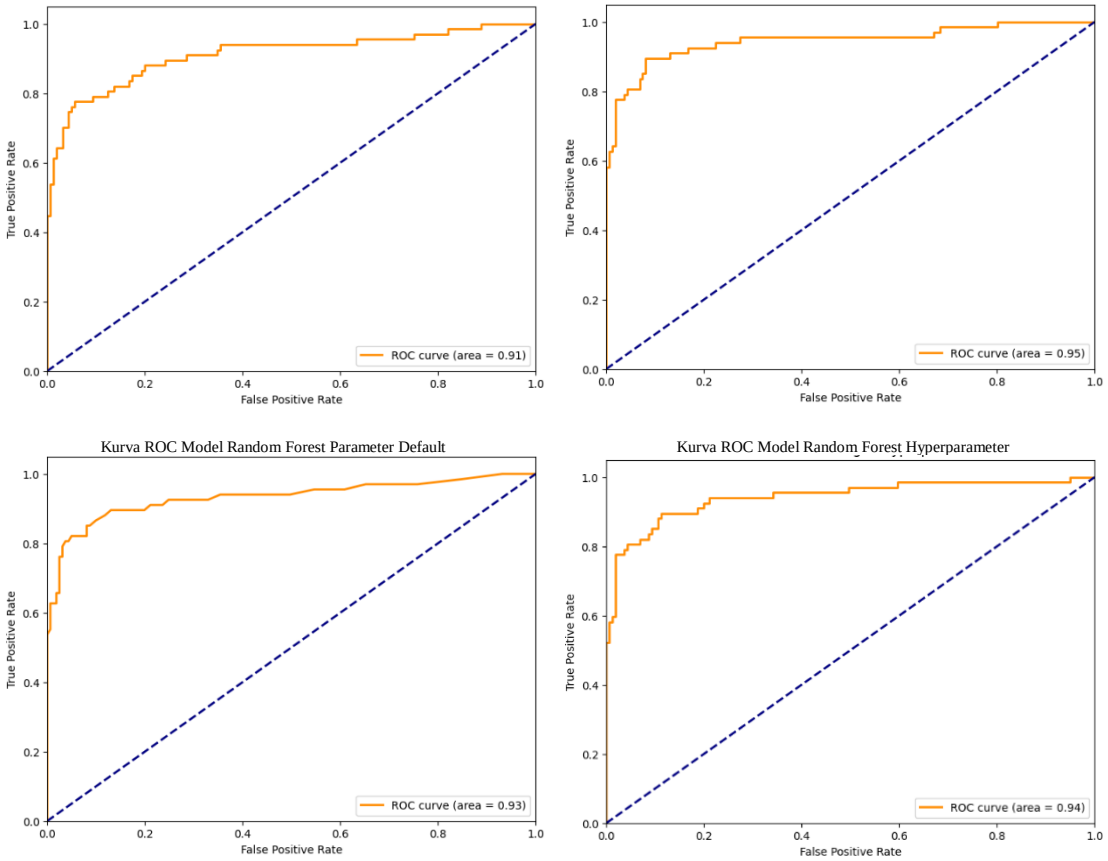
Gambar 3 menyajikan visualisasi dari *confusion matrix* yang dihasilkan oleh model klasifikasi XGBoost dan *Random Forest* parameter *default*, serta klasifikasi menggunakan penyetelan *hyperparameter*. Matriks ini menunjukkan bahwa klasifikasi XGBoost menggunakan penyetelan *hyperparameter* menunjukkan hasil prediksi yang lebih baik dibandingkan model Random Forest. Sebagai contoh untuk kelas 0 sangat akurat, hanya mencatat tiga kesalahan klasifikasi.




Gambar 3. Perbandingan *confusion matrix*

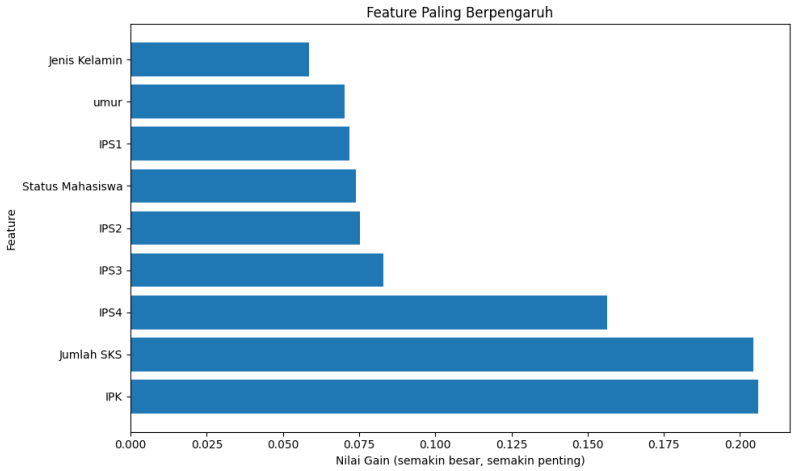
Perbandingan visual dari performa diskriminatif model diilustrasikan secara jelas dalam Gambar 4, yang menyajikan nilai *Area Under the Curve* (AUC) untuk setiap skenario pengujian. Di antara semua eksperimen, model yang menggunakan XGBoost penyetelan *hyperparameter* membuktikan keunggulannya dengan mencapai nilai AUC yang lebih tinggi, yaitu 0,95.

