Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

Perbandingan Kinerja *Random Forest*, *Decision Tree*, dan *Naive Bayes* Menggunakan Metode 10-*Fold* Cross Validation untuk Prediksi Status Akademik Mahasiswa

Jeprinus Purba*, Rahul Sinurat

Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan alam, Universitas HKBP Nommensen, Kota Pematangsiantar, Indonesia
Jl. Sangnawaluh No.4, Siopat Suhu, Kec. Siantar Timur, Kota Pematangsiantar, Sumatra Utara, Indonesia
Email: jeprinuspurba01@gmail.com
Email Penulis Korespondensi: jeprinuspurba01@gmail.com

Abstrak−Prediksi status akademik mahasiswa merupakan salah satu upaya penting yang dapat dilakukan perguruan tinggi untuk mengidentifikasi mahasiswa yang berpotensi mengalami penurunan performa akademik, keterlambatan penyelesaian studi, maupun dropout. Identifikasi secara dini memungkinkan institusi memberikan intervensi akademik dan non-akademik secara lebih tepat sasaran sehingga dapat meningkatkan tingkat keberhasilan mahasiswa dalam menyelesaikan pendidikan. Penelitian ini bertujuan membandingkan kinerja algoritma Random Forest, Decision Tree, dan Naive Bayes dalam memprediksi status akademik mahasiswa menggunakan dataset Predict Students Dropout and Academic Success yang tersedia pada UCI Machine Learning Repository. Dataset terdiri atas 4.424 data mahasiswa dengan 36 atribut prediktor dan tiga kelas target, yaitu Dropout, Enrolled, dan Graduate. Tahapan penelitian meliputi eksplorasi data, seleksi atribut, pembangunan model klasifikasi, serta evaluasi menggunakan metode 10-Fold Cross Validation. Kinerja model dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), Confusion Matrix, dan ROC Curve. Hasil pengujian menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dengan nilai Accuracy sebesar 77,4%, Precision 0,760, Recall 0,774, F1-Score 0,759, dan AUC 0,897, sehingga mampu mengungguli Decision Tree dan Naive Bayes. Kontribusi penelitian ini adalah memberikan analisis komparatif terhadap tiga algoritma klasifikasi menggunakan dataset pendidikan tinggi yang bersifat publik serta menunjukkan bahwa Random Forest merupakan algoritma yang paling efektif untuk mendukung sistem prediksi status akademik mahasiswa secara lebih akurat dan konsisten.

Kata Kunci: Machine Learning; Random Forest; Decision Tree; Naive Bayes; Prediksi Status Akademik Mahasiswa

Abstract−*Predicting students' academic status is an important effort that higher education institutions can undertake to identify students who are at risk of experiencing academic decline, delayed graduation, or dropout. Early identification enables institutions to provide appropriate academic and non-academic interventions, thereby improving student retention and graduation rates. This study aims to compare the performance of the Random Forest, Decision Tree, and Naive Bayes algorithms in predicting students' academic status using the Predict Students Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. The dataset consists of 4,424 student records with 36 predictor attributes and three target classes: Dropout, Enrolled, and Graduate. The research methodology includes data exploration, feature selection, model development using Orange Data Mining, and model evaluation through the 10-Fold Cross Validation method. Model performance was assessed using Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), Matthews Correlation Coefficient (MCC), Confusion Matrix, and ROC Curve. The experimental results indicate that the Random Forest algorithm achieved the best performance, with an Accuracy of 77.4%, Precision of 0.760, Recall of 0.774, F1-Score of 0.759, AUC of 0.897, and MCC of 0.624, outperforming both Decision Tree and Naive Bayes. This study contributes a comparative evaluation of three widely used classification algorithms on a publicly available higher education dataset and demonstrates that Random Forest is the most effective algorithm for supporting accurate and reliable student academic status prediction systems.*

Keywords*: Machine Learning; Random Forest; Decision Tree; Naive Bayes; Student Academic Status Prediction*

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong transformasi di berbagai sektor, termasuk pada bidang pendidikan tinggi. Perguruan tinggi saat ini menghasilkan data akademik dalam jumlah yang sangat besar, mulai dari data identitas mahasiswa, riwayat perkuliahan, nilai akademik, kehadiran, hingga aktivitas pembelajaran pada sistem informasi akademik. Data tersebut memiliki potensi yang besar untuk dimanfaatkan sebagai sumber informasi dalam mendukung proses pengambilan keputusan berbasis data (data-driven decision making). Salah satu penerapan yang berkembang pesat adalah pemanfaatan teknik Machine Learning untuk melakukan prediksi terhadap status akademik mahasiswa. Melalui pendekatan tersebut, institusi pendidikan dapat mengidentifikasi pola-pola yang tersembunyi dalam data sehingga mampu memprediksi kondisi akademik mahasiswa secara lebih cepat dan objektif dibandingkan metode evaluasi konvensional [1], [2].

Status akademik mahasiswa merupakan salah satu indikator penting dalam mengevaluasi keberhasilan proses pembelajaran di perguruan tinggi. Status tersebut umumnya diklasifikasikan menjadi mahasiswa yang masih aktif (Enrolled), mahasiswa yang berhasil menyelesaikan studi (Graduate), dan mahasiswa yang tidak mampu menyelesaikan studi atau mengalami Dropout. Tingginya jumlah mahasiswa yang mengalami dropout menjadi salah satu tantangan yang dihadapi oleh berbagai institusi pendidikan karena berdampak pada kualitas lulusan, efektivitas penyelenggaraan pendidikan, serta capaian indikator kinerja perguruan tinggi. Oleh karena itu, diperlukan suatu mekanisme yang mampu mendeteksi mahasiswa yang berpotensi mengalami penurunan status akademik sejak tahap awal sehingga pihak perguruan tinggi dapat memberikan pendampingan akademik maupun nonakademik secara lebih tepat sasaran [3], [4]. Selama ini, proses identifikasi mahasiswa yang berisiko mengalami penurunan prestasi akademik masih banyak dilakukan

Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

melalui evaluasi manual berdasarkan nilai semester, kehadiran, atau hasil konsultasi dengan dosen pembimbing akademik. Pendekatan tersebut memiliki beberapa keterbatasan, di antaranya membutuhkan waktu yang relatif lama, bergantung pada penilaian subjektif, serta kurang mampu memanfaatkan data akademik dalam jumlah besar yang tersimpan pada sistem informasi perguruan tinggi. Seiring berkembangnya konsep Educational Data Mining dan Learning Analytics, pemanfaatan algoritma klasifikasi menjadi salah satu solusi yang banyak digunakan untuk membantu proses prediksi status akademik mahasiswa secara otomatis dengan tingkat akurasi yang tinggi [1], [5].

Berbagai algoritma klasifikasi telah dikembangkan dan diterapkan untuk menyelesaikan permasalahan prediksi status akademik mahasiswa. Di antara algoritma yang paling banyak digunakan adalah Random Forest, Decision Tree, dan Naive Bayes. Random Forest merupakan algoritma berbasis ensemble learning yang membangun sejumlah pohon keputusan (decision tree) secara acak, kemudian menggabungkan hasil prediksi masing-masing pohon melalui mekanisme majority voting. Pendekatan tersebut mampu meningkatkan kemampuan generalisasi model serta mengurangi risiko overfitting. Sebaliknya, Decision Tree membangun satu struktur pohon keputusan yang mudah dipahami dan diinterpretasikan sehingga banyak digunakan dalam berbagai penelitian klasifikasi. Sementara itu, Naive Bayes merupakan algoritma probabilistik yang menerapkan Teorema Bayes dengan asumsi bahwa setiap atribut bersifat independen terhadap atribut lainnya. Algoritma ini memiliki proses komputasi yang sederhana dan cepat sehingga tetap menjadi salah satu metode klasifikasi yang banyak digunakan pada berbagai penelitian machine learning [4], [6], [7].

Sejumlah penelitian sebelumnya telah membuktikan bahwa algoritma machine learning mampu memberikan performa yang baik dalam memprediksi keberhasilan akademik maupun risiko dropout mahasiswa. Villar dan de Andrade [1] melakukan penelitian mengenai perbandingan beberapa algoritma machine learning untuk memprediksi keberhasilan akademik mahasiswa dan menunjukkan bahwa algoritma berbasis ensemble menghasilkan performa yang lebih tinggi dibandingkan algoritma klasifikasi tunggal. Penelitian tersebut menegaskan bahwa kombinasi beberapa model klasifikasi mampu meningkatkan kemampuan generalisasi model dalam mengenali pola data pendidikan yang kompleks.

Penelitian lain yang dilakukan oleh Putra et al. [5] membandingkan algoritma Random Forest dan XGBoost untuk memprediksi mahasiswa yang berpotensi mengalami dropout. Hasil penelitian menunjukkan bahwa Random Forest memberikan tingkat akurasi yang tinggi serta mampu mengidentifikasi atribut-atribut akademik yang paling berpengaruh terhadap keberhasilan studi mahasiswa. Selain itu, Rebelo Marcolino et al. [6] memanfaatkan data aktivitas pembelajaran pada platform Moodle untuk membangun model prediksi student dropout. Penelitian tersebut menunjukkan bahwa algoritma berbasis ensemble memiliki kemampuan yang lebih baik dalam menangkap hubungan antar atribut dibandingkan algoritma klasifikasi konvensional.

Selanjutnya, Utami et al. [4] menerapkan kombinasi metode Recursive Feature Elimination Cross Validation (RFE-CV) dengan beberapa algoritma klasifikasi untuk meningkatkan performa prediksi dropout mahasiswa. Hasil penelitian menunjukkan bahwa proses seleksi fitur mampu meningkatkan akurasi model karena hanya menggunakan atribut yang memiliki kontribusi terbesar terhadap proses klasifikasi. Sementara itu, Jain et al. [6] mengembangkan model prediksi student dropout menggunakan pendekatan Particle Swarm Optimization (PSO), Synthetic Minority Over-sampling Technique (SMOTE), dan metode ensemble. Penelitian tersebut membuktikan bahwa kombinasi teknik optimasi dan algoritma ensemble mampu meningkatkan performa model pada dataset pendidikan yang memiliki karakteristik kompleks.

Meskipun berbagai penelitian terdahulu telah berhasil menerapkan algoritma machine learning dalam memprediksi status akademik mahasiswa, masih terdapat beberapa perbedaan yang menjadi celah penelitian (research gap). Sebagian besar penelitian hanya berfokus pada penggunaan satu atau dua algoritma klasifikasi sehingga belum memberikan gambaran yang komprehensif mengenai perbandingan performa beberapa algoritma pada dataset yang sama. Selain itu, beberapa penelitian menggunakan dataset yang berbeda maupun metode evaluasi yang tidak seragam, sehingga hasil yang diperoleh sulit dibandingkan secara objektif. Penggunaan teknik evaluasi yang berbeda juga dapat menghasilkan nilai performa yang bervariasi sehingga belum memberikan informasi yang cukup mengenai algoritma yang paling sesuai untuk diterapkan pada prediksi status akademik mahasiswa.

Berdasarkan kondisi tersebut, penelitian ini menawarkan perbandingan tiga algoritma klasifikasi, yaitu Random Forest, Decision Tree, dan Naive Bayes, menggunakan dataset Predict Students Dropout and Academic Success yang diperoleh dari UCI Machine Learning Repository. Seluruh algoritma dievaluasi menggunakan metode 10-Fold Cross Validation sehingga setiap model memperoleh kondisi pengujian yang sama. Pendekatan ini diharapkan mampu menghasilkan proses evaluasi yang lebih objektif serta memberikan gambaran yang lebih akurat mengenai performa masing-masing algoritma berdasarkan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), Matthews Correlation Coefficient (MCC), Confusion Matrix, dan Receiver Operating Characteristic (ROC) Curve.

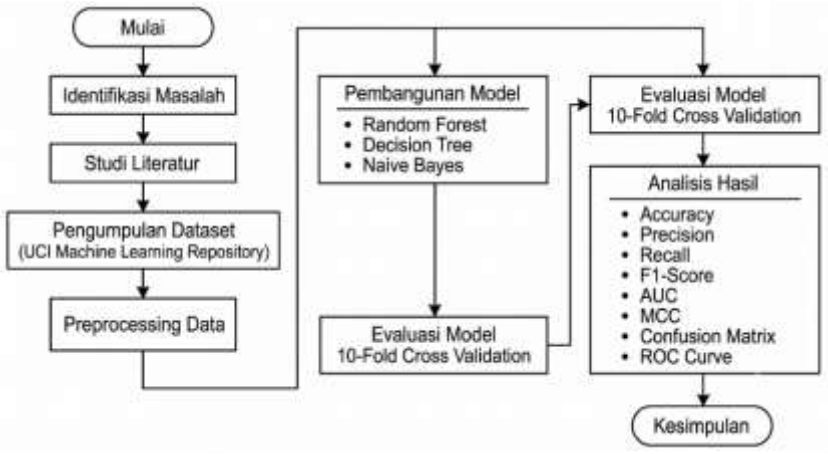
Berdasarkan uraian tersebut, penelitian ini bertujuan untuk membandingkan kinerja algoritma Random Forest, Decision Tree, dan Naive Bayes menggunakan metode 10-Fold Cross Validation dalam memprediksi status akademik mahasiswa. Hasil penelitian diharapkan dapat memberikan rekomendasi mengenai algoritma yang memiliki performa terbaik untuk diterapkan pada sistem prediksi status akademik mahasiswa, serta menjadi referensi bagi pengembangan sistem early warning di perguruan tinggi dalam mengidentifikasi mahasiswa yang berisiko mengalami dropout secara lebih dini.

2. METODOLOGI PENELITIAN

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

2.1 Tahapan Penelitian

Penelitian ini menggunakan metode eksperimen dengan pendekatan *supervised machine learning* untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu *Random Forest*, *Decision Tree*, dan *Naive Bayes*, dalam memprediksi status akademik mahasiswa. Proses penelitian dimulai dari identifikasi permasalahan, studi literatur, pengumpulan *dataset*, praproses data, pembangunan model, evaluasi model menggunakan metode 10-*Fold* Cross Validation, hingga analisis hasil pengujian. Tahapan penelitian tersebut dirancang agar proses eksperimen dapat dilakukan secara sistematis dan menghasilkan evaluasi yang objektif terhadap performa setiap algoritma [1], [4]. Alur penelitian yang dilakukan ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, penelitian diawali dengan identifikasi permasalahan mengenai tingginya risiko mahasiswa mengalami *dropout* maupun perubahan status akademik selama masa studi. Selanjutnya dilakukan studi literatur untuk memperoleh informasi mengenai penerapan algoritma klasifikasi pada bidang pendidikan tinggi. *Dataset* yang digunakan berasal dari *UCI Machine Learning Repository* dengan nama *Predict Students Dropout and Academic Success* yang terdiri atas 4.424 data mahasiswa dan 36 atribut prediktor [1], [5].

Tahap berikutnya adalah *preprocessing* data untuk memastikan atribut yang digunakan sesuai dengan kebutuhan pemodelan. Setelah proses tersebut selesai, dilakukan pembangunan model menggunakan tiga algoritma klasifikasi, yaitu *Random Forest*, *Decision Tree*, dan *Naive Bayes*. Ketiga algoritma dipilih karena mewakili pendekatan *ensemble learning*, pohon keputusan, dan probabilistik yang memiliki karakteristik berbeda dalam melakukan proses klasifikasi [2], [4], [6].

Evaluasi model dilakukan menggunakan metode 10-*Fold* Cross Validation. Metode ini membagi *dataset* menjadi sepuluh bagian dengan sembilan bagian digunakan sebagai data pelatihan dan satu bagian sebagai data pengujian secara bergantian hingga seluruh data memperoleh kesempatan menjadi data uji. Teknik tersebut banyak digunakan karena mampu menghasilkan estimasi performa model yang lebih stabil dibandingkan metode pembagian data sederhana (*hold-out validation*) [8].

Hasil evaluasi kemudian dianalisis menggunakan beberapa metrik, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, *Area Under Curve* (*AUC*), *Matthews Correlation Coefficient* (*MCC*), *Confusion Matrix*, dan *ROC Curve*. Penggunaan beberapa metrik evaluasi bertujuan memberikan gambaran yang lebih komprehensif mengenai kemampuan masing-masing algoritma dalam mengklasifikasikan status akademik mahasiswa sehingga dapat ditentukan algoritma yang memiliki performa terbaik [3], [8], [9].

2.2 Kajian Metode Penelitian

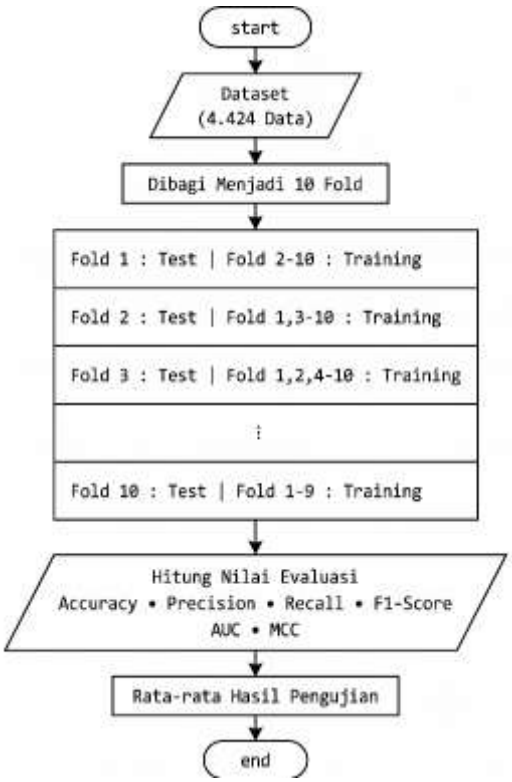
2.2.1 Metode 10-*Fold* Cross Validation

Metode 10-Fold Cross Validation merupakan salah satu teknik validasi yang paling banyak digunakan dalam penelitian machine learning untuk mengevaluasi performa model klasifikasi. Teknik ini bertujuan memperoleh estimasi kemampuan model secara lebih objektif dengan memanfaatkan seluruh data sebagai data pelatihan (training data) maupun data pengujian (testing data) secara bergantian [1], [10]. Berbeda dengan metode hold-out validation yang hanya melakukan satu kali pembagian data, 10-Fold Cross Validation melakukan proses evaluasi berulang sehingga hasil pengujian menjadi lebih stabil dan memiliki tingkat generalisasi yang lebih baik.

Pada metode ini, seluruh dataset dibagi menjadi 10 bagian (fold) yang memiliki ukuran relatif sama. Pada iterasi pertama, satu fold digunakan sebagai data pengujian sedangkan sembilan fold lainnya digunakan sebagai data pelatihan. Setelah proses pelatihan dan pengujian selesai dilakukan, fold berikutnya digunakan sebagai data pengujian hingga seluruh fold memperoleh kesempatan menjadi data uji tepat satu kali. Dengan demikian setiap data akan digunakan sebagai data pelatihan sebanyak sembilan kali dan sebagai data pengujian sebanyak satu kali. Hasil evaluasi dari seluruh iterasi kemudian dirata-ratakan sehingga menghasilkan nilai performa akhir model yang lebih representatif dibandingkan

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

hanya menggunakan satu kali pembagian data [9], [10]. Pada penelitian ini metode 10-Fold Cross Validation diterapkan terhadap dataset Predict Students Dropout and Academic Success yang terdiri atas 4.424 data mahasiswa. Dataset tersebut dibagi menjadi sepuluh fold dengan ukuran yang hampir sama. Selanjutnya algoritma Random Forest, Decision Tree, dan Naive Bayes dilatih menggunakan sembilan fold, kemudian diuji menggunakan satu fold yang tersisa. Setelah iterasi pertama selesai, proses diulang dengan mengganti fold pengujian hingga seluruh fold pernah digunakan sebagai data pengujian. Prosedur tersebut memastikan bahwa setiap algoritma memperoleh kondisi pelatihan dan pengujian yang sama sehingga hasil perbandingan menjadi lebih adil dan objektif. Ilustrasi penerapan metode 10-Fold Cross Validation pada penelitian ini ditunjukkan pada Gambar 2.



Gambar 2. Proses 10-Fold Cross Validation

Berdasarkan Gambar 2, dataset terlebih dahulu dibagi menjadi sepuluh fold dengan ukuran yang relatif sama. Pada setiap iterasi satu fold digunakan sebagai data pengujian sedangkan sembilan fold lainnya digunakan sebagai data pelatihan. Proses tersebut diulang sebanyak sepuluh kali hingga seluruh fold memperoleh kesempatan menjadi data uji. Seluruh nilai evaluasi kemudian dirata-ratakan untuk memperoleh performa akhir masing-masing algoritma. Tahapan pelaksanaan metode 10-Fold Cross Validation pada penelitian ini adalah sebagai berikut:

a. Dataset dibagi menjadi sepuluh fold yang berukuran relatif sama.
b. Pada iterasi pertama, sembilan fold digunakan sebagai data pelatihan dan satu fold digunakan sebagai data pengujian.
c. Model klasifikasi dibangun menggunakan data pelatihan sesuai algoritma yang digunakan.
d. Model yang telah terbentuk kemudian diuji menggunakan data pengujian.
e. Nilai evaluasi berupa Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC) dihitung.
f. Proses diulangi hingga seluruh fold pernah digunakan sebagai data pengujian.

Seluruh nilai evaluasi dirata-ratakan sehingga diperoleh performa akhir masing-masing algoritma. Secara matematis, nilai rata-rata hasil Cross Validation dapat dihitung menggunakan Persamaan (1).

$$CV = \frac{1}{k}\sum_{i=1}^{k} M_i$$

Dengan CV, merupakan nilai rata-rata hasil evaluasi (Cross Validation) yang digunakan sebagai nilai performa akhir model klasifikasi. Nilai ini diperoleh dari rata-rata seluruh hasil pengujian pada setiap fold sehingga mampu menggambarkan performa model secara keseluruhan. k, menyatakan jumlah fold yang digunakan pada proses Cross Validation. Pada penelitian ini digunakan k = 10, sehingga proses pelatihan dan pengujian dilakukan sebanyak sepuluh iterasi. Mi, merupakan nilai metrik evaluasi yang diperoleh pada iterasi atau fold ke-i. Nilai tersebut dapat berupa Accuracy, Precision, Recall, F1-Score, AUC, maupun Matthews Correlation Coefficient (MCC) yang dihasilkan pada setiap proses pengujian. Berdasarkan Persamaan (1), nilai evaluasi akhir diperoleh dengan menjumlahkan seluruh nilai hasil pengujian pada sepuluh iterasi, kemudian dibagi dengan jumlah fold yang digunakan. Proses perhitungan rata-rata

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

ini bertujuan untuk mengurangi pengaruh pembagian data secara acak sehingga hasil evaluasi menjadi lebih stabil, objektif, dan mampu merepresentasikan kemampuan model ketika diterapkan pada data baru. Oleh karena itu, metode 10-Fold Cross Validation dipilih karena memberikan estimasi performa yang lebih andal dibandingkan metode pembagian data tunggal (hold-out validation) serta telah menjadi salah satu teknik evaluasi yang paling banyak digunakan dalam penelitian machine learning.

Pada penelitian ini digunakan nilai k=10 sehingga setiap algoritma dievaluasi sebanyak sepuluh kali. Nilai evaluasi akhir diperoleh dari rata-rata seluruh iterasi sehingga mampu mengurangi bias akibat pembagian data secara acak dan menghasilkan estimasi performa model yang lebih konsisten [9], [11]. Oleh karena itu, metode 10-Fold Cross Validation dipilih karena mampu memberikan keseimbangan yang baik antara efisiensi komputasi dan keandalan evaluasi, serta telah banyak digunakan sebagai standar dalam penelitian klasifikasi berbasis machine learning.

3. HASIL DAN PEMBAHASAN.

3.1 Gambaran Dataset

Untuk melihat distribusi kelas target pada *dataset* penelitian, hasil visualisasi disajikan pada Gambar 1.

Tabel 1. Karakteristik *Dataset* Penelitian

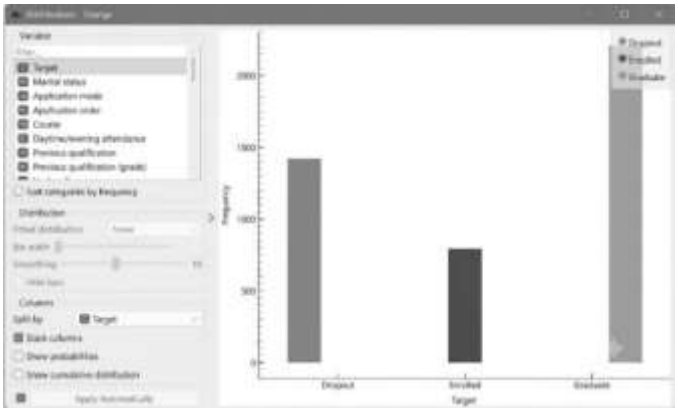
Karakteristik	Keterangan
Sumber *Dataset*	*UCI Machine Learning Repository*
Nama *Dataset*	*Predict Students Dropout and Academic Success*
Jumlah Data	4.424
Jumlah Atribut Prediktor	36
Jumlah Target	1
Jumlah Kelas	3
Label Target	*Dropout*, *Enrolled*, *Graduate*
Missing Value	Tidak ada

Berdasarkan Tabel 1, *dataset* yang digunakan tidak memiliki nilai missing value, sehingga data dapat langsung digunakan pada tahap pemodelan tanpa memerlukan proses imputasi data. Seluruh atribut numerik maupun kategorikal telah tersedia dalam format yang sesuai sehingga proses *preprocessing* hanya difokuskan pada pemilihan atribut target dan atribut prediktor menggunakan perangkat lunak *Orange Data Mining*. Kondisi tersebut memberikan keuntungan karena model dapat dibangun menggunakan data yang bersih sehingga hasil evaluasi lebih merepresentasikan kemampuan algoritma dibandingkan kualitas data.

Selain memiliki jumlah data yang cukup besar, *dataset* ini juga mempunyai distribusi atribut yang beragam, mulai dari informasi demografi mahasiswa, latar belakang pendidikan, kondisi sosial ekonomi, hingga performa akademik pada semester pertama dan semester kedua. Keragaman atribut tersebut memungkinkan algoritma klasifikasi mempelajari pola hubungan antarvariabel yang memengaruhi status akademik mahasiswa. Oleh karena itu, *dataset* ini banyak digunakan sebagai acuan dalam penelitian prediksi *dropout* maupun prediksi keberhasilan akademik pada pendidikan tinggi.

3.2 Analisis Karakteristik *dataset*

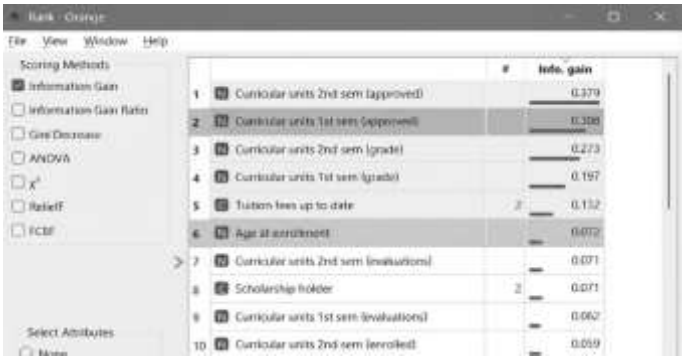
Sebelum dilakukan proses klasifikasi, dilakukan analisis terhadap karakteristik *dataset* untuk memperoleh gambaran mengenai distribusi kelas target dan tingkat kepentingan setiap atribut terhadap proses prediksi. Analisis ini bertujuan untuk mengetahui apakah terdapat atribut yang memiliki pengaruh dominan terhadap status akademik mahasiswa sehingga dapat membantu proses interpretasi hasil klasifikasi. Untuk mengetahui distribusi data pada masing-masing kelas target, dilakukan visualisasi menggunakan *widget Distributions* pada *Orange Data Mining*. Hasil visualisasi ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Kelas Target *Dataset*

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

Berdasarkan Gambar 3, *dataset* terdiri atas tiga kelas target, yaitu *Dropout*, *Enrolled*, dan *Graduate*. Distribusi data menunjukkan bahwa kelas *Graduate* memiliki jumlah data yang lebih banyak dibandingkan kelas lainnya, sedangkan kelas *Enrolled* memiliki jumlah data yang relatif lebih sedikit. Perbedaan distribusi tersebut menunjukkan bahwa *dataset* memiliki tingkat ketidakseimbangan (*class imbalance*) yang masih berada dalam batas wajar sehingga algoritma klasifikasi masih mampu mempelajari pola masing-masing kelas tanpa memerlukan teknik penyeimbangan data (*resampling*). Selain distribusi kelas, dilakukan pula analisis terhadap tingkat kepentingan atribut (*feature importance*) menggunakan metode *Information Gain*. Analisis ini bertujuan untuk mengetahui atribut yang memberikan kontribusi paling besar dalam proses klasifikasi. Hasil pemeringkatan atribut menggunakan metode *Information Gain* disajikan pada Gambar 4.



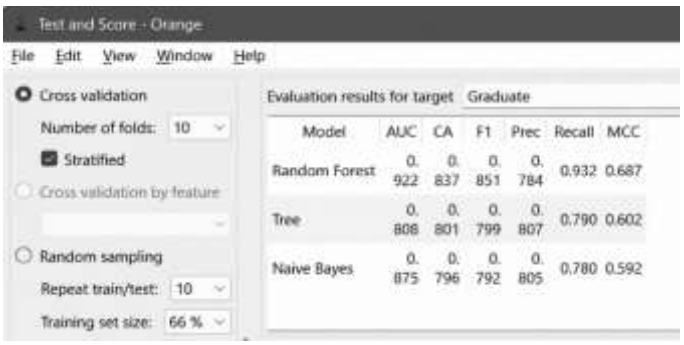
Gambar 4. Hasil Pemeringkatan Atribut Menggunakan *Information Gain*

Berdasarkan Gambar 4, atribut Curricular units 2nd semester (approved) memperoleh nilai *Information Gain* tertinggi sebesar 0,379, diikuti oleh Curricular units 1st semester (approved) sebesar 0,308, Curricular units 2nd semester (grade) sebesar 0,273, dan Curricular units 1st semester (grade) sebesar 0,197. Hasil tersebut menunjukkan bahwa performa akademik mahasiswa pada semester awal merupakan faktor yang paling berpengaruh terhadap status akademik mahasiswa. Selain atribut akademik, atribut Tuition fees up to date juga memiliki nilai *Information Gain* yang relatif tinggi, yaitu sebesar 0,132, sehingga menunjukkan bahwa kondisi pembayaran biaya kuliah juga berkontribusi terhadap prediksi status akademik mahasiswa. Temuan ini sejalan dengan beberapa penelitian terdahulu yang menyatakan bahwa faktor akademik dan faktor ekonomi merupakan dua variabel utama yang memengaruhi risiko mahasiswa mengalami *dropout*.

3.3 Penerapan Metode 10-Fold Cross Validation

Setelah proses preprocessing dan analisis karakteristik dataset selesai dilakukan, tahap berikutnya adalah mengevaluasi performa model menggunakan metode 10-Fold Cross Validation. Pada penelitian ini, metode tersebut diterapkan melalui widget Test & Score pada perangkat lunak Orange Data Mining untuk memastikan bahwa proses pengujian dilakukan secara objektif serta memanfaatkan seluruh data sebagai data pelatihan maupun data pengujian secara bergantian. Dataset Predict Students Dropout and Academic Success yang terdiri atas 4.424 data mahasiswa secara otomatis dibagi menjadi sepuluh fold dengan ukuran yang relatif sama. Pada iterasi pertama, satu fold digunakan sebagai data pengujian (testing data), sedangkan sembilan fold lainnya digunakan sebagai data pelatihan (training data). Selanjutnya model Random Forest, Decision Tree, dan Naive Bayes dibangun menggunakan data pelatihan, kemudian diuji menggunakan data pengujian untuk memperoleh nilai Accuracy, Precision, Recall, F1-Score, AUC, dan MCC.

Setelah iterasi pertama selesai dilakukan, sistem secara otomatis mengganti fold pengujian hingga seluruh fold memperoleh kesempatan menjadi data uji sebanyak satu kali. Dengan demikian, setiap data digunakan sebagai data pelatihan sebanyak sembilan kali dan sebagai data pengujian sebanyak satu kali. Nilai evaluasi dari seluruh iterasi kemudian dirata-ratakan sehingga menghasilkan estimasi performa model yang lebih stabil serta mengurangi pengaruh pembagian data secara acak. Proses penerapan metode 10-Fold Cross Validation pada penelitian ini ditunjukkan pada Gambar 5.



Gambar 5. Penerapan Metode 10-Fold Cross Validation pada Orange Data Mining

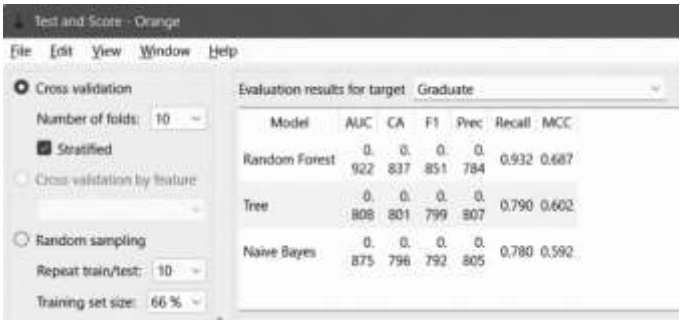
Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

Berdasarkan Gambar 5, seluruh algoritma dievaluasi menggunakan konfigurasi yang sama, yaitu metode Cross Validation dengan jumlah 10 fold. Penggunaan konfigurasi yang identik memastikan bahwa perbandingan performa antaralgoritma dilakukan pada kondisi eksperimen yang sama sehingga hasil evaluasi dapat dibandingkan secara adil. Nilai rata-rata hasil pengujian dari sepuluh iterasi selanjutnya digunakan sebagai dasar untuk menentukan algoritma yang memiliki performa terbaik dalam memprediksi status akademik mahasiswa.

3.4 Hasil Evaluasi Algoritma

Setelah proses pembangunan model selesai dilakukan, tahap berikutnya adalah mengevaluasi performa masing-masing algoritma klasifikasi menggunakan metode 10-*Fold* Cross Validation. Metode ini dipilih karena mampu menghasilkan estimasi performa model yang lebih objektif dengan memanfaatkan seluruh data sebagai data pelatihan maupun data pengujian secara bergantian. Dengan demikian, hasil evaluasi yang diperoleh tidak bergantung pada satu pembagian data tertentu sehingga dapat mengurangi risiko bias dalam proses pengujian.

Pada penelitian ini, tiga algoritma klasifikasi, yaitu *Random Forest*, *Decision Tree*, dan *Naive Bayes*, diterapkan pada *dataset* yang sama menggunakan konfigurasi eksperimen yang identik. Seluruh model dibangun menggunakan perangkat lunak *Orange Data Mining* dan dievaluasi berdasarkan beberapa metrik performa, yaitu *Accuracy*, *Area Under Curve* (*AUC*), *Precision*, *Recall*, *F1-Score*, dan *Matthews Correlation Coefficient* (*MCC*). Penggunaan beberapa metrik evaluasi bertujuan memberikan gambaran yang lebih komprehensif mengenai kemampuan setiap algoritma dalam melakukan klasifikasi status akademik mahasiswa. Untuk mengetahui hasil evaluasi masing-masing algoritma klasifikasi, dilakukan pengujian menggunakan *widget Test & Score* pada *Orange Data Mining*. Hasil pengujian ditunjukkan pada Gambar 6.



Gambar 6. Hasil Evaluasi Algoritma Menggunakan Widget Test & Score

Berdasarkan Gambar 6, setiap algoritma menghasilkan nilai performa yang berbeda pada seluruh metrik evaluasi. Perbedaan tersebut menunjukkan bahwa karakteristik masing-masing algoritma memengaruhi kemampuan model dalam mengenali pola hubungan antar atribut pada *dataset* prediksi status akademik mahasiswa. Untuk mempermudah proses analisis, hasil evaluasi masing-masing algoritma disajikan kembali pada Tabel 2.

Tabel 2. Hasil Evaluasi Algoritma

Algoritma	*AUC*	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*MCC*
Random Forest	0.897	0.774	0.760	0.774	0.759	0.624
Decision Tree	0.753	0.702	0.700	0.702	0.701	0.517
Naive Bayes	0.847	0.697	0.710	0.697	0.703	0.514

Berdasarkan Tabel 2, algoritma *Random Forest* memperoleh performa terbaik dibandingkan algoritma lainnya pada hampir seluruh metrik evaluasi. Algoritma ini menghasilkan nilai *Accuracy* sebesar 77,4%, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data mahasiswa ke dalam kelas yang benar. Selain itu, nilai *AUC* sebesar 0,897 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan ketiga kelas target, yaitu *Dropout*, *Enrolled*, dan *Graduate*. Nilai *AUC* yang mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang tinggi dan mampu menghasilkan prediksi yang konsisten pada berbagai nilai ambang (*threshold*).

Dari sisi *Precision*, algoritma *Random Forest* memperoleh nilai sebesar 0,760, yang menunjukkan bahwa sebagian besar prediksi yang dihasilkan oleh model sesuai dengan kondisi sebenarnya. Nilai *Recall* sebesar 0,774 menunjukkan bahwa model mampu mengenali sebagian besar data pada setiap kelas target, sedangkan nilai *F1-Score* sebesar 0,759 menunjukkan keseimbangan yang baik antara *Precision* dan *Recall*. Keseimbangan tersebut menjadi indikator bahwa model tidak hanya memiliki tingkat ketepatan prediksi yang tinggi, tetapi juga mampu meminimalkan kesalahan klasifikasi pada masing-masing kelas.

Nilai *Matthews Correlation Coefficient* (*MCC*) yang diperoleh *Random Forest* sebesar 0,624 juga merupakan nilai tertinggi dibandingkan dua algoritma lainnya. *MCC* merupakan salah satu metrik evaluasi yang mampu menggambarkan kualitas model secara menyeluruh karena mempertimbangkan seluruh elemen pada *confusion matrix*. Semakin tinggi nilai *MCC*, semakin baik kemampuan model dalam melakukan klasifikasi secara konsisten. Dengan demikian, hasil penelitian menunjukkan bahwa *Random Forest* memiliki tingkat stabilitas yang lebih baik dibandingkan algoritma lainnya.

Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

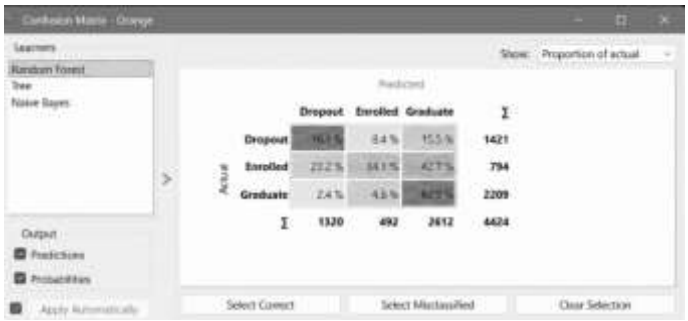
Sementara itu, algoritma *Decision Tree* memperoleh nilai *Accuracy* sebesar 70,2%, dengan *AUC* sebesar 0,753. Walaupun algoritma ini memiliki keunggulan dalam menghasilkan model yang mudah dipahami dan diinterpretasikan, performanya masih berada di bawah *Random Forest*. Hal tersebut disebabkan karena *Decision Tree* hanya membangun satu struktur pohon keputusan sehingga lebih sensitif terhadap variasi data pelatihan dan lebih rentan mengalami *overfitting*. Akibatnya, kemampuan generalisasi model menjadi lebih rendah ketika diterapkan pada data yang belum pernah dipelajari sebelumnya.

Algoritma *Naive Bayes* memperoleh nilai *Accuracy* sebesar 69,7%, *Precision* sebesar 0,710, dan *AUC* sebesar 0,847. Walaupun nilai *AUC* yang dihasilkan relatif tinggi, nilai *Accuracy* masih lebih rendah dibandingkan *Random Forest*. Kondisi tersebut menunjukkan bahwa asumsi independensi antar atribut yang digunakan oleh *Naive Bayes* belum sepenuhnya sesuai dengan karakteristik *dataset* pendidikan tinggi, karena sebagian besar atribut akademik memiliki hubungan yang saling memengaruhi. Akibatnya, model tidak mampu mengenali seluruh pola hubungan antar atribut secara optimal.

Secara keseluruhan, hasil pengujian menunjukkan bahwa algoritma *Random Forest* merupakan algoritma dengan performa terbaik pada penelitian ini. Hal tersebut dibuktikan melalui nilai *Accuracy*, *AUC*, *Recall*, *F1-Score*, dan *MCC* yang lebih tinggi dibandingkan *Decision Tree* maupun *Naive Bayes*. Pendekatan *ensemble learning* yang digunakan oleh *Random Forest* memungkinkan model membangun banyak pohon keputusan (*decision tree*) secara independen, kemudian menggabungkan hasil prediksi melalui mekanisme *majority voting*. Strategi tersebut mampu meningkatkan kemampuan generalisasi model sekaligus mengurangi risiko *overfitting*, sehingga menghasilkan performa klasifikasi yang lebih stabil pada *dataset* prediksi status akademik mahasiswa.

3.5 Analisis *Confusion Matrix*

Setelah diperoleh hasil evaluasi menggunakan metode 10-*Fold* Cross Validation, dilakukan analisis lebih lanjut terhadap algoritma dengan performa terbaik, yaitu *Random Forest*, menggunakan *Confusion Matrix*. Analisis ini bertujuan untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kelas target secara lebih rinci serta mengidentifikasi pola kesalahan klasifikasi yang masih terjadi. Melalui *Confusion Matrix*, dapat diketahui jumlah data yang berhasil diprediksi dengan benar maupun data yang mengalami kesalahan klasifikasi pada masing-masing kelas sehingga memberikan gambaran yang lebih komprehensif mengenai kualitas model. Untuk menganalisis kemampuan klasifikasi algoritma *Random Forest* pada setiap kelas target, hasil *Confusion Matrix* ditunjukkan pada Gambar 7.



Gambar 7. Hasil *Confusion Matrix* Algoritma *Random Forest*

Berdasarkan Gambar 7, algoritma *Random Forest* mampu mengklasifikasikan sebagian besar data mahasiswa ke dalam kelas yang sesuai dengan kondisi sebenarnya. Nilai yang berada pada diagonal utama *confusion matrix* menunjukkan jumlah data yang berhasil diprediksi dengan benar, sedangkan nilai di luar diagonal menunjukkan jumlah data yang mengalami kesalahan klasifikasi (*misclassification*). Semakin besar nilai pada diagonal utama, semakin baik kemampuan model dalam melakukan klasifikasi. Pada kelas *Graduate*, model menunjukkan kemampuan klasifikasi yang sangat baik. Sebagian besar mahasiswa yang benar-benar berada pada kelas *Graduate* berhasil diprediksi secara tepat oleh model. Hal ini menunjukkan bahwa karakteristik mahasiswa yang berhasil menyelesaikan studi memiliki pola yang cukup jelas sehingga mudah dikenali oleh algoritma *Random Forest*. Faktor-faktor seperti jumlah mata kuliah yang lulus, nilai akademik pada semester pertama dan kedua, serta status pembayaran biaya kuliah memberikan kontribusi besar terhadap keberhasilan proses klasifikasi.

Pada kelas *Dropout*, algoritma juga mampu menghasilkan tingkat klasifikasi yang tinggi. Sebagian besar mahasiswa yang benar-benar mengalami *dropout* berhasil dikenali sebagai kelas *Dropout*. Hasil ini menunjukkan bahwa *Random Forest* mampu memanfaatkan kombinasi berbagai atribut akademik dan sosial ekonomi untuk membedakan mahasiswa yang berisiko berhenti kuliah dari mahasiswa yang masih aktif maupun yang telah lulus. Kemampuan tersebut sangat penting apabila model diterapkan sebagai dasar pengembangan *early warning system* pada perguruan tinggi. Berbeda dengan dua kelas sebelumnya, kelas *Enrolled* masih menunjukkan beberapa kesalahan klasifikasi. Sebagian data mahasiswa yang seharusnya termasuk dalam kelas *Enrolled* diprediksi sebagai *Graduate* maupun *Dropout*. Kondisi tersebut dapat terjadi karena karakteristik mahasiswa yang masih aktif sering kali memiliki atribut yang berada di antara kedua kelas lainnya. Sebagai contoh, mahasiswa yang masih aktif dapat memiliki performa akademik yang baik sehingga menyerupai mahasiswa *Graduate*, atau sebaliknya memiliki performa akademik yang rendah sehingga lebih mendekati

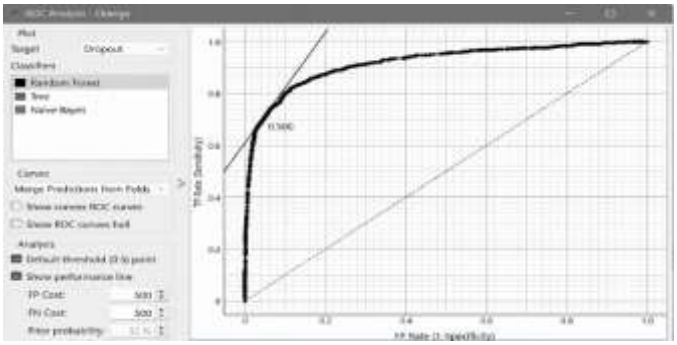
Jurnal Ilmu Komputer, Teknologi Dan Informasi
Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

karakteristik mahasiswa *Dropout*. Akibatnya, proses pemisahan kelas *Enrolled* menjadi lebih kompleks dibandingkan dua kelas lainnya.

Secara keseluruhan, hasil *Confusion Matrix* menunjukkan bahwa algoritma *Random Forest* memiliki kemampuan klasifikasi yang baik terhadap ketiga kelas target. Walaupun masih terdapat beberapa kesalahan prediksi pada kelas *Enrolled*, jumlah kesalahan tersebut relatif kecil dibandingkan jumlah prediksi yang benar. Temuan ini konsisten dengan nilai *Accuracy*, *Recall*, dan *F1-Score* yang diperoleh pada tahap evaluasi sebelumnya. Dengan demikian, analisis *Confusion Matrix* memperkuat bahwa *Random Forest* merupakan algoritma yang paling efektif dalam memprediksi status akademik mahasiswa pada *dataset* yang digunakan.

3.6 Analisis ROC Curve

Selain menggunakan *Confusion Matrix*, penelitian ini juga mengevaluasi performa algoritma *Random Forest* melalui analisis *Receiver Operating Characteristic* (ROC) Curve. Kurva ROC digunakan untuk mengukur kemampuan model dalam membedakan setiap kelas target pada berbagai nilai ambang (*threshold*). Semakin dekat kurva ROC terhadap sudut kiri atas grafik, semakin baik kemampuan model dalam melakukan klasifikasi. Selain itu, nilai *Area Under Curve* (*AUC*) digunakan sebagai indikator kuantitatif untuk mengukur kualitas model secara keseluruhan [6], [7]. Untuk mengevaluasi kemampuan diskriminasi model pada berbagai nilai *threshold*, hasil analisis *ROC Curve* algoritma *Random Forest* disajikan pada Gambar 8.



Gambar 8. Kurva ROC Algoritma *Random Forest*

Berdasarkan Gambar 8, kurva ROC algoritma *Random Forest* berada jauh di atas garis diagonal yang merepresentasikan klasifikasi acak (*random classifier*). Posisi kurva tersebut menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan setiap kelas target berdasarkan atribut yang dimiliki mahasiswa. Semakin dekat kurva terhadap sudut kiri atas, semakin tinggi sensitivitas model dalam mengenali kelas yang benar dengan tingkat kesalahan klasifikasi yang rendah. Nilai *AUC* sebesar 0,897 yang diperoleh algoritma *Random Forest* menunjukkan bahwa model termasuk dalam kategori *excellent classification*. Nilai tersebut mengindikasikan bahwa probabilitas model dalam membedakan data mahasiswa yang berasal dari kelas yang berbeda mencapai hampir 90%. Dengan kata lain, model memiliki kemampuan diskriminasi yang tinggi sehingga mampu menghasilkan prediksi yang konsisten pada berbagai kondisi data.

Jika dibandingkan dengan algoritma *Decision Tree* dan *Naive Bayes*, *Random Forest* menghasilkan nilai *AUC* yang lebih tinggi. Hal ini menunjukkan bahwa pendekatan *ensemble learning* yang digunakan *Random Forest* mampu meningkatkan kemampuan model dalam mengenali pola hubungan antar atribut dibandingkan algoritma pohon keputusan tunggal maupun algoritma probabilistik. Penggabungan banyak pohon keputusan melalui mekanisme *majority voting* membuat model lebih stabil terhadap variasi data dan mengurangi risiko *overfitting*. Hasil analisis ROC juga memperkuat temuan pada *Confusion Matrix* dan hasil evaluasi *Test & Score*. Ketiga analisis tersebut secara konsisten menunjukkan bahwa *Random Forest* memiliki performa terbaik dibandingkan *Decision Tree* dan *Naive Bayes*. Dengan demikian, algoritma *Random Forest* layak dipertimbangkan sebagai metode klasifikasi untuk pengembangan sistem prediksi status akademik mahasiswa maupun sistem peringatan dini (*early warning system*) di lingkungan perguruan tinggi.

3.7 Pembahasan

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memberikan performa terbaik dibandingkan *Decision Tree* dan *Naive Bayes* dalam memprediksi status akademik mahasiswa. Berdasarkan hasil pengujian menggunakan metode 10-*Fold* Cross Validation, algoritma *Random Forest* memperoleh nilai *Accuracy* sebesar 77,4%, *Precision* sebesar 0,760, *Recall* sebesar 0,774, *F1-Score* sebesar 0,759, *Area Under Curve* (*AUC*) sebesar 0,897, serta *Matthews Correlation Coefficient* (*MCC*) sebesar 0,624. Seluruh nilai tersebut lebih tinggi dibandingkan algoritma *Decision Tree* maupun *Naive Bayes*, sehingga menunjukkan bahwa pendekatan *ensemble learning* yang diterapkan oleh *Random Forest* mampu menghasilkan model klasifikasi yang lebih akurat dan lebih stabil. Keunggulan *Random Forest* pada penelitian ini disebabkan oleh mekanisme pembentukan banyak pohon keputusan (*decision tree*) yang dibangun menggunakan teknik *bootstrap sampling* dan pemilihan atribut secara acak (*random feature selection*). Setiap pohon menghasilkan prediksi secara independen, kemudian seluruh hasil prediksi digabungkan melalui mekanisme *majority voting*. Pendekatan tersebut mampu mengurangi variansi model serta meminimalkan risiko *overfitting* yang sering terjadi pada

Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

algoritma *Decision Tree* tunggal. Oleh karena itu, *Random Forest* mampu menghasilkan kemampuan generalisasi yang lebih baik ketika diterapkan pada data yang belum pernah dipelajari sebelumnya.

Hasil penelitian ini sejalan dengan penelitian yang dilakukan oleh Villar dan de Andrade, [2] yang melaporkan bahwa algoritma berbasis *ensemble learning* menghasilkan performa klasifikasi yang lebih tinggi dibandingkan algoritma klasifikasi tunggal pada prediksi student *dropout*. Penelitian tersebut menunjukkan bahwa kombinasi beberapa pohon keputusan mampu meningkatkan stabilitas model serta mengurangi kesalahan klasifikasi pada data pendidikan tinggi yang memiliki karakteristik kompleks. Temuan tersebut mendukung hasil penelitian ini yang menunjukkan bahwa *Random Forest* memiliki nilai *Accuracy* dan *AUC* tertinggi dibandingkan dua algoritma lainnya.

Hasil penelitian ini juga konsisten dengan penelitian Putra et al.[7] yang membandingkan beberapa algoritma klasifikasi pada *dataset* prediksi mahasiswa *dropout*. Penelitian tersebut menyimpulkan bahwa *Random Forest* memiliki kemampuan yang lebih baik dalam mengenali pola hubungan antar atribut akademik dibandingkan algoritma berbasis pohon keputusan tunggal. Kemampuan tersebut berasal dari proses pembentukan banyak pohon keputusan sehingga model menjadi lebih tahan terhadap perubahan data pelatihan dan menghasilkan prediksi yang lebih konsisten.

Penelitian Rebelo Marcolino et al.[5] juga menunjukkan hasil yang serupa. Mereka menyatakan bahwa data pendidikan memiliki hubungan antar atribut yang kompleks, terutama pada atribut yang berkaitan dengan performa akademik mahasiswa. Oleh karena itu, algoritma berbasis ensemble lebih mampu menangkap hubungan antar atribut dibandingkan algoritma yang hanya menggunakan satu model klasifikasi. Pada penelitian ini, atribut seperti Curricular units 2nd semester (approved), Curricular units 1st semester (approved), serta Curricular units 2nd semester (grade) terbukti menjadi atribut dengan nilai *Information Gain* tertinggi, sehingga memberikan kontribusi besar terhadap keberhasilan proses klasifikasi.

Sementara itu, algoritma *Decision Tree* memperoleh nilai *Accuracy* sebesar 70,2%, lebih rendah dibandingkan *Random Forest*. Walaupun algoritma ini memiliki keunggulan dalam menghasilkan model yang mudah dipahami melalui struktur pohon keputusan, kemampuan generalisasi model masih terbatas karena hanya bergantung pada satu pohon keputusan. Kondisi tersebut menyebabkan model lebih sensitif terhadap perubahan data pelatihan sehingga berpotensi mengalami *overfitting*. Hasil penelitian ini mendukung penelitian Utami et al.[4], yang menyatakan bahwa performa *Decision Tree* cenderung menurun ketika diterapkan pada *dataset* yang memiliki jumlah atribut cukup banyak dan pola hubungan data yang kompleks.

Algoritma *Naive Bayes* memperoleh nilai *Accuracy* sebesar 69,7% dengan *AUC* sebesar 0,847. Walaupun algoritma ini memiliki proses komputasi yang cepat serta mampu menghasilkan model dengan kompleksitas rendah, performanya masih berada di bawah *Random Forest*. Hal tersebut disebabkan oleh asumsi independensi antar atribut yang digunakan oleh *Naive Bayes*. Pada *dataset* pendidikan tinggi, sebagian besar atribut memiliki hubungan yang saling memengaruhi, misalnya antara jumlah mata kuliah yang lulus, nilai akademik, serta status pembayaran biaya kuliah. Akibatnya, asumsi independensi tersebut tidak sepenuhnya terpenuhi sehingga kemampuan klasifikasi model menjadi berkurang. Temuan ini sejalan dengan penelitian Atika [3] yang menyatakan bahwa *Naive Bayes* lebih sesuai digunakan pada *dataset* dengan hubungan antar atribut yang relatif sederhana.

Analisis *Confusion Matrix* pada penelitian ini menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan sebagian besar data mahasiswa ke dalam kelas yang benar. Kesalahan klasifikasi yang masih terjadi terutama ditemukan pada kelas *Enrolled*, karena karakteristik mahasiswa yang masih aktif sering kali memiliki atribut yang menyerupai kelas *Graduate* maupun *Dropout*. Kondisi tersebut menyebabkan proses pemisahan kelas menjadi lebih sulit dibandingkan dua kelas lainnya. Walaupun demikian, jumlah kesalahan klasifikasi relatif kecil sehingga tidak memberikan pengaruh yang signifikan terhadap performa model secara keseluruhan.

Hasil analisis *ROC Curve* juga memperkuat hasil evaluasi sebelumnya. Nilai *AUC* sebesar 0,897 menunjukkan bahwa algoritma *Random Forest* memiliki kemampuan diskriminasi yang sangat baik dalam membedakan ketiga kelas target. Kurva ROC yang berada jauh di atas garis diagonal menunjukkan bahwa model memiliki sensitivitas dan spesifisitas yang tinggi. Dengan demikian, hasil *ROC Curve* konsisten dengan hasil *Accuracy*, *F1-Score*, dan *Confusion Matrix*, sehingga memperkuat kesimpulan bahwa *Random Forest* merupakan algoritma yang paling sesuai untuk prediksi status akademik mahasiswa pada *dataset* yang digunakan.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa penggunaan algoritma *Random Forest* memberikan keseimbangan terbaik antara akurasi, kemampuan generalisasi, serta stabilitas model dibandingkan *Decision Tree* dan *Naive Bayes*. Selain menghasilkan nilai evaluasi yang lebih tinggi, algoritma ini juga menunjukkan kemampuan yang lebih baik dalam mengenali pola hubungan antar atribut akademik, sosial, dan ekonomi mahasiswa. Oleh karena itu, *Random Forest* memiliki potensi yang besar untuk diterapkan sebagai dasar pengembangan Sistem Pendukung Keputusan (SPK) maupun *Early Warning System* (EWS) di perguruan tinggi guna mengidentifikasi mahasiswa yang berisiko mengalami penurunan status akademik atau *dropout* secara lebih dini.

4. KESIMPULAN

Penelitian ini berhasil membandingkan kinerja algoritma Random Forest, Decision Tree, dan Naive Bayes dalam memprediksi status akademik mahasiswa menggunakan dataset Predict Students Dropout and Academic Success dengan pendekatan 10-Fold Cross Validation sebagai metode evaluasi. Penerapan metode tersebut memungkinkan seluruh data memperoleh kesempatan yang sama sebagai data pelatihan dan data pengujian sehingga menghasilkan proses evaluasi

Vol. 4, No. 2, Juli 2026, Hal 86 - 96
ISSN: **2963-0169** (Online - Elektronik)
DOI: 10.62866/jurikti.v4i2.367
https://journal.grahamitra.id/index.php/jurikti

yang lebih objektif, konsisten, dan mampu mengurangi bias akibat pembagian data secara acak. Berdasarkan keseluruhan proses penelitian, algoritma Random Forest terbukti memberikan performa terbaik dibandingkan dua algoritma lainnya karena mampu menghasilkan model yang lebih stabil dalam mengklasifikasikan tiga kategori status akademik mahasiswa, yaitu Dropout, Enrolled, dan Graduate. Hasil penelitian juga menunjukkan bahwa atribut yang berkaitan dengan performa akademik pada semester awal serta status pembayaran biaya kuliah merupakan faktor yang paling berpengaruh terhadap keberhasilan proses prediksi, sehingga informasi tersebut dapat dimanfaatkan sebagai indikator dalam mendeteksi mahasiswa yang berpotensi mengalami penurunan status akademik sejak dini. Temuan ini menunjukkan bahwa pemanfaatan teknik machine learning, khususnya algoritma Random Forest, memiliki potensi yang baik untuk mendukung pengambilan keputusan berbasis data di lingkungan perguruan tinggi melalui pengembangan Early Warning System (EWS) atau sistem pendukung keputusan yang mampu membantu pihak akademik dalam merancang strategi pencegahan dropout secara lebih tepat sasaran. Dengan demikian, penelitian ini tidak hanya memberikan perbandingan performa tiga algoritma klasifikasi, tetapi juga memberikan kontribusi praktis dalam mendukung peningkatan kualitas layanan akademik melalui penerapan analisis prediktif pada data mahasiswa.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas HKBP Nommensen Pematangsiantar, atas dukungan akademik yang diberikan selama pelaksanaan penelitian ini. Penulis juga menyampaikan apresiasi kepada dosen pengampu mata kuliah *Artificial Intelligence* atas arahan, masukan, dan bimbingan yang diberikan selama proses penyusunan penelitian. Selain itu, penulis mengucapkan terima kasih *UCI Machine Learning Repository* yang telah menyediakan *Dataset Predict Students Dropout and Academic Success* sehingga penelitian ini dapat dilaksanakan dengan baik.

REFERENCES

[1] D. P. Morán, M. M. Goméz, and V. Yeste, *Identifying Key Factors of Student Dropout through Random Forest: A Data-Driven Approach*, vol. 2025, no. Icomta. Atlantis Press International BV, 2025. doi: 10.2991/978-94-6463-868-4_46.

[2] A. Villar and C. R. V. de Andrade, "*Supervised Machine Learning* algorithms for predicting student *dropout* and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, 2024, doi: 10.1007/s44163-023-00079-z.

[3] P. D. Atika, "A Comparative Study of *Machine Learning*-Based Student *Dropout* Risk Prediction," vol. 14, no. 225, pp. 167–174, 2026, doi: 10.33558/piksel.v14i1.12299.

[4] S. G. A. Utami, H. Setiadi, and A. Rohmadi, "Comparative Analysis of *Machine Learning* Algorithms with RFE-CV for Student *Dropout* Prediction," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 3, pp. 1319–1338, 2025, doi: 10.52436/1.jutif.2025.6.3.4695.

[5] M. Rebelo Marcolino *et al.*, "Student *dropout* prediction through *Machine Learning* optimization: insights from moodle log data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-025-93918-1.

[6] A. Jain, A. K. Dubey, S. Khan, A. Panwar, M. Alkhatib, and A. M. Alshahrani, "A PSO weighted ensemble framework with SMOTE balancing for student *dropout* prediction in smart education systems," *Sci. Rep.*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.1038/s41598-025-97506-1.

[7] L. G. R. Putra, D. D. Prasetya, and M. Mayadi, "Student *Dropout* Prediction Using *Random Forest* and *XGBoost* Method," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.

[8] S. A. Sulak and N. Köklü, "Predicting Student *Dropout* Using *Machine Learning* Algorithms," Intelligent Methods in Engineering Sciences, vol. 3, no. 3, pp. 91–98, 2024. doi: 10.58190/imiens.2024.103

[9] W. Winarsih, H. Sutanto, and A. P. Widodo, "Implementation of the Ensemble *Machine Learning* Algorithm for Student *Dropout* Prediction Analysis," Jurnal Sistem Informasi Bisnis, vol. 15, no. 2, pp. 159–166, 2025. doi: 10.14710/vol15iss2pp159-166

[10] T. A. Marzuqi, E. Kristiani, and Marcel, "Prediksi Mahasiswa Drop-Out di Universitas XYZ," Jurnal Teknologi Informasi dan Ilmu Komputer, 2024. doi: 10.25126/jtiik.2024118689

[11] N. Puspitasari, M. A. Wibowo, and B. Warsito, "*Dropout* Prediction Using KNN, *Decision Tree*, *Naive Bayes*, and *Ensemble Learning*," Jurnal Sisfokom, vol. 15, no. 2, 2025. doi: 10.32736/sisfokom.v15i02.2591

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

[13] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal, "An Explainable *Machine Learning* Approach for Student *Dropout* Prediction," Expert Systems with Applications, vol. 233, 2023. doi: 10.1016/j.eswa.2023.120933

[14] M. Vaarma and H. Li, "Predicting Student *Dropout*s with *Machine Learning*: An Empirical Study in Finnish Higher Education," Technology in Society, 2024. doi: 10.1016/j.techsoc.2024.102474

[15] J. Nieuwoudt and M. Pedler, "Student Retention in Higher Education: Why Students Choose to Remain at University," Journal of College Student Retention: Research, Theory & Practice, vol. 25, no. 2, pp. 326–349, 2023. doi: 10.1177/1521025120985228