Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

Penerapan Metode GA-RU Pada Algoritma Random Forest Untuk Mengatasi Class Imbalance Data Beasiswa KIP-Kuliah

Febrian Nor Rahman, Taghfirul Azhima Yoga Siswa*, Rudiman

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
Email: [1]2011102441187@umkt.ac.id, [2,*]tay758@umkt.ac.id, [3]rud959@umkt.ac.id,
Email Penulis Korespondensi: tay758@umkt.ac.id
Submitted: **16/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Ketidakseimbangan kelas (class imbalance) merupakan tantangan umum dalam analisis data, di mana jumlah data pada kelas mayoritas jauh lebih besar dibandingkan kelas minoritas. Kondisi ini menyebabkan model klasifikasi cenderung memprediksi kelas mayoritas, sehingga tingkat akurasi untuk mengidentifikasi kelas minoritas menjadi rendah. Penelitian ini mengusulkan penerapan Genetic Algorithm (GA) yang dikombinasikan dengan Random Undersampling (RU) pada algoritma Random Forest untuk menangani masalah ketidakseimbangan kelas pada data penerima Beasiswa Kartu Indonesia Pintar (KIP) di Universitas Muhammadiyah Kalimantan Timur. Data yang digunakan berjumlah 1.080 record dengan 37 fitur terkait faktor sosial-ekonomi penerima beasiswa, yang setelah proses pembersihan data menyisakan 1.075 record. Hasil penelitian menunjukkan bahwa metode Random Undersampling meningkatkan akurasi model Random Forest dari 84,27% menjadi 85,06%. Meskipun peningkatan ini terlihat kecil, hal ini signifikan karena menunjukkan peningkatan stabilitas model dalam mengklasifikasikan kelas minoritas, yang sebelumnya memiliki akurasi rendah. Kombinasi GA dan RU terbukti efektif dalam meningkatkan performa model, menghasilkan klasifikasi yang lebih stabil untuk kelas minoritas. Penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem seleksi penerima beasiswa yang lebih akurat dan efisien serta menjadi acuan untuk penelitian di bidang data mining dan machine learning.

Kata Kunci: Genetic Algorithm (GA); Random Undersampling (RU); Random Forest (RF); Klasifikasi.

Abstract−Class imbalance is a common challenge in data analysis, where the majority class significantly outnumbers the minority class. This condition causes classification models to lean toward predicting the majority class, resulting in low accuracy in identifying the minority class. This study proposes the application of Genetic Algorithm (GA) combined with Random Undersampling (RU) on the Random Forest algorithm to address class imbalance issues in the dataset of Indonesia Smart Card (KIP) scholarship recipients at Universitas Muhammadiyah Kalimantan Timur. The dataset comprises 1,080 records with 37 features related to the socio-economic factors of the scholarship recipients. After data cleaning, 1,075 records were retained. The results indicate that the Random Undersampling method improved the accuracy of the Random Forest model from 84.27% to 85.06%. Although this improvement appears modest, it is significant as it demonstrates increased model stability in classifying the minority class, which previously had low accuracy. The combination of GA and RU proved effective in enhancing model performance, resulting in more stable classification for the minority class. This study is expected to contribute to the development of more accurate and efficient scholarship selection systems and serve as a reference for research in data mining and machine learning.

Keywords: Genetic Algorithm (GA); Random Undersampling (RU); Random Forest (RF); Classification.

1. PENDAHULUAN

Beasiswa Kartu Indonesia Pintar (KIP) Kuliah adalah program pemerintah Indonesia yang bertujuan memberikan dukungan finansial kepada mahasiswa dari keluarga kurang mampu. Program ini merupakan transformasi dari Bidikmisi, dengan tujuan meningkatkan akses pendidikan tinggi tanpa terbebani oleh masalah ekonomi melalui pemberian bantuan biaya pendidikan dan biaya hidup [1].