Kurva ROC Model XGBoost Parameter Default Kurva ROC Model XGBoost Hyperparameter




Gambar 4. Perbandingan Kurva ROC

Prosedur analisis *feature importance* dijalankan untuk mengidentifikasi variabel-variabel yang memiliki kontribusi paling signifikan terhadap model. Tahapan evaluasi ini didasarkan pada model yang menunjukkan kinerja terbaik, yaitu XGBoost yang dilakukan penyetelan *hyperparameter*, guna menilai seberapa besar peran setiap atribut dalam menghasilkan prediksi oleh algoritma tersebut. Visualisasi yang terdapat pada Gambar 5 menguraikan atribut-atribut yang memiliki pengaruh terbesar dalam memprediksi model. Model mengidentifikasi dua fitur utama yang paling dominan, yaitu IPK dan jumlah SKS.




Gambar 5. Fitur Paling Berpengaruh

## 4. KESIMPULAN DAN SARAN

Hasil penelitian yang telah dipaparkan menunjukkan kesimpulan penting mengenai kinerja klasifikasi. Pengujian awal, yang menerapkan algoritma XGBoost dengan parameter *default* pada dataset mahasiswa, menghasilkan model yang dinilai cukup baik. Model ini mencatatkan akurasi sebesar 88,15%. Pada pengujian kedua, algoritma XGBoost dilakukan teknik optimasi melalui proses penyetelan *hyperparameter* yang menggunakan tujuh *hyperparameter*. Setelah optimasi, kinerja model menujukkan peningkatan. Dimana akurasi menjadi 91,66%. Selain itu untuk pengujian algoritma *Random Forest* dengan parameter *default* didapatkan akurasi sebesar 89,91%. Sedangkan algoritma *Random Forest* setelah dilakukan proses penyetelan *hyperparameter* didapatkan akurasi sebesar 90,78%. Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa optimasi penyetelan *hyperparameter* adalah solusi utama yang direkomendasikan untuk meningkatkan efisiensi kinerja algoritma XGBoost dan *Random Forest* dalam tugas klasifikasi.

Meskipun algoritma XGBoost dengan penyetelan *hyperparameter* telah mencapai akurasi tinggi sebesar 91,66%, terdapat beberapa aspek yang perlu diperhatikan antara lain ketidakseimbangan kelas dan potensi bias data. Visualisasi distribusi *target* menunjukkan adanya ketidakseimbangan data antara jumlah mahasiswa yang lulus "Tepat Waktu" dan "Terlambat". Ketimpangan ini berisiko membuat model cenderung lebih akurat dalam memprediksi kelas mayoritas namun kurang sensitif terhadap mahasiswa yang benar-benar berisiko terlambat. Disamping itu, model sangat bergantung pada fitur akademik seperti IPK dan Jumlah SKS. Hal ini beresiko mengabaikan faktor eksternal non-akademik yang signifikan, seperti status mahasiswa, umur, atau jenis kelamin.

Untuk pengembangan selanjutnya, disarankan untuk melakukan eksperimen lebih mendalam pada teknik *stacking* atau penggabungan algoritma XGBoost, *Random Forest*, dan algoritma boosting lainnya seperti LightGBM atau AdaBoost untuk mengeksplorasi apakah kombinasi model-model tersebut dapat menghasilkan akurasi yang lebih konsisten di berbagai angkatan mahasiswa. Selain itu, untuk mengatasi ketidakseimbangan kelas dapat menggunakan teknik *oversampling* antara lain SMOTE dan ADASYN. Serta untuk mengatasi potensi bias data, jumlan kelas non-akademik dapat ditambahkan.

## DAFTAR PUSTAKA

[1] N. Hasanah, F. Syahfitri and T. Pujahadi, “Sosialisasi Tentang Pentingnya Pendidikan Tingkat Perguruan Tinggi Kepada Masyarakat Desa Jaring Halus,” *Jurnal Pengabdian Kepada Masyarakat*, vol. 2, no. 1, pp. 23-29, 2021.

[2] H. Mudarti and Y. Fatrisna, “Sistem Penjaminan Mutu Eksternal dan Akreditasi Dalam Lembaga Pendidikan Indonesia,” vol. 10, 2025.

[3] E. Haryatmi and S. P. Hervianti, “Penerapan Algoritma Support Vector Machine untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi*, vol. 5, no. 2, pp. 386-392, 2021, doi: 10.29207/resti.v5i2.3007.

