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

Prediksi Tunggakan Pembayaran Mahasiswa Universitas Muhammadiyah Ahmad Dahlan Cirebon menggunakan Algoritma *Random Forest*

Prediction of Unpaid Student Fees at Muhammadiyah Ahmad Dahlan University Cirebon using the Random Forest Algorithm

[1]Suherman*, [2]Domy Kristomo
[1,2]Program Studi Magister Teknologi Informasi, Fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia
[1,2]Jl. Raya Janti Karang Jambe no 143, Yogyakarta, D.I.Y
*e-mail: student.suherman24@mti.utdi.ac.id

(***received*:** 6 August 2025, ***revised*:** 13 October 2025, ***accepted*:** 14 October 2025)

Abstrak

Penelitian ini bertujuan mengembangkan model prediksi tunggakan pembayaran mahasiswa di Universitas Muhammadiyah Ahmad Dahlan Cirebon menggunakan algoritma Random Forest. Dataset yang digunakan berasal dari Sistem Informasi Akademik dengan total 490 data mahasiswa dari empat angkatan (2018–2021), yang dibagi menjadi 80% data latih dan 20% data uji. Proses pengolahan mencakup data cleaning, transformasi, dan seleksi fitur dengan Recursive Feature Elimination. Model dioptimalkan melalui GridSearchCV untuk memperoleh konfigurasi terbaik. Hasil evaluasi menunjukkan kinerja yang tinggi dengan AUC 0,980, akurasi 88,8%, precision 90,4%, recall 88,8%, dan F1-score 0,875. Analisis feature importance menempatkan variabel *jumlah_tunggakan* sebagai faktor dominan. Rekomendasi strategis yang dapat diterapkan universitas meliputi: (1) penerapan sistem peringatan dini berbasis data bagi mahasiswa berisiko, (2) program keringanan atau cicilan pembayaran khusus bagi mahasiswa dengan tunggakan tinggi, serta (3) monitoring berkala melalui dashboard keuangan untuk pengambilan keputusan cepat. Dengan demikian, penelitian ini tidak hanya menghasilkan model prediksi, tetapi juga solusi praktis dalam pengelolaan keuangan kampus.

Kata kunci: *GridSearchCV*, manajemen keuangan perguruan tinggi, *machine learning*, evaluasi model (*Auc*, *F1*, akurasi), *random forest*, *recursive feature elimination*, prediksi tunggakan mahasiswa

Abstract

This study aims to develop a predictive model for student fee payment arrears at Universitas Muhammadiyah Ahmad Dahlan Cirebon using the Random Forest algorithm. The dataset was obtained from the Academic Information System and consisted of 490 student records from four cohorts (2018–2021), which were divided into 80% training data and 20% testing data. The data processing stages included data cleaning, transformation, and feature selection using Recursive Feature Elimination (RFE). The model was optimized using GridSearchCV to obtain the best configuration. The evaluation results indicate strong performance, with an AUC of 0.980, accuracy of 88.8%, precision of 90.4%, recall of 88.8%, and an F1-score of 0.875. Feature importance analysis identified the amount of arrears variable as the most dominant factor influencing prediction outcomes. Strategic recommendations for university implementation include: (1) deploying a data-driven early warning system to identify at-risk students, (2) offering payment relief or installment programs for students with high arrears, and (3) conducting regular financial monitoring through a dashboard to support timely decision-making. Therefore, this study not only produces an effective predictive model but also provides practical solutions for improving university financial management.

Keywords: *GridSearchCV, higher education financial management, machine learning, model evaluation (AUC, F1, accuracy), random forest, recursive feature elimination, student loan default prediction*

1 Pendahuluan

Tingginya angka tunggakan pembayaran kuliah merupakan permasalahan yang umum dihadapi perguruan tinggi. [1][2] Studi sebelumnya menunjukkan bahwa metode konvensional untuk memprediksi tunggakan pembayaran seringkali tidak akurat dan tidak tepat waktu, sehingga menyulitkan upaya pencegahan. [1][2] Penelitian tentang prediksi keterlambatan pembayaran telah dilakukan di berbagai sektor, termasuk industri Fintech [1] dan lembaga pendidikan [2]. Namun, aplikasi *machine learning* untuk memprediksi tunggakan pembayaran mahasiswa di perguruan tinggi masih perlu dieksplorasi lebih lanjut. Data historis pembayaran mahasiswa yang kompleks membutuhkan analisis yang lebih canggih untuk mengidentifikasi pola dan faktor-faktor yang berkontribusi terhadap tunggakan. [5] Algoritma *Random Forest*, terbukti efektif dalam prediksi berbagai fenomena, termasuk prediksi kelulusan mahasiswa [3][4] dan lama studi [5]. Kemampuannya dalam menangani data berdimensi tinggi dan menghasilkan akurasi prediksi yang tinggi [4] [5] menjadikannya pilihan yang tepat untuk mengatasi permasalahan tunggakan pembayaran di Universitas Muhammadiyah Ahmad Dahlan (UMMADA) Cirebon. Meskipun beberapa penelitian telah menggunakan *Random Forest* untuk memprediksi kinerja akademik mahasiswa [6], aplikasi algoritma ini untuk memprediksi tunggakan pembayaran masih relatif jarang, karena itu, diharapkan bahwa penelitian ini akan menghasilkan temuan baru dalam bidang ini. Tujuan dari penelitian ini adalah untuk mengembangkan model prediksi yang lebih akurat dan efisien untuk membantu UMMADA Cirebon dalam pengelolaan keuangan dan keberlangsungan operasionalnya.