Ketidakseimbangan kelas merupakan tantangan signifikan dalam analisis data, terutama dalam konteks klasifikasi. Dalam banyak aplikasi, seperti deteksi penipuan, diagnosis medis, dan seleksi penerima beasiswa, jumlah data pada kelas mayoritas sering kali jauh lebih besar dibandingkan kelas minoritas. Hal ini dapat menyebabkan model klasifikasi cenderung memprediksi kelas mayoritas, sehingga mengabaikan kelas minoritas yang penting.Dalam penelitian ini, kami mengusulkan pendekatan inovatif yang menggabungkan Genetic Algorithm (GA) dan Random Undersampling (RU) pada algoritma Random Forest untuk mengatasi masalah ketidakseimbangan kelas pada data penerima Beasiswa Kartu Indonesia Pintar (KIP) di Universitas Muhammadiyah Kalimantan Timur. Pendekatan ini tidak hanya bertujuan untuk meningkatkan akurasi model, tetapi juga untuk meningkatkan stabilitas klasifikasi pada kelas minoritas, yang sering kali terabaikan dalam penelitian sebelumnya.

Keunggulan dari pendekatan ini terletak pada kombinasi dua teknik: GA yang berfungsi untuk mengoptimalkan pemilihan fitur dan RU yang membantu menyeimbangkan jumlah data antara kelas mayoritas dan minoritas. Dengan menggunakan GA, kami dapat memilih fitur-fitur yang paling relevan, sehingga meningkatkan efisiensi model. Sementara itu, RU membantu mengurangi bias yang dihasilkan oleh ketidakseimbangan kelas, sehingga model dapat lebih fokus pada klasifikasi kelas minoritas.

Penelitian ini berbeda dari studi-studi sebelumnya yang hanya menggunakan satu teknik untuk menangani ketidakseimbangan kelas. Dengan menggabungkan GA dan RU, kami berharap dapat memberikan kontribusi yang signifikan dalam pengembangan sistem seleksi penerima beasiswa yang lebih akurat dan efisien. Selain itu, hasil dari penelitian ini diharapkan dapat menjadi acuan bagi penelitian lebih lanjut di bidang data mining dan machine learning, khususnya dalam konteks klasifikasi dengan ketidakseimbangan kelas.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

Dalam beberapa tahun terakhir, penelitian mengenai penerapan machine learning dalam proses seleksi penerima KIP Kuliah semakin berkembang, menunjukkan potensinya dalam meningkatkan efisiensi dan akurasi proses seleksi. [2] meneliti model algoritma yang akurat dalam memprediksi kelayakan mahasiswa sebagai penerima beasiswa, sedangkan penelitian oleh (Budiarto et al., 2022) menggunakan machine learning untuk menganalisis atribut seperti pendapatan orang tua, jumlah saudara, status pemegang KIP, dan jarak rumah ke sekolah. Namun, masalah ketidakseimbangan data (class imbalance) sering menjadi hambatan, di mana model cenderung bias terhadap kelas mayoritas sehingga memengaruhi performa prediksi [4].

Untuk mengatasi permasalahan ketidakseimbangan kelas pada data maka diterapkan data balancing class Imbalance. Class Imbalance adalah ketidakseimbangan kelas dapat menghasilkan akurasi prediksi yang baik untuk kelas mayoritas, tetapi menjadi kurang efektif dalam memprediksi kelas minoritas, sehingga nilai akurasi dari pengklasifikasi menjadi tidak optimal. Masalah ketidak seimbangan kelas umumnya dapat diatasi melalui dua pendekatan, yaitu pada tingkat data dan tingkat algoritma. Pendekatan tingkat data bertujuan untuk meningkatkan keseimbangan kelas, sementara pendekatan tingkat algoritma berfokus pada perbaikan algoritma atau penggabungan (ensemble) pengklasifikasi agar lebih efektif dalam mengenali kelas minoritas [5].

Dalam konteks seleksi penerima Beasiswa Kartu Indonesia Pintar (KIP) Kuliah, algoritma Random Forest (RF) telah menunjukkan potensi signifikan dalam meningkatkan akurasi dan efisiensi proses seleksi. RF, sebagai algoritma ensemble, menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan akurat. Algoritma ini mampu menganalisis berbagai faktor, seperti prestasi akademik, kondisi ekonomi, dan keterlibatan dalam kegiatan ekstrakurikuler [6]. Salah satu keