P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Klasifikasi Risiko Dropout Mahasiswa Menggunakan Algoritma Random Forest pada Dataset Predict Students' Dropout and Academic Success

Novita Sari Siagian[1],Monica Sari Batubara [2], Leony Sinaga[3]

[1]Program Studi Ilmu Komputer , Fakultas Matematika dan Ilmu Pengetahuan, Universitas HKBP Nommensen Pematang Siantar, Indonesia

[2] Program Studi Ilmu Komputer , Fakultas Matematika dan Ilmu Pengetahuan, Universitas HKBP Nommensen Pematang Siantar, Indonesia

Email: [1] novitasarisiagian44@gmail.com,[2] batubaramonica9@gmail.com, [3] leonysinaga30@gmail.com

Email Responden : (novitasarisiagian44@gmail.com)

ABSTRAK

Tingginya angka dropout mahasiswa merupakan salah satu permasalahan yang dihadapi perguruan tinggi karena berdampak terhadap kualitas lulusan, efektivitas penyelenggaraan pendidikan, serta tingkat retensi mahasiswa. Oleh karena itu, diperlukan suatu metode yang mampu mengidentifikasi mahasiswa yang berisiko dropout secara lebih dini berdasarkan data akademik yang tersedia. Penelitian ini bertujuan mengklasifikasikan risiko dropout mahasiswa menggunakan algoritma Random Forest dengan memanfaatkan dataset Predict Students' Dropout and Academic Success yang diperoleh dari UCI Machine Learning Repository. Penelitian menggunakan pendekatan kuantitatif dengan tahapan meliputi pengumpulan dataset, data cleaning, transformasi variabel target menjadi klasifikasi biner (Dropout dan Tidak Dropout), pembagian data, pembangunan model Random Forest, serta evaluasi model menggunakan metode 10-Fold Cross Validation. Dataset yang digunakan terdiri atas 4.424 data mahasiswa dengan 35 atribut, yang meliputi 34 atribut prediktor dan 1 atribut target. Kinerja model dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC). Hasil penelitian menunjukkan bahwa model Random Forest memperoleh Accuracy sebesar 86,1%, Precision 85,8%, Recall 86,1%, F1-Score 85,8%, AUC 0,902, dan MCC 0,673. Nilai AUC menunjukkan kemampuan klasifikasi yang sangat baik dalam membedakan mahasiswa yang berisiko dropout dan yang tidak, sedangkan nilai MCC menunjukkan kualitas prediksi yang cukup kuat meskipun data memiliki distribusi kelas yang tidak seimbang. Hasil penelitian menunjukkan bahwa algoritma Random Forest memiliki performa yang baik dan berpotensi diterapkan sebagai sistem pendukung keputusan untuk membantu perguruan tinggi melakukan deteksi dini terhadap mahasiswa yang berisiko dropout sehingga intervensi akademik dapat dilakukan secara lebih cepat dan tepat.

Kata Kunci: *Random Forest, machine learning, klasifikasi, dropout mahasiswa, data akademik*

ABSTRACT

The high rate of student dropout remains a significant challenge for higher education institutions because it affects graduate quality, educational effectiveness, and student retention. Therefore, an effective approach is required to identify students at risk of dropping out at an early stage based on available academic data. This study aims to classify

student dropout risk using the Random Forest algorithm and the Predict Students' Dropout and Academic Success dataset obtained from the UCI Machine Learning Repository. A quantitative approach was employed through several stages, including dataset collection, data cleaning, binary target transformation (Dropout and Non-Dropout), data partitioning, Random Forest model development, and model evaluation using 10-Fold Cross Validation. The dataset consists of 4,424 student records with 35 attributes, including 34 predictor attributes and one target attribute. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC). The experimental results achieved an Accuracy of 86.1%, Precision of 85.8%, Recall of 86.1%, F1-Score of 85.8%, AUC of 0.902, and MCC of 0.673. The AUC value indicates excellent classification capability in distinguishing students at risk of dropping out from those who are not, while the MCC value demonstrates good predictive quality despite the class imbalance in the dataset. These findings indicate that the Random Forest algorithm provides strong classification performance and has the potential to be implemented as a decision support system for the early identification of students at risk of dropping out, enabling universities to provide timely academic interventions.

Keywords: *Random Forest, machine learning, student dropout, classification, academic data.*

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pemanfaatan kecerdasan buatan (Artificial Intelligence atau AI) dalam berbagai bidang, termasuk pendidikan. Salah satu cabang AI yang berkembang pesat adalah machine learning, yaitu metode yang memungkinkan komputer mempelajari pola dari data sehingga mampu menghasilkan prediksi maupun klasifikasi secara otomatis tanpa harus diprogram secara eksplisit [1]. Dalam bidang pendidikan, machine learning dimanfaatkan untuk membantu institusi pendidikan menganalisis data akademik, memprediksi capaian belajar mahasiswa, mengidentifikasi mahasiswa yang berisiko mengalami kegagalan studi, serta mendukung pengambilan keputusan berbasis data. Pemanfaatan teknologi tersebut menjadi semakin penting seiring meningkatnya jumlah data akademik yang tersimpan dalam sistem informasi perguruan tinggi.

Salah satu permasalahan yang masih menjadi tantangan dalam penyelenggaraan pendidikan tinggi adalah tingginya angka dropout mahasiswa. Mahasiswa yang mengalami dropout tidak dapat menyelesaikan studinya sehingga berdampak pada kualitas lulusan, efektivitas penyelenggaraan pendidikan, akreditasi program studi, serta reputasi institusi. Berdasarkan Buku Statistik Pendidikan Tinggi yang diterbitkan oleh Pusat Data dan Teknologi Informasi Kementerian Pendidikan Tinggi, data mahasiswa putus kuliah terus dipantau melalui Pangkalan Data Pendidikan Tinggi (PDDikti) sebagai salah satu indikator penting dalam evaluasi penyelenggaraan pendidikan tinggi (Pusat Data dan Teknologi Informasi, 2025). Selain itu, penelitian Handini dkk. Sebagaimana dikutip dalam [2] melaporkan bahwa sekitar 602.208 mahasiswa atau sekitar 7% dari total mahasiswa terdaftar tidak berhasil menyelesaikan studinya dan sebagian besar kasus dropout terjadi pada perguruan tinggi swasta. Kondisi tersebut

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

menunjukkan bahwa deteksi dini terhadap mahasiswa yang berisiko dropout menjadi kebutuhan penting agar perguruan tinggi dapat memberikan intervensi akademik secara lebih cepat dan tepat.

Risiko dropout dipengaruhi oleh berbagai faktor, baik faktor akademik maupun nonakademik. Faktor akademik meliputi Indeks Prestasi Kumulatif (IPK), jumlah mata kuliah yang lulus, tingkat kehadiran, nilai akademik, serta riwayat pengambilan mata kuliah. Sementara itu, faktor nonakademik meliputi kondisi ekonomi, status pembayaran uang kuliah, motivasi belajar, dukungan keluarga, dan lingkungan sosial. Seluruh informasi tersebut umumnya telah tersedia dalam sistem informasi akademik perguruan tinggi. Namun, proses identifikasi mahasiswa yang berisiko dropout masih banyak dilakukan secara manual sehingga membutuhkan waktu yang relatif lama dan berpotensi menghasilkan keputusan yang kurang optimal. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengolah data akademik secara cepat, akurat, dan objektif sehingga mahasiswa yang berisiko dropout dapat diidentifikasi sejak dini [3].

Perkembangan machine learning memberikan peluang untuk membangun sistem prediksi risiko dropout yang lebih efektif. Penelitian yang dilakukan oleh Tumbilung dan Efendi (2025) [3] menerapkan algoritma Decision Tree, Naïve Bayes, dan K-Nearest Neighbor (KNN) untuk memprediksi risiko dropout mahasiswa berdasarkan data akademik. Hasil penelitian menunjukkan bahwa algoritma Decision Tree memberikan performa terbaik dibandingkan algoritma lainnya. Penelitian Sitinjak (2026) [4] juga menunjukkan bahwa penerapan algoritma machine learning mampu menghasilkan klasifikasi status mahasiswa dengan performa yang baik berdasarkan evaluasi menggunakan confusion matrix. Selanjutnya, Kapila Devi dan Ratnoo (2022) menerapkan algoritma Random Forest untuk memprediksi dropout siswa dengan mempertimbangkan faktor akademik dan sosial ekonomi. Penelitian tersebut memperoleh tingkat akurasi sebesar 86% dan menunjukkan bahwa Random Forest mampu mengidentifikasi faktor-faktor penting yang memengaruhi risiko dropout. Penelitian Putra dkk. (2025) [5] kemudian membandingkan algoritma Random Forest dan XGBoost menggunakan dataset yang terdiri atas 4.424 data mahasiswa dengan 34 atribut. Hasil penelitian menunjukkan bahwa Random Forest menghasilkan performa yang lebih baik dibandingkan XGBoost dalam memprediksi mahasiswa yang berisiko dropout.

Meskipun berbagai penelitian mengenai prediksi risiko dropout telah banyak dilakukan, masih terdapat beberapa kesenjangan penelitian (research gap). Sebagian besar penelitian lebih berfokus pada perbandingan beberapa algoritma klasifikasi, sedangkan penerapan Random Forest sebagai model utama masih relatif terbatas. Selain itu, beberapa penelitian menggunakan dataset yang berasal dari satu institusi sehingga hasil model belum tentu dapat digeneralisasikan pada karakteristik data yang lebih beragam. Di samping itu, evaluasi model pada beberapa penelitian masih terbatas pada metrik Accuracy atau Confusion Matrix, sehingga performa model belum dianalisis secara menyeluruh menggunakan Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC).

Berdasarkan kesenjangan tersebut, penelitian ini menggunakan dataset Predict Students' Dropout and Academic Success yang diperoleh dari UCI Machine Learning Repository [6]. Dataset tersebut merupakan dataset publik yang memiliki karakteristik lebih beragam

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

sehingga dapat menghasilkan model klasifikasi yang lebih representatif. Penelitian ini menerapkan algoritma Random Forest sebagai model utama karena memiliki kemampuan menghasilkan akurasi yang tinggi, mengurangi risiko overfitting, serta membangun model yang lebih stabil melalui pendekatan ensemble learning [7]. Selain itu, evaluasi model dilakukan menggunakan metode 10-Fold Cross Validation dengan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC), sehingga memberikan evaluasi performa model yang lebih komprehensif dibandingkan penelitian sebelumnya.

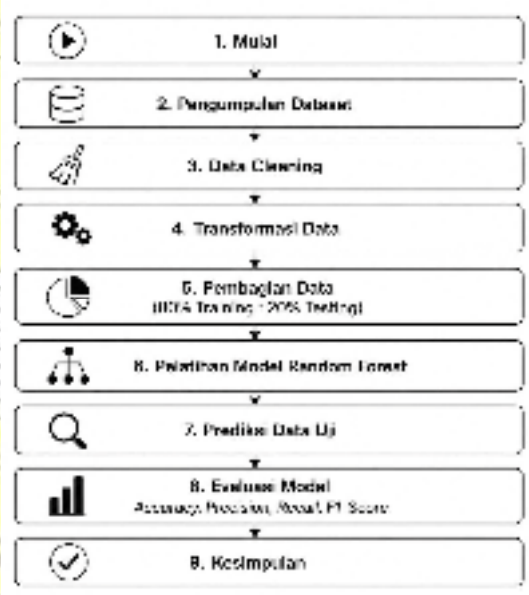
Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan algoritma Random Forest untuk mengklasifikasikan risiko dropout mahasiswa berdasarkan data akademik serta mengevaluasi kinerja model menggunakan berbagai metrik evaluasi. Hasil penelitian diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan yang mampu membantu perguruan tinggi mendeteksi mahasiswa yang berisiko dropout sejak dini sehingga tindakan pendampingan akademik dapat dilakukan secara lebih cepat, tepat, dan berbasis data.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian dilakukan melalui beberapa tahapan, yaitu pengumpulan dataset, data cleaning, transformasi data, pembagian data menjadi data latih dan data uji, pembangunan model Random Forest, pengujian model, serta evaluasi hasil klasifikasi. Pada tahap preprocessing, dilakukan pemeriksaan missing value, pemilihan atribut yang digunakan, serta transformasi variabel target menjadi klasifikasi biner. Selanjutnya, dataset dibagi menjadi 80% data latih dan 20% data uji untuk proses pembentukan model. Model klasifikasi dibangun menggunakan algoritma Random Forest. Algoritma ini membentuk sejumlah decision tree melalui teknik bootstrap sampling dan pemilihan atribut secara acak pada setiap percabangan. Prediksi akhir diperoleh menggunakan mekanisme majority voting sehingga menghasilkan model yang lebih stabil dan mampu mengurangi risiko overfitting [7]. Alur penelitian dapat dilihat pada Gambar 1.

Gambar 1. Diagram Alur Penelitian



Berdasarkan Gambar 1, penelitian diawali dengan proses pengumpulan dataset Predict Students' Dropout and Academic Success dari UCI Machine Learning Repository. Selanjutnya dilakukan tahap data cleaning dan transformasi data untuk memastikan kualitas data sebelum diproses oleh algoritma Random Forest. Dataset kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Model Random Forest dibangun menggunakan data latih, kemudian digunakan untuk memprediksi data uji. Hasil prediksi selanjutnya dievaluasi menggunakan metrik Accuracy, Precision, Recall, F1-Score, AUC, dan Matthews Correlation Coefficient (MCC)

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

untuk mengetahui kinerja model dalam mengklasifikasikan risiko dropout mahasiswa.

2.2 Algoritma Random Forest

Random Forest merupakan salah satu algoritma supervised machine learning berbasis ensemble learning yang dikembangkan oleh Breiman (2001) [7]. Algoritma ini membangun sejumlah decision tree dari sampel data yang dipilih menggunakan teknik bootstrap sampling. Setiap pohon menghasilkan prediksi terhadap data uji, kemudian hasil akhir [8] ; [7] ditentukan melalui mekanisme majority voting, yaitu kelas yang memperoleh suara terbanyak dari seluruh pohon keputusan ditetapkan sebagai hasil klasifikasi. Konsep tersebut menjadikan Random Forest mampu meningkatkan stabilitas model dibandingkan penggunaan satu decision tree saja. Proses kerja Random Forest terdiri atas beberapa tahapan, yaitu pembentukan sampel data secara acak (bootstrap sampling), pembangunan sejumlah decision tree dengan pemilihan atribut secara acak pada setiap percabangan (random feature selection), kemudian penggabungan hasil prediksi seluruh pohon menggunakan mekanisme majority voting. Pendekatan ensemble learning tersebut mampu meningkatkan kemampuan generalisasi model sehingga menghasilkan prediksi yang lebih akurat dan lebih stabil dibandingkan metode klasifikasi tunggal. Random Forest memiliki beberapa keunggulan, di antaranya mampu menangani data berukuran besar, bekerja dengan baik pada data berdimensi tinggi, relatif tahan terhadap noise dan outlier, serta mampu mengurangi risiko overfitting [9]

Pada penelitian ini, algoritma Random Forest diterapkan untuk mengklasifikasikan mahasiswa ke dalam dua kelas, yaitu Dropout dan Tidak Dropout, berdasarkan data akademik mahasiswa. Model dibangun menggunakan perangkat lunak Orange Data Mining dan dievaluasi menggunakan metode 10-Fold Cross Validation. Kinerja model diukur menggunakan metrik Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC) sehingga kemampuan model dalam mengklasifikasikan risiko dropout dapat dievaluasi secara komprehensif. Penggunaan Random Forest pada penelitian ini dipilih karena berbagai penelitian menunjukkan bahwa algoritma tersebut merupakan salah satu metode supervised learning yang efektif untuk permasalahan prediksi dan klasifikasi pada data pendidikan [9].

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Penelitian ini menggunakan dataset Predict Students' Dropout and Academic Success yang diperoleh dari UCI Machine Learning Repository [6]. Dataset tersebut terdiri atas 4.424 data mahasiswa dengan 35 atribut, yaitu 34 atribut prediktor dan 1 atribut target. Dataset ini memuat informasi akademik, demografis, dan sosial ekonomi mahasiswa sehingga banyak dimanfaatkan dalam penelitian Educational Data Mining untuk membangun model prediksi risiko dropout mahasiswa [6] ; [10].

Pada penelitian ini, variabel target ditransformasikan menjadi klasifikasi biner dengan menggabungkan kelas Graduate dan Enrolled menjadi Tidak Dropout, sedangkan kelas Dropout tetap dipertahankan. Transformasi tersebut dilakukan agar proses klasifikasi sesuai dengan tujuan penelitian, yaitu mengidentifikasi mahasiswa yang berisiko mengalami dropout. Tahapan transformasi target juga bertujuan menyederhanakan proses klasifikasi sehingga model Random Forest dapat

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

memfokuskan prediksi pada dua kelas utama, yaitu mahasiswa yang berisiko dropout dan mahasiswa yang tidak berisiko dropout [10]. Distribusi data setelah proses transformasi target dapat dilihat pada Tabel 1.

Tabel 1. Distribusi Dataset

Kelas	Jumlah Data	Persentase
Dropout	1.421	32,12%
Tidak Dropout	3.003	67,88%
Total	4.424	100%

Berdasarkan Tabel 1 terlihat bahwa jumlah mahasiswa pada kategori Tidak Dropout lebih banyak dibandingkan kategori Dropout. Perbedaan jumlah tersebut menunjukkan adanya ketidakseimbangan kelas (class imbalance). Meskipun demikian, Random Forest dikenal mampu memberikan performa yang baik pada data dengan distribusi kelas yang tidak seimbang karena menggunakan pendekatan ensemble learning melalui pembentukan banyak decision tree dan mekanisme majority voting [7] ; [1].

3.2 Tahap Preprocessing Data

Sebelum dilakukan proses klasifikasi, dataset terlebih dahulu melalui tahap preprocessing. Tahapan ini bertujuan agar data yang digunakan sesuai dengan kebutuhan penelitian dan dapat diproses oleh algoritma Random Forest secara optimal. Tahapan preprocessing yang dilakukan meliputi:

1. Mengimpor dataset ke dalam perangkat lunak Orange Data Mining.
2. Memeriksa atribut dan memastikan tidak terdapat data yang hilang (missing value).
3. Mengubah variabel target menjadi klasifikasi biner, yaitu:
   - Dropout
   - Tidak Dropout
4. Menentukan seluruh atribut selain Target sebagai Features.
5. Menentukan atribut Target sebagai variabel kelas (class variable). Setelah preprocessing selesai, dataset siap digunakan untuk proses pelatihan dan pengujian model klasifikasi.

3.3 Implementasi Algoritma Random Forest

Setelah tahap preprocessing selesai, dataset diproses menggunakan algoritma Random Forest melalui perangkat lunak Orange Data Mining. Seluruh atribut prediktor dijadikan sebagai features, sedangkan atribut Target digunakan sebagai class variable. Proses pelatihan model diawali dengan pembentukan sejumlah sampel data menggunakan teknik bootstrap sampling, yaitu pengambilan data latih secara acak dengan pengembalian (sampling with replacement) sehingga setiap decision tree memperoleh kombinasi data yang berbeda [7]. Selanjutnya, setiap decision tree dibangun menggunakan sebagian atribut yang dipilih secara acak (random feature selection) pada setiap percabangan. Teknik ini bertujuan meningkatkan keragaman antar pohon keputusan sehingga mampu mengurangi risiko overfitting dan meningkatkan kemampuan generalisasi model [1] ; [10]. Setelah seluruh decision tree terbentuk, masing-masing pohon melakukan prediksi terhadap data uji. Hasil prediksi dari seluruh pohon kemudian digabungkan menggunakan mekanisme majority voting. Kelas yang memperoleh jumlah suara terbanyak ditetapkan sebagai hasil klasifikasi akhir. Pendekatan ensemble learning tersebut menghasilkan model yang lebih stabil dibandingkan penggunaan satu decision tree karena keputusan akhir merupakan gabungan dari banyak model [7] ; [11].

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Evaluasi model dilakukan menggunakan metode 10-Fold Cross Validation dengan teknik Stratified Sampling. Dataset dibagi menjadi sepuluh bagian dengan proporsi kelas yang tetap terjaga pada setiap lipatan (fold). Pada setiap iterasi, sembilan bagian digunakan sebagai data latih dan satu bagian sebagai data uji. Proses ini diulang sebanyak sepuluh kali hingga seluruh data memperoleh kesempatan menjadi data uji. Selanjutnya, nilai evaluasi dihitung berdasarkan rata-rata hasil seluruh iterasi sehingga menghasilkan pengukuran performa model yang lebih objektif dan mampu meminimalkan bias akibat pembagian data secara acak [1] ; [5].

3.4 Hasil Evaluasi Model

Hasil evaluasi model Random Forest diperoleh menggunakan widget Test & Score pada Orange Data Mining.

Tabel 2 Hasil Evaluasi Random Forest

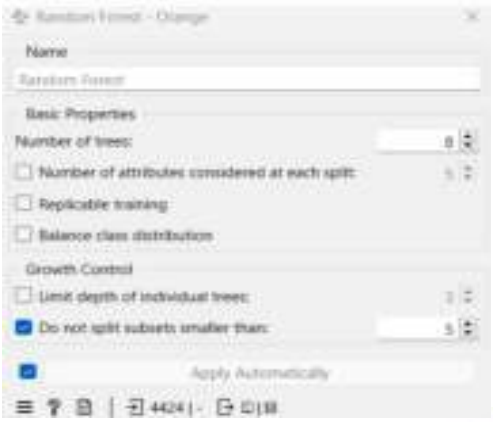
Metrik	Nilai
Accuracy	86,1%
Precision	85,8%
Recall	86,1%
F1-Score	85,8%
AUC	0,902
MCC	0,673

Berdasarkan hasil pengujian menggunakan metode 10-Fold Cross Validation, algoritma Random Forest menghasilkan nilai Accuracy sebesar 86,1%, Precision sebesar 85,8%, Recall sebesar 86,1%, F1-Score sebesar 85,8%, Area Under Curve (AUC) sebesar 0,902, dan Matthews Correlation Coefficient (MCC) sebesar 0,673. Nilai Accuracy menunjukkan bahwa model mampu mengklasifikasikan sekitar 86 dari setiap 100 data mahasiswa dengan benar [1]. Nilai Precision yang tinggi menunjukkan bahwa sebagian besar mahasiswa yang diprediksi mengalami dropout memang termasuk ke dalam kelas tersebut, sedangkan Recall sebesar 86,1% menunjukkan bahwa model mampu mendeteksi sebagian besar mahasiswa yang benar-benar berisiko dropout [11]. Nilai F1-Score sebesar 85,8% menunjukkan keseimbangan antara Precision dan Recall, sehingga model tidak hanya akurat tetapi juga konsisten dalam melakukan klasifikasi [12]. Selain itu, nilai AUC sebesar 0,902 termasuk dalam kategori excellent classification karena memiliki kemampuan yang sangat baik dalam membedakan kedua kelas, sedangkan nilai MCC sebesar 0,673 menunjukkan hubungan yang cukup kuat antara hasil prediksi dengan kondisi sebenarnya, terutama pada data yang memiliki distribusi kelas tidak seimbang [13].

3.5 Analisis Confusion Matrix

Confusion Matrix digunakan untuk mengetahui kemampuan model dalam mengklasifikasikan setiap kelas secara lebih rinci. Matriks ini terdiri atas empat komponen, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP menunjukkan jumlah mahasiswa Dropout yang berhasil diprediksi dengan benar, TN menunjukkan mahasiswa Tidak Dropout yang diprediksi dengan benar, FP menunjukkan mahasiswa Tidak Dropout yang diprediksi sebagai Dropout, sedangkan FN menunjukkan mahasiswa Dropout yang diprediksi sebagai Tidak Dropout. Implementasi algoritma Random Forest dilakukan menggunakan perangkat lunak Orange Data Mining. Pada tahap ini ditentukan parameter model yang akan digunakan, antara lain jumlah decision tree sebanyak delapan pohon (Number of Trees = 8) dan jumlah minimum data pada setiap simpul (Do not split subsets smaller than = 5). Pengaturan parameter tersebut digunakan sebagai dasar dalam proses pembentukan model klasifikasi. Untuk mengetahui kemampuan model secara lebih rinci digunakan Confusion Matrix.

Gambar 2. Algoritma Random Forest



Berdasarkan Gambar 2, algoritma Random Forest dikonfigurasi untuk membentuk delapan decision tree yang dibangun dari sampel data hasil bootstrap sampling. Setiap pohon melakukan proses klasifikasi secara independen, kemudian hasil prediksi digabungkan menggunakan mekanisme majority voting untuk menghasilkan keputusan akhir. Pendekatan ini menghasilkan model yang lebih stabil dan mampu mengurangi risiko overfitting [7] ; [1].

Berdasarkan confusion matrix diperoleh bahwa:

- Sebanyak 82,1% mahasiswa yang benar-benar mengalami Dropout berhasil diprediksi dengan benar oleh model.
- Sebanyak 87,6% mahasiswa yang termasuk kategori Tidak Dropout berhasil diprediksi dengan benar.
- Sebagian kecil data masih mengalami kesalahan klasifikasi, yaitu mahasiswa Dropout yang diprediksi sebagai Tidak Dropout dan sebaliknya.

Hasil tersebut menunjukkan bahwa algoritma Random Forest memiliki kemampuan yang cukup baik dalam membedakan mahasiswa yang memiliki risiko dropout dengan mahasiswa yang tidak berisiko.

3.6 Pembahasan

Berdasarkan hasil penelitian, algoritma Random Forest mampu memberikan performa klasifikasi yang baik dalam mengidentifikasi mahasiswa yang berisiko mengalami dropout. Hasil evaluasi menunjukkan nilai Accuracy sebesar 86,1%, Precision sebesar 85,8%, Recall sebesar 86,1%, F1-Score sebesar 85,8%, AUC sebesar 0,902, dan MCC sebesar 0,673. Nilai tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan mahasiswa yang berpotensi mengalami dropout dan mahasiswa yang tidak berisiko. Tingginya nilai akurasi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar, sedangkan nilai Precision dan Recall yang relatif seimbang mengindikasikan bahwa model mampu mengidentifikasi mahasiswa berisiko tanpa menghasilkan terlalu banyak kesalahan prediksi. Kondisi ini menunjukkan bahwa Random Forest efektif digunakan pada permasalahan klasifikasi yang melibatkan banyak atribut dan hubungan antarvariabel yang kompleks [1] ; [14].

Performa yang baik tersebut tidak terlepas dari mekanisme kerja Random Forest yang membangun sejumlah decision tree menggunakan teknik bootstrap sampling dan random feature selection. Setiap pohon keputusan dibangun dari kombinasi data latih yang berbeda sehingga menghasilkan model yang beragam. Selanjutnya, seluruh hasil prediksi digabungkan menggunakan mekanisme majority voting untuk menentukan kelas akhir. Pendekatan ensemble learning tersebut mampu mengurangi variansi model dan meminimalkan risiko overfitting, sehingga menghasilkan model yang lebih stabil dibandingkan penggunaan satu decision tree [1] ; [15].

Pada penelitian ini proses evaluasi dilakukan menggunakan metode 10-Fold Cross Validation dengan teknik Stratified Sampling. Metode tersebut membagi dataset menjadi sepuluh bagian dengan proporsi kelas yang tetap terjaga pada setiap lipatan. Setiap bagian secara bergantian digunakan sebagai data uji,

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

sedangkan sembilan bagian lainnya digunakan sebagai data latih. Pendekatan ini menghasilkan proses evaluasi yang lebih objektif karena seluruh data memperoleh kesempatan menjadi data uji dan hasil evaluasi merupakan rata-rata dari seluruh proses pengujian [13]. Oleh karena itu, nilai performa yang diperoleh lebih mampu merepresentasikan kemampuan model ketika diterapkan pada data baru.

Nilai AUC sebesar 0,902 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan mahasiswa yang berisiko dropout dan mahasiswa yang tidak berisiko. Menurut klasifikasi performa model, nilai AUC di atas 0,90 termasuk dalam kategori excellent classification, sehingga menunjukkan bahwa Random Forest mampu menghasilkan probabilitas prediksi yang sangat baik [16]. Selain itu, nilai Matthews Correlation Coefficient (MCC) sebesar 0,673 menunjukkan adanya korelasi yang cukup kuat antara hasil prediksi dan kondisi sebenarnya. Metrik MCC dinilai lebih representatif dibandingkan Accuracy ketika dataset memiliki distribusi kelas yang tidak seimbang karena mempertimbangkan seluruh komponen pada confusion matrix [17].

Hasil confusion matrix menunjukkan bahwa model berhasil mengklasifikasikan 81,7% mahasiswa yang benar-benar mengalami dropout dan 87,8% mahasiswa yang termasuk kategori tidak dropout dengan benar. Hasil tersebut menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik pada kedua kelas, meskipun performa pada kelas tidak dropout sedikit lebih tinggi. Perbedaan tersebut diduga dipengaruhi oleh distribusi data yang tidak seimbang, di mana jumlah mahasiswa pada kategori tidak dropout lebih banyak dibandingkan kategori dropout. Walaupun demikian, Random Forest tetap mampu menghasilkan performa yang baik karena mekanisme ensemble learning membuat model lebih tahan terhadap pengaruh ketidakseimbangan data dibandingkan beberapa algoritma klasifikasi lainnya [15].

Hasil penelitian ini sejalan dengan penelitian Andrianof et al. (2025) [10] yang melaporkan bahwa Random Forest mampu menghasilkan tingkat akurasi yang tinggi dalam memprediksi keberhasilan studi mahasiswa. Penelitian tersebut menunjukkan bahwa Random Forest efektif memanfaatkan data akademik mahasiswa untuk menghasilkan model prediksi yang stabil sehingga layak diterapkan sebagai sistem pendukung keputusan di lingkungan perguruan tinggi. Kesamaan hasil tersebut menunjukkan bahwa Random Forest memiliki kemampuan yang konsisten dalam menyelesaikan permasalahan klasifikasi pada bidang pendidikan.

Temuan penelitian ini juga mendukung penelitian Sulehu et al. (2025) [18] yang menyatakan bahwa Random Forest mampu meningkatkan kualitas prediksi terhadap status akademik mahasiswa dibandingkan beberapa algoritma klasifikasi lainnya. Menurut penelitian tersebut, kombinasi banyak decision tree mampu menghasilkan model dengan tingkat generalisasi yang lebih baik sehingga dapat mengurangi kesalahan prediksi pada data baru. Hasil tersebut sesuai dengan penelitian ini yang memperoleh nilai Accuracy sebesar 86,1% dan AUC sebesar 0,902, sehingga menunjukkan kemampuan model dalam membedakan mahasiswa yang berisiko dropout dan yang tidak.

Selain itu, hasil penelitian ini juga konsisten dengan penelitian Gustian dan Mahardika (2025) [12] yang membandingkan algoritma Decision Tree dan Random Forest. Penelitian tersebut menunjukkan bahwa Random Forest menghasilkan performa klasifikasi yang lebih baik karena mampu mengurangi risiko overfitting melalui mekanisme bootstrap sampling dan majority voting. Pada penelitian ini, keunggulan tersebut

juga terlihat dari tingginya nilai Accuracy, F1-Score, serta MCC yang menunjukkan bahwa model memiliki kemampuan prediksi yang stabil meskipun dataset memiliki karakteristik yang cukup kompleks.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa algoritma Random Forest layak digunakan sebagai metode klasifikasi risiko dropout mahasiswa. Kemampuan model dalam menghasilkan nilai Accuracy, Precision, Recall, F1-Score, AUC, dan MCC yang tinggi menunjukkan bahwa Random Forest mampu mendeteksi mahasiswa yang berpotensi mengalami dropout secara efektif. Informasi tersebut dapat dimanfaatkan oleh perguruan tinggi sebagai dasar dalam memberikan intervensi dini, seperti pembinaan akademik, layanan konseling, maupun program pendampingan belajar sehingga risiko dropout dapat diminimalkan. Dengan demikian, penerapan Random Forest berpotensi mendukung pengambilan keputusan berbasis data dalam meningkatkan keberhasilan studi mahasiswa.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Random Forest mampu diterapkan dengan baik untuk mengklasifikasikan risiko dropout mahasiswa menggunakan dataset Predict Students' Dropout and Academic Success. Penerapan algoritma ini dilakukan melalui tahapan preprocessing data, pembentukan model, dan evaluasi menggunakan metode 10-Fold Cross Validation. Hasil pengujian menunjukkan bahwa model memiliki performa yang baik dengan nilai Accuracy sebesar 86,1%, Precision sebesar 85,8%, Recall sebesar 86,1%, F1-Score sebesar 85,8%, Area Under Curve (AUC) sebesar 0,902, serta Matthews Correlation Coefficient (MCC) sebesar 0,673. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan mahasiswa yang berisiko dropout maupun tidak dropout dengan tingkat ketepatan yang tinggi sehingga dapat digunakan sebagai model klasifikasi yang andal. Hasil penelitian juga menunjukkan bahwa algoritma Random Forest memiliki kemampuan yang baik dalam menangani data akademik mahasiswa karena menggunakan pendekatan ensemble learning yang menggabungkan banyak decision tree melalui mekanisme majority voting. Pendekatan tersebut menghasilkan model yang lebih stabil dan mampu mengurangi risiko overfitting, sehingga performa klasifikasi menjadi lebih baik. Selain itu, hasil evaluasi melalui confusion matrix menunjukkan bahwa model mampu mengidentifikasi sebagian besar mahasiswa pada kedua kelas dengan tingkat prediksi yang baik, meskipun masih terdapat sejumlah kecil kesalahan klasifikasi yang dipengaruhi oleh distribusi data yang tidak seimbang. Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma Random Forest berpotensi dimanfaatkan sebagai sistem pendukung keputusan dalam lingkungan perguruan tinggi untuk mendeteksi mahasiswa yang memiliki risiko dropout sejak dini. Informasi yang dihasilkan dari model klasifikasi dapat membantu pihak perguruan tinggi dalam merancang strategi pendampingan akademik, layanan konseling, maupun kebijakan lain yang bertujuan meningkatkan retensi mahasiswa dan menurunkan angka dropout. Dengan demikian, tujuan penelitian untuk menerapkan algoritma Random Forest dalam mengklasifikasikan risiko dropout mahasiswa serta mengevaluasi kinerja model telah tercapai dengan hasil yang memuaskan .

5. UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, masukan, dan bimbingan selama proses

penyusunan penelitian ini. Penulis juga menyampaikan terima kasih kepada UCI Machine Learning Repository yang telah menyediakan dataset Predict Students' Dropout and Academic Success sehingga penelitian ini dapat dilaksanakan. Selain itu, penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, baik secara langsung maupun tidak langsung, sehingga penelitian ini dapat diselesaikan dengan baik.

REFERENCES

[1] K. Jefri Junifer Pangaribuan1*), Henry Tanjaya2), *Machine learning 분야 소개 및 주요 방법론 학습 기본 machine learning 알고리즘에 대한 이해 및 응용 관련 최신 연구 동향 습득*, vol. 45, no. 13. 2017. [Online]. Available: https://books.google.ca/books?id=EoYBngEACAAJ&dq=mitchell+machine+learning+1997&hl=en&sa=X&ved=0ahUKEwiomdqfj8TkAhWGslkKHRCbAtoQ6AEIKjAA

[2] P. Bimbingan, D. A. N. Konseling, U. Mencegah, S. Attrition, and D. I. Perguruan, "Ariestio Widyaseno, 2023 PROGRAM BIMBINGAN DAN KONSELING UNTUK MENCEGAH STUDENT ATTRITION DI PERGURUAN TINGGI Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu," pp. 1–9, 2023.

[3] C. F. Tumbilung and Rissal Efendi, "Prediksi Risiko Drop Out Mahasiswa Menggunakan Model Machine Learning Berbasis Data Akademik (Studi Kasus : Universitas XYZ)," *JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 16, no. 2, pp. 176–183, 2025, [Online]. Available: https://ejurnal.provisi.ac.id/index.php/JTIKP/article/view/1257/878

[4] D. A. Sitinjak, "Analisis Prediksi Drop out Mahasiswa dengan Random Forest Studi Kasus Fakultas Ilmu Komputer Universitas Esa Unggul," vol. 1, no. 1, pp. 8740–8748, 2022.

[5] L. G. R. Putra, D. D. Prasetya, and M. Mayadi, "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.

[6] F. A. Putra, S. Mirajdandi, B. Okmarizal, and S. Mulyanda, "Prediksi Dropout Mahasiswa : Early-Warning Berbasis Enrollment dengan Machine Learning," vol. 15, no. 3, pp. 465–473, 2025.

[7] N. K. Dewi, S. Y. Mulyadi, and U. D. Syafitri, "Penerapan Metode Random Forest Dalam Driver Analysis," *Forum Stat. Dan Komputasi*, vol. 16, no. 1, pp. 35–43, 2012, [Online]. Available: http://journal.ipb.ac.id/index.php/statistika/article/view/5443

[8] G. M. Angelo and E. Mailoa, "Deteksi Phishing Website Menggunakan Algoritma Random Forest dengan Hyperparameter Tuning GridSearchCV Abstrak," vol. 7, no. 2, pp. 322–331, 2026.

[9] M. W. Nugroho, "Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 9, no. 4, pp. 1562–1571, 2025, doi: 10.35870/jtik.v9i4.4236.

[10] P. S. Saputra, "Analisis Prediktif Dropout Mahasiswa Berdasarkan Kinerja Akademik Semester Awal Menggunakan Machine Learning," *J. Ris. dan Apl. Mhs. Inform.*, vol. 07, no. 01, pp. 164–171, 2026.

[11] A. Amin, S. Basri, M. F. Hassan, and M. Rehman, "Systematic Literature Review on Creativity in Software Engineering," pp. 370–396, 2016.

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

[13] I. I. J. Rifka Alkhilyatul Ma'rifat, I Made Suraharta, *METODOLOGI PENELITIAN (Panduan Praktis Untuk Penelitian Berkualitas)*, vol. 2. 2024.

[14] S. Lina, M. Sitio, S. Kom, and M. Kom, *MACHINE LEARNING BERBASIS POHON KEPUTUSAN : Penulis : Sartika Lina Mulani Sitio , S . Kom ., M . Kom*. 2026.

[15] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[16] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[17] I. N. Mulyanan, M. Yusril, H. Setyawan, and W. I. Rahayu, "Penerapan Metode Naïve Bayes Untuk Merekomendasikan," vol. 7, no. 5, pp. 3453–3460, 2023.

[18] M. Sulehu, W. Wisda, F. Wanita, and M. Markani, "Optimasi Prediksi Kelulusan Mahasiswa Menggunakan Random Forest untuk Meningkatkan Tingkat Retensi," *J. Minfo Polgan*, vol. 13, no. 2, pp. 2364–2374, 2025, doi: 10.33395/jmp.v13i2.14472.

P-ISSN : 2580-4316
https://doi.org/10.37817/ikraith-informatika.v10i2 E-ISSN : 2654-8054