Penelitian ini diarahkan untuk memberikan kontribusi praktis bagi UMMADA Cirebon dalam rangka optimalisasi pengelolaan keuangan. Model prediksi yang akurat akan memfasilitasi intervensi dini terhadap mahasiswa yang berpotensi menunggak pembayaran, sehingga dapat meminimalisir kerugian finansial dan mempertahankan stabilitas keuangan kampus. Hal ini sejalan dengan tujuan penelitian-penelitian sebelumnya yang menekankan pentingnya prediksi dini untuk meningkatkan efisiensi pengelolaan keuangan [1][2]. Akan tetapi, penelitian ini menawarkan kontribusi yang lebih spesifik dengan berfokus pada prediksi tunggakan pembayaran mahasiswa di lingkungan perguruan tinggi dan memanfaatkan algoritma *Random Forest* yang telah teruji efikasinya dalam berbagai konteks prediksi [5] [6]. Lebih dari itu, penelitian ini juga diharapkan dapat menemukan penyebab tunggakan pembayaran di UMMADA Cirebon, memberikan landasan empiris bagi pengembangan kebijakan dan strategi yang lebih efektif dalam mencegah dan mengelola tunggakan di masa mendatang.

Penelitian ini bertujuan memberikan rekomendasi strategi pengelolaan tunggakan pembayaran mahasiswa di UMMADA Cirebon berdasarkan hasil prediksi model, sehingga penelitian ini tidak hanya menghasilkan model prediksi tetapi juga solusi praktis bagi permasalahan yang dihadapi UMMADA Cirebon. Rekomendasi ini akan mempertimbangkan temuan dari studi sebelumnya tentang variabel yang memengaruhi keterlambatan pembayaran [1].

2 Tinjauan Literatur

Penelitian tentang prediksi tunggakan pembayaran telah dilakukan pada berbagai sektor dengan pendekatan beragam. Pada sektor listrik/ utilitas, menggunakan *Random Forest Regressor* dan BiLSTM untuk memprediksi keterlambatan pembayaran listrik, dengan hasil yang cukup akurat [7]. Di sektor perbankan/keuangan, membandingkan *XGBoost*, *Logistic Regression*, dan *Random Forest* untuk memprediksi gagal bayar pinjaman berbasis data nasabah, dan menunjukkan bahwa *XGBoost* memiliki performa terbaik [8]. Penelitian lain oleh [9] berfokus pada prediksi tanggal pembayaran kartu kredit menggunakan *Linear Regression*, *XGBoost*, *Random Forest*, dan *KNN*, yang menghasilkan akurasi sangat tinggi (hingga 99% pada *XGBoost*). Pada sektor kesehatan sosial, [10] telah membuktikan bahwa *Random Forest* lebih unggul dibanding *AdaBoost* dalam memprediksi peserta BPJS yang berpotensi menunggak, terutama ketika digabungkan dengan *SMOTE* untuk menangani data tidak seimbang.

Di bidang pendidikan, *Random Forest* juga digunakan untuk memprediksi kelulusan mahasiswa oleh [3] dengan hasil akurasi, presisi, dan *recall* yang sangat tinggi. Namun, meskipun memberikan gambaran kuat tentang kemampuan algoritma, penelitian-penelitian tersebut tidak berfokus pada persoalan tunggakan pembayaran mahasiswa. Dari sisi algoritma, *Random Forest* terbukti andal pada berbagai domain karena mampu menangani data berdimensi tinggi dan memberikan interpretabilitas

melalui *feature importance*. *XGBoost* sering menunjukkan performa terbaik dalam konteks keuangan, sementara *Logistic Regression* yang dikombinasikan dengan *feature selection* seperti *Recursive Feature Elimination (RFE)* dan teknik penyeimbangan data seperti *SMOTE* juga mampu menghasilkan kinerja yang kompetitif meskipun kurang menangkap non-linearitas yang kompleks.

Teknik optimasi juga mendapat perhatian dalam literatur. Seleksi fitur menggunakan *RFE* dan *Genetic Algorithm* terbukti meningkatkan akurasi dan efisiensi model, sementara *hyperparameter tuning* melalui *GridSearchCV* sering digunakan untuk memaksimalkan kinerja *Random Forest*, sebagaimana ditunjukkan dalam penelitian [11] yang berhasil mencapai akurasi 91,41% pada data klinis. Namun, meskipun berbagai teknik ini efektif, sebagian besar penelitian berhenti pada evaluasi performa model dan belum menghubungkannya secara langsung dengan rekomendasi kebijakan operasional.