[4] E. Novianto, A. Hermawan, and D. Avianto, “Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1,” *rabit*, vol. 8, no. 2, pp. 146–154, 2023, doi: 10.36341/rabit.v8i2.3434.

[5] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, vol. 110, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.

[6] S. A. Alwarthan, N. Aslam, and I. U. Khan, “Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review,” *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–26, 2022, doi: 10.1155/2022/8924028.

[7] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, “Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review,” *IEEE Access*, vol. 10, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[8] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K.

Phasinam, “Classification and prediction of student performance data using various machine learning algorithms,” *Materials Today: Proceedings*, vol. 80, pp. 3782–3785, 2023, doi: 10.1016/j.matpr.2021.07.382.

[9] M. Kumar, N. Singh, J. Wadhwa, P. Singh, G. Kumar, and A. Qtaishat, "Utilizing Random Forest and XGBoost Data Mining Algorithms for Anticipating Students’ Academic Performance," *Int. J. Mod. Educ. Comput. Sci.*, vol. 16, no. 2, pp. 29–44, 2024, doi: 10.5815/ijmecs.2024.02.03.

[10] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, “Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit,” *Jomta*, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792

[11] L. G. A. Putri, S. A. Wicaksono, dan B. Rahayudi, "Analisis Klasifikasi Spam Email Menggunakan Metode Extreme Gradient Boosting (XGBoost)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 2, pp. 1-8, 2025.

[12] F. A. P. Prasetya and P. H. P. Rosa, “Klasifikasi Kegagalan Pembayaran Kredit Nasabah Bank dengan Algoritma XGBoost,” Seminar Nasional Informatika Bela Negara (Santika) 2024, vol. 4, 2024, pp. 366-371.

[13] A. Agustiningsih, Y. Findawati, and I. Alnarus Kautsar, “Classification Of Vocational High School Graduates’ Ability In Industry Using Extreme Gradient Boosting (XGBoost), Random Forest, And Logistic Regression,” *J. Tek. Inform. (JUTIF)*, vol. 4, no. 4, pp. 977–985, 2023, doi: 10.52436/1.jutif.2023.4.4.945.

[14] Ali, Z.H.; Burhan, A.M. Hybrid Machine Learning Approach for Construction Cost Estimation: An Evaluation of Extreme Gradient Boosting Model. *Asian J. Civ. Eng*. 2023, 24, 2427–2442.

[15] A. A. Khalil, Z. Liu, A. Fathalla, A. Ali, and A. Salah, “Machine Learning Based Method for Insurance Fraud Detection on Class Imbalance Datasets With Missing Values*,” IEEE Access*, vol. 12, pp. 155451–155468, 2024, doi: 10.1109/ACCESS.2024.3468993.

[16] Q. A. Hidayaturrohman and E. Hanada, “Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure,” *BioMedInformatics*, vol. 4, no. 4, pp. 2201–2212, 2024, doi: 10.3390/biomedinformatics4040118.

[17] R. Wang, J. Zhang, B. Shan, M. He, and J. Xu, “XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage,” *NDT*, vol. Volume 18, pp. 659–667, 2022, doi: 10.2147/NDT.S349956.

[18] T. Kavzoglu and A. Teke, “Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost),” *Arab. J. Sci. Eng.*, vol. 47, no. 6, pp. 7367–7385, 2022, doi: 10.1007/s13369-022-06560-8.

[19] A. Asselman, M. Khaldi, and S. Aammou, “Enhancing the prediction of student performance based on the machine learning XGBoost algorithm,” Interact. Learn. Environ., vol. 31, no. 6, pp. 3360–3379, Aug. 2023, doi: 10.1080/10494820.2021.1928235.

[20] Q. T. Phan, Y. K. Wu, and Q. D. Phan, “A Hybrid Wind Power Forecasting Model with XGBoost, Data Preprocessing Considering Different NWPs,” Applied Sciences, vol. 11, no. 3, p. 1100, 2021, doi: 10.3390/app11031100.

[21] J. Quist, L. Taylor, J. Staaf, and A. Grigoriadis, “Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification,” *Cancers*, vol. 13, no. 5, p. 991, 2021, doi: 10.3390/cancers13050991.

[22] E. Fitri, “Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah,” *J. Appl. Comput. Sci. Technol*., vol. 4, no. 1, pp. 58–64, July 2023, doi: 10.52158/jacost.v4i1.491.

[23] S. Wang, J. Zhuang, J. Zheng, H. Fan, J. Kong, and J. Zhan, “Application of Bayesian Hyperparameter Optimized Random Forest and XGBoost Model for Landslide Susceptibility Mapping,” *Front. Earth Sci.*, vol. 9, p. 712240, 2021, doi: 10.3389/feart.2021.712240.

[24] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, 2023, doi: 10.3390/computers12050091.