Tabel 1 Perbandingan penggunaan algoritma dari berbagai sektor

Sektor	Algoritma	Dataset	Optimasi	Hasil
Listrik	*RF Regressor, BiLSTM*	Riwayat pembayaran listrik	–	Model berhasil prediksi keterlambatan
Perbankan	*XGBoost, LR, RF*	Data nasabah *(Kaggle)*	–	*XGBoost* terbaik, *RF & LR* baik
Perbankan	*LinReg, XGBoost, RF, KNN*	Data kartu kredit	–	Akurasi hingga 99% *(XGBoost)*
Kesehatan	*RF vs AdaBoost*	Big Data BPJS	*SMOTE*	*RF* unggul (*AUC & F1* lebih tinggi)
Pendidikan	*Random Forest*	Data mahasiswa	–	*Acc/Prec/Recall = 100%*
Pendidikan	*Random Forest*	Data akademik	–	*Acc 88%, Prec 81%, Recall 97%*
Klinis	*RF + RFE*	Data klinis UCI	*GridSearchCV*	*Acc 91,41%*
Optimasi	*LR+RFE+SMOTE*	UCI (3 dataset)	*RFE, SMOTE*	*AUC 93%*

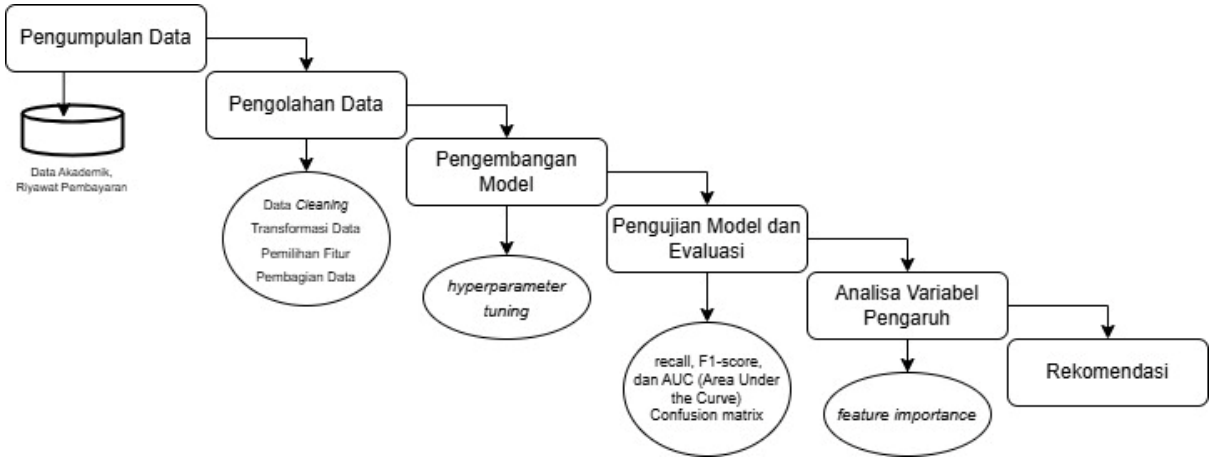
Berdasarkan Tabel 1 di atas, dapat disimpulkan bahwa meskipun berbagai algoritma dan teknik optimasi telah terbukti efektif dalam meningkatkan akurasi prediksi, masih terdapat celah penelitian yang belum tergarap. Pertama, penelitian terdahulu lebih banyak berfokus pada sektor listrik, perbankan, dan kesehatan, sementara konteks tunggakan pembayaran mahasiswa di perguruan tinggi masih jarang dikaji. Kedua, sebagian besar studi hanya mengklasifikasikan dua kelas (menunggak/tidak menunggak), belum menerapkan klasifikasi multi-kelas yang lebih relevan dengan kondisi nyata. Ketiga, meskipun nilai *feature importance* sering dilaporkan, hasil tersebut belum banyak diterjemahkan menjadi strategi operasional yang konkret.

Penelitian ini hadir untuk mengisi celah tersebut dengan tiga kontribusi utama. Pertama, menggunakan data riil mahasiswa (490 data) dari UMMADA Cirebon dengan pembagian latih-uji 80:20. Kedua, membangun model Random Forest multi-kelas (Tunggak, Mungkin, Tepat Waktu) dengan optimasi *hyperparameter* melalui *GridSearchCV* serta evaluasi metrik komprehensif (AUC, akurasi, *precision*, *recall*, *F1-score*). Ketiga, menurunkan hasil analisis *feature importance* ke dalam rekomendasi strategis konkrit berupa sistem peringatan dini berbasis data, skema cicilan adaptif bagi mahasiswa berisiko, dan dashboard keuangan untuk monitoring *real-time*. Dengan demikian, penelitian ini tidak hanya menambah bukti empiris penggunaan machine learning di pendidikan tinggi, tetapi juga menyediakan landasan praktis untuk pengelolaan keuangan kampus yang lebih proaktif dan berbasis data.

3 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *data mining* untuk mengembangkan model prediksi tunggakan pembayaran mahasiswa di UMMADA Cirebon. Data primer yang digunakan berasal dari sistem informasi akademik (SIMAK-UMMADA), yang mencakup data historis pembayaran mahasiswa. Penggunaan data dari sistem akademik memungkinkan akses ke data yang komprehensif dan akurat, mengurangi bias dan meningkatkan reliabilitas hasil penelitian [12].

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



Gambar 1 Alur proses metode penelitian

3.1 Pengumpulan Data

Dari Gambar 1 di atas Data primer diambil dari Sistem Akademik UMMADA Cirebon. Variabel-variabel yang akan dikumpulkan meliputi data demografis mahasiswa, data akademik dan data riwayat pembayaran. Mengingat kampus UMMADA Cirebon merupakan penggabungan dari 3 kampus Sekolah Tinggi untuk itu data yang mempunyai riwayat adalah dari program studi yang lama yaitu program studi D3 Keperawatan dengan data dari 4 angkatan (2018,2019,2020,2021). Data yang telah dikumpulkan akan dibersihkan (*data cleaning*) untuk menangani *missing values* dan *outliers* menggunakan teknik yang sesuai, seperti imputasi (misalnya, imputasi rata-rata, median, atau modus) atau penghapusan data. Keputusan untuk menggunakan teknik imputasi atau penghapusan data akan didasarkan pada analisis karakteristik data dan proporsi *missing values.*

3.2 Pengolahan Data

Setelah proses *data cleaning*, data akan diolah untuk mempersiapkannya untuk pemodelan. Proses pengolahan data meliputi:

a. Transformasi Data
   Variabel-variabel kategorikal akan ditransformasikan ke dalam bentuk numerik menggunakan teknik *label encoding* .
b. Pemilihan Fitur
   Teknik seleksi fitur, menggunakan *recursive feature elimination* dari algoritma *Random Forest*, digunakan untuk memilih variabel-variabel yang paling berpengaruh terhadap prediksi tunggakan pembayaran. Hal ini bertujuan untuk meningkatkan efisiensi dan akurasi model serta mengurangi kompleksitas model. [13][14]
c. Pembagian Data
   Data yang telah diolah akan terdiri dari tiga bagian: data pelatihan (training data), data pengujian (testing data), dan perbandingan dua puluh persen untuk masing-masing. Untuk menjamin representasi yang seimbang dari semua data, pembagiannya dilakukan secara acak.

3.3 Pengembangan Model

Algoritma *Random Forest* akan digunakan untuk mengembangkan model prediksi tunggakan pembayaran mahasiswa. Parameter-parameter algoritma akan dioptimalkan menggunakan teknik *hyperparameter tuning* dengan *grid search* menggunakan data test untuk mengevaluasi performa model pada setiap iterasi *hyperparameter tuning*. Tujuannya adalah untuk menemukan kombinasi parameter yang meminimalkan kesalahan prediksi.[14]

3.4 Pengujian Model dan Evaluasi

Data pengujian yang belum pernah digunakan selama proses pelatihan dan validasi akan digunakan untuk menguji model *Random Forest* yang telah dioptimalkan. Berbagai metrik, seperti akurasi, presisi, *recall*, skor F1, dan *AUC (Area Under the Curve)*, akan digunakan untuk mengukur kemampuan model untuk memprediksi tunggakan pembayaran dengan

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

tepat. Selain itu, untuk memahami jenis kesalahan yang dibuat oleh model, seperti false positive dan *false negative*, analisis *confusion matrix* akan dilakukan. Penggunaan data validasi dan pengujian yang berbeda menjamin generalisasi model dan mengurangi risiko *overfitting*.

3.5 Analisa Variabel Pengaruh

Setelah model dikembangkan dan dievaluasi, analisis akan dilakukan untuk menemukan variabel-variabel yang paling berpengaruh terhadap prediksi tunggakan pembayaran. Berdasarkan fitur penting dari algoritma *Random Forest*, hasil analisis ini akan memberikan wawasan tentang faktor-faktor yang berkontribusi terhadap tunggakan pembayaran mahasiswa UMMADA Cirebon..

3.6 Rekomendasi

Berdasarkan hasil penelitian, rekomendasi akan diberikan kepada UMMADA Cirebon terkait strategi pengelolaan tunggakan pembayaran mahasiswa. Rekomendasi ini akan mencakup langkah-langkah preventif dan kuratif yang dapat diambil untuk meminimalisir angka tunggakan dan menjaga stabilitas keuangan kampus.

4 Hasil dan Pembahasan

Bagian ini menyajikan hasil dan pembahasan dalam membuat model prediksi tunggakan mahasiswa menggunakan model Random Forest dengan Feature Selection. Pembahasan mencangkup pengolahan dataset, pengembangan model, pengujian model dan evaluasi, analisa variabel pengaruh dan memberikan rekomendasi pada UMMADA Cirebon. Menggunakan machine learning aplikasi orange untuk pembuatan modelnya.

4.1 Pengolahan Data

4.1.1 Transformasi Data

Beberapa variabel kategorikal telah diubah menjadi numeric menggunakan label encoding, berikut Tabel 2 merupakan hasilnya :

Tabel 2 Label encoding

Fitur	Nilai Asli	Nilai Numerik
Jenis Kelamin	L, P	0, 1
Status Sosial	Sangat Miskin	0
	Miskin	1
	Cukup	2
	Kaya	3
Pekerjaan Orangtua	PNS	1
	Pegawai Swasta	2
	Wiraswasta	3
	Guru/Dosen Negeri	4
	Lainnya	5

4.1.2 Pemilihan Fitur

Hasil seleksi fitur, menggunakan *recursive feature elimination* dari algoritma *Random Forest*

Tabel 3 Hasil seleksi fitur

Fitur	Score Importance
Rata-rata Tunggakan	0.0572
Asal Daerah	0.0261
Status Sosial	0.0192
Pembayaran Awal	0.0177
Pekerjaan Orangtua	0.0157
Jenis Kelamin	0.0084
Nilai Ipk	0.0059

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

Hasil seleksi fitur pada Tabel 3 di atas menunjukkan bahwa variabel jumlah_tunggakan memiliki kontribusi paling besar terhadap prediksi, dengan nilai *feature importance* tertinggi (0.8499). Hal ini dapat dimaknai bahwa fitur ini secara langsung merepresentasikan risiko keterlambatan atau ketidakmampuan membayar, mirip dengan temuan dalam penelitian [10] dan [14], di mana fitur-fitur yang merefleksikan perilaku pembayaran terbukti sangat berpengaruh dalam model prediktif berbasis Random Forest. Selain itu, rata_rata_tunggakan, asal_daerah, status_sosial, dan pekerjaan_orangtua juga menunjukkan pengaruh meskipun relatif kecil. Ini sejalan dengan temuan pada [9] dan [13], bahwa variabel demografi dan perilaku historis masih relevan namun bukan faktor dominan. Fitur seperti jenis_kelamin dan nilai_ipk memiliki pengaruh rendah, menguatkan konsep bahwa dalam konteks prediksi risiko finansial, variabel akademik dan identitas gender kurang memberikan informasi prediktif yang signifikan, sebagaimana juga tercermin dalam penelitian [8] dan [11].

4.1.3 Pembagian Data

Sebelum menggunakan algoritma klasifikasi *Random Forest* untuk membangun model, langkah pertama yang perlu dilakukan adalah membagi kumpulan data menjadi dua bagian: data pelatihan dan data pengujian. Tujuan dari bagian ini adalah untuk mengevaluasi kinerja model dengan menilai tingkat kesalahan prediksi yang mungkin terjadi. Data pelatihan digunakan untuk melatih algoritma dan membentuk model, dan data pengujian digunakan untuk menguji akurasi model yang telah dibangun. Jika model berhasil, maka dianggap layak digunakan untuk melakukan prediksi terhadap data baru. Dari total data yang tersedia, pembagian dilakukan sebesar 80 persen untuk data pelatihan dan 20 persen untuk data pengujian. Proporsi pembagian data 80:20 antara data latih dan data uji merupakan pendekatan umum dan terbukti efektif dalam banyak penelitian, seperti yang digunakan dalam [3] dan [15]. Hal ini memberikan keseimbangan antara pelatihan model dan validasi performa prediktifnya. Lihat Tabel 4 berikut :

Tabel 4 Pembagian data

Nama Data	Presentase	Jumlah Data
Data Latih	80%	392
Data Pengujian	20%	98
Total	100%	490

4.2 Pengembangan Model

Setelah membagi data menjadi data pelatihan dan data uji, langkah selanjutnya adalah melakukan analisis klasifikasi pada sampel pelatihan yang telah ditentukan dengan menggunakan *Random Forest*. Pengembangan model prediksi dilakukan dengan menggunakan widget *Random Forest* yang tersedia dalam aplikasi *Orange*. Untuk meningkatkan kinerja model, proses penyesuaian *Hyperparameter* dilakukan menggunakan fungsi *gridsearchCV* pada *scikit-learn* dengan *widget Python Script*. Kode program *Python* terhubung langsung dengan widget evaluasi dan data latih. *Number of Trees (n_estimators)*, *Maximum Tree Depth (max_depth)*, dan *Minimum Samples in Leaves (min_samples_leaf)* adalah parameter yang dioptimalkan. Setelah proses tuning, diperoleh konfigurasi model optimal sebagai berikut:

Tabel 5 *Tuning hyperparameter*

Parameter	Rentang Nilai	Nilai Terbaik	Penjelasan
Number of Trees (n_estimators)	[50, 100, 150]	100	Menunjukkan jumlah pohon (tree) dalam *Random Forest*. Nilai 100 memberikan keseimbangan antara akurasi dan waktu komputasi.

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

Maximum Tree Depth (max_depth)	[5, 10, 15]	15	Merupakan kedalaman maksimum setiap pohon keputusan. Nilai 15 berarti model dapat membentuk pohon yang cukup dalam untuk menangkap kompleksitas data.
Minimum Samples in Leaves (min_samples_leaf)	[1, 2, 4]	1	Menentukan jumlah minimum sampel di tiap daun pohon. Nilai 1 memberikan fleksibilitas tinggi untuk membagi node sehingga model lebih kompleks.
Number of Trees (n_estimators)	[50, 100, 150]	100	Menunjukkan jumlah pohon (tree) dalam *Random Forest*. Nilai 100 memberikan keseimbangan antara akurasi dan waktu komputasi.

Kombinasi pada Tabel 5 diatas menghasilkan model Random Forest dengan kemampuan prediksi yang optimal, berdasarkan evaluasi metrik akurasi pada data uji (dari proses GridSearchCV). Nilai-nilai ini menunjukkan bahwa model cenderung membutuhkan pohon yang dalam dan kompleks (depth 15, leaf size 1) dengan jumlah pohon yang moderat (100) agar dapat menangkap pola dalam data secara efektif, ini sesuai dengan karakteristik data keuangan dan risiko yang cenderung memiliki pola tidak linier dan kompleks, sebagaimana dijelaskan dalam [7], [10], dan [16]. Strategi tuning ini juga sejalan dengan pendekatan dalam prediksi energi menggunakan LSTM dalam [17], yang menekankan pentingnya optimasi hiperparameter untuk meminimalkan kesalahan prediksi.

4.3 Pengujian Model dan Evaluasi

Setelah model diperoleh, langkah selanjutnya adalah evaluasi model. Langkah ini bertujuan untuk mengidentifikasi keakuratan model. Ukuran yang digunakan untuk mengevaluasi hasil prediksi model meliputi nilai *accuracy*, *precision*, *recall/sensitivity*, dan *specificity* dengan menggunakan *confusion matrix*.

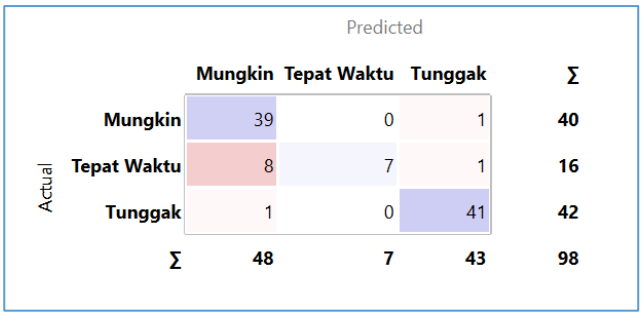
Tabel 6 Hasil test skor

Matrik	Nilai	Penjelasan
AUC	0.980	Area Under Curve: sangat tinggi, menunjukkan kemampuan model membedakan kelas sangat baik.
CA	0.888	Classification Accuracy: akurasi sebesar **88.8%**, artinya model memprediksi benar sekitar 89 dari 100 kasus.
F1	0.875	F1-score: kombinasi *Precision* dan *Recall*, baik untuk data tidak seimbang. Nilai ini menunjukkan keseimbangan yang bagus.
Prec	0.904	Precision: dari semua prediksi "positif", 90.4% memang benar. Model tidak banyak salah deteksi.
Recall	0.888	Recall: dari semua kasus yang benar-benar positif, 88.8% berhasil ditemukan. Model cukup sensitif.
MCC	0.825	Matthews Correlation Coefficient: metrik korelasi yang bagus untuk data tidak seimbang. Nilai 0.825

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

	menunjukkan korelasi kuat antara prediksi dan label asli.

Model pada Tabel 6 di atas menunjukkan kinerja yang sangat baik dengan AUC 0.980, accuracy 88.8%, dan F1-score 0.875. Ini konsisten dengan penelitian pada prediksi fraud dan pinjaman menggunakan Random Forest di [15], [8], dan [9], di mana algoritma ini memberikan performa superior dibandingkan model lain. Tingginya nilai Precision (0.904) dan Recall (0.888) menunjukkan keseimbangan antara menghindari kesalahan deteksi dan menjangkau seluruh kasus positif — sesuai konteks deteksi risiko, seperti pada [15] dan [9]

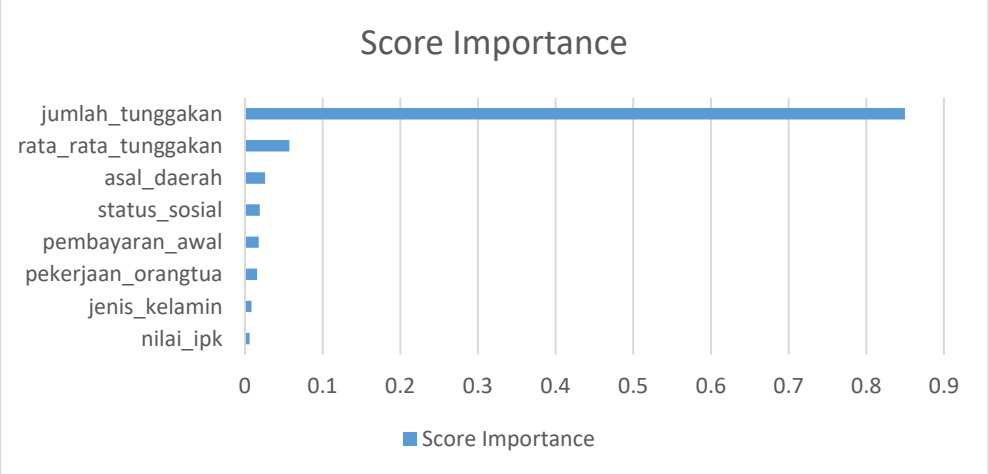


Gambar 2 Hasil *confusion matrix*

Dari hasil confusion matrix pada Gambar 2 di atas, prediksi "Tunggak" sangat akurat: 41 dari 42 benar, hanya ada 1 yang salah. Kemudian untuk Kelas "Mungkin" juga sangat akurat: 39 dari 40 benar, dan Kelas "Tepat Waktu" memiliki kesalahan terbesar: hanya 7 dari 16 yang benar, sisanya sering disalahartikan sebagai "Mungkin".

4.4 Analisa Variabel Pengaruh

Setelah dilakukan pengujian dan evaluasi terhadap model dengan hasil yang sangat baik menunjukkan bahwa model mampu memprediksi dengan tepat dan akurat. Selanjutnya, untuk memahami lebih dalam mengenai kinerja model, penting untuk melakukan analisis variabel yang berpengaruh. Analisis ini akan membantu mengidentifikasi faktor-faktor mana yang paling signifikan dalam mempengaruhi hasil prediksi.



Gambar 3 Grafik hasil *feature importance*

Hasil proses seleksi fitur pada Gambar 3 di atas secara langsung merefleksikan hasil analisis *feature importance* dari algoritma *Random Forest*. Nilai-nilai *importance* yang diperoleh menunjukkan seberapa besar kontribusi masing-masing fitur terhadap akurasi prediksi model. Dalam hal ini, jumlah_tunggakan menempati peringkat teratas dengan nilai feature importance sebesar 0.8499, yang berarti fitur ini memiliki pengaruh dominan dalam menentukan hasil prediksi risiko tunggakan mahasiswa. Analisis feature importance

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

menegaskan bahwa *Random Forest* tidak hanya unggul dalam akurasi, tetapi juga dalam interpretabilitas. Penggunaan *feature importance* untuk menyederhanakan model dengan menghilangkan fitur tidak signifikan (jenis_kelamin dan nilai_ipk) juga direkomendasikan dalam studi [11], [18] di mana seleksi fitur berperan penting dalam menurunkan kompleksitas tanpa mengorbankan akurasi. Pemanfaatan teknik ini mendukung praktik *model refinement* untuk meningkatkan efisiensi komputasi serta akurasi prediksi jangka panjang, seperti disarankan dalam studi-studi yang menggunakan *Recursive Feature Elimination* (RFE) dan *Genetic Algorithm* [14], [18].

4.5 Rekomendasi

Model Random Forest telah diterapkan untuk mengklasifikasikan status mahasiswa ke dalam tiga kategori: "Mungkin", "Tepat Waktu", dan "Tunggak". Hasil evaluasi yang dilakukan menggunakan *metrik performa* dan *confusion matrix* menunjukkan bahwa secara umum model bekerja dengan sangat baik. Kelas "Mungkin" menunjukkan kinerja yang sangat baik dengan nilai *Recall* sebesar 0.975, yang berarti hampir semua data "Mungkin" berhasil dikenali oleh model. Namun, *Precision-nya* hanya 0.813, mengindikasikan bahwa masih ada data dari kelas lain (seperti "Tepat Waktu" atau "Tunggak") yang secara keliru diklasifikasikan sebagai "Mungkin". Sementara itu, kelas "Tepat Waktu" menunjukkan kelemahan yang cukup signifikan. Meskipun memiliki Precision sempurna (1.000)—artinya semua prediksi "Tepat Waktu" memang benar—namun nilai Recall-nya rendah (0.438). Ini berarti lebih dari separuh data "Tepat Waktu" gagal dikenali, dan sebagian besar justru diprediksi sebagai "Mungkin". Hal ini bisa disebabkan oleh jumlah data kelas ini yang relatif lebih sedikit dibandingkan kelas lainnya, atau adanya kemiripan karakteristik dengan kelas "Mungkin". Di sisi lain, model menunjukkan performansi yang sangat kuat pada kelas "Tunggak", dengan Precision sebesar 0.953 dan *Recall* sebesar 0.976. Hal ini mencerminkan kemampuan model yang sangat baik dalam mengenali dan memprediksi data mahasiswa yang menunggak.

5 Kesimpulan

Model Random Forest yang dikembangkan dengan *feature selection* dan *tuning hyperparameter* pada aplikasi *Orange* berhasil memprediksi risiko tunggakan mahasiswa dengan sangat baik. Variabel paling dominan dalam prediksi adalah jumlah tunggakan, diikuti oleh rata-rata tunggakan, asal daerah, status sosial, dan pekerjaan orangtua. Pengujian menunjukkan metrik kinerja model yang tinggi seperti *AUC* 0.980, akurasi 88.8%, dan *F1-score* 0.875, dengan kinerja terbaik pada kelas "Tunggak" dan cukup baik untuk kelas "Mungkin". Kelas "Tepat Waktu" memiliki tantangan dalam *recall* yang rendah karena data lebih sedikit dan kemiripan karakteristiknya dengan kelas lain. Model ini menunjukkan bahwa *Random Forest* unggul dalam menangkap pola kompleks untuk prediksi risiko finansial mahasiswa, sejalan dengan studi lain yang menggunakan *Random Forest* untuk prediksi dalam bidang pendidikan dan keuangan. [19]
Penelitian ini menunjukkan bahwa model *Random Forest* efektif dalam memprediksi risiko tunggakan mahasiswa dengan akurasi tinggi dan kemampuan interpretasi fitur penting seperti jumlah tunggakan. Namun, penelitian terbatas pada data dari satu institusi dan menghadapi tantangan data tidak seimbang yang memengaruhi kemampuan prediksi terutama pada kelas "Tepat Waktu". Kontribusinya terletak pada penerapan *tuning hyperparameter* dan *feature selection* dalam *platform Orange* yang memudahkan implementasi praktis serta memberikan wawasan pengelolaan risiko finansial mahasiswa yang lebih baik. Saran untuk penelitian selanjutnya mencakup penanganan ketidakseimbangan data melalui teknik sampling, pengembangan model dengan data multisumber dari berbagai institusi, serta eksplorasi fitur baru dan metode lain seperti ensemble atau deep learning untuk meningkatkan performa model. Studi lain seperti [5] dan [7] mendukung temuan bahwa *Random Forest* sangat baik dalam menangani masalah klasifikasi risiko di bidang pendidikan dan keuangan, sehingga penelitian ini mengisi gap dengan fokus pada prediksi tunggakan mahasiswa dan penyediaan solusi yang aplikatif bagi institusi pendidikan tinggi.

Referensi

[1] B. S. R. Sudirman, I. Oktavia, F. H. Sarumaha, “Analisis Keterlambatan Pembayaran dalam Industri *Fintech* menggunakan Algoritma C4.5,” Vol. 11, No. 2, pp. 166–177, 2024.

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

[2] M. Nurhasanah, I. Zufria, U. Islam, and N. Sumatera, "Implementasi Algoritma C5.0 untuk memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan pada SMP Swasta An-Naas Binjai," Vol. 9, No. 1, pp. 107–116, 2024.

[3] F. Riskiyono and D. Mahdiana, "*Implementation of Random Forest Algorithm for Graduation Prediction*," *Sinkron*, Vol. 8, No. 3, pp. 1662–1670, 2024, DOI: 10.33395/sinkron.v8i3.13750.

[4] R. Bakri, N. P. Astuti, and A. S. Ahmar, "*Machine Learning Algorithms with Parameter Tuning to Predict Students' Graduation-on-time: A Case Study in Higher Education*," *J. Appl. Sci. Eng. Technol. Educ.*, Vol. 4, No. 2, pp. 259–265, 2022, DOI: 10.35877/454ri.asci1581.

[5] A. Akbar, Z. Indra, Y. Andriyani, and T. Melia, "*Implementation of the Random Forest Method for Predicting Students' Length of Study*," *J. Stat. Methods Data SCI.*, Vol. 1, No. 2, pp. 32–43, 2024, DOI: 10.31258/jsmds.v1i2.15.

[6] Y. Abubakar, N. Bahiah, and H. Ahmad, "*Prediction of Students' Performance in E-Learning Environment using Random Forest*," *Int. J. Innov. Comput.*, Vol. 7, No. 2, pp. 1–5, 2017, [Online]. Available: http://se.fsksm.utm.my/ijic/index.php/ijic

[7] D. Puspita, S. Nilam, M. I. Arifyanto, P. Studi, and M. Sains, "*Prediction of Electricity Bill Payment Delays for Customers using A Machine Learning Approach* Listrik Pelanggan dengan Pendekatan *Machine*," Vol. 10, No. 1, pp. 446–457, 2025.

[8] R. A. Zuama, N. Ichsan, A. B. Pohan, M. S. Azis, and M. Lase, "*An Implementation of Machine Learning on Loan Default Prediction based on Customer Behavior*," *J. Info Sains Inform. dan Sains*, Vol. 14, No. 01, pp. 157–164, 2024, DOI: 10.54209/infosains.v14i01.

[9] R. Pullela and K. Vaddepally, "*Payment Date Prediction using Machine Learning*," *Int. J. Res. Trends Innov.*, Vol. 8, No. 5, pp. 224–228, 2023, [Online]. Available: https://www.ijrti.org/papers/IJRTI2305034.pdf

[10] I. A. Rahmi, F. M. Afendi, and A. Kurnia, "Metode AdaBoost dan Random Forest untuk Prediksi Peserta JKN-KIS yang Menunggak," *Jambura J. Math.*, Vol. 5, No. 1, pp. 83–94, 2023, doi: 10.34312/jjom.v5i1.15869.

[11] D. Germandy, C. Putra, and A. T. Putra, "*Optimizing Random Forest for Predicting Thoracic Surgery Success in Lung Cancer using Recursive Feature Elimination and GridSearchCV*," Vol. 2, No. 2, pp. 97–105, 2024, DOI: 10.15294/rji.v2i2.73154.

[12] R. S. Baker and P. S. Inventado, "*Emergence and Innovation in Digital Learning: Foundations and Applications*," *Emerg. Innov. Digit. Learn. Found. Appl.*, pp. 1–13, 2016, DOI: 10.15215/aupress/9781771991490.01.

[13] Sutarman, R. Siringoringo, D. Arisandi, E. Kurniawan, and E. B. Nababan, "Model Klasifikasi dengan *Logistic Regression* dan *Recursive Feature Elimination* pada Data Tidak Seimbang," *J. Teknol. Inf. dan Ilmu Komput.*, Vol. 11, No. 4, pp. 735–742, 2024, DOI: 10.25126/jtiik.1148198.

[14] Y. Priyatno and D. Wirantanu, "*Comparison of Genetic Algorithm and Recursive Feature Elimination on*," *J. Resti*, Vol. 5, No. 158, pp. 189–198, 2024.

[15] L. Moumeni, M. Saber, I. Slimani, I. Elfarissi, and Z. Bougroun, "*Machine Learning for Credit Card Fraud Detection*," *Lect. Notes Electr. Eng.*, Vol. 745, No. 24, pp. 211–221, 2022, DOI: 10.1007/978-981-33-6893-4_20.

[16] V. Sheth, U. Tripathi, and A. Sharma, "*A Comparative Analysis of Machine Learning Algorithms for Classification Purpose*," *Procedia Comput. SCI.*, Vol. 215, pp. 422–431, 2022, DOI: 10.1016/j.procs.2022.12.044.

[17] P. S. Pravin, J. Z. M. Tan, K. S. Yap, and Z. Wu, "*Hyperparameter Optimization Strategies for Machine Learning-based Stochastic Energy Efficient Scheduling in Cyber-Physical Production Systems*," *Digit. Chem. Eng.*, Vol. 4, No. July, p. 100047, 2022, DOI: 10.1016/j.dche.2022.100047.

[18] D. Wicaksono, A. Mareta, and A. Erdiyanto, "*Machine Learning-based Cow Milk Quality Classification using Recursive Feature Elimination Cross-Validation*," Vol. 14, No. 2, 2024.

[19] 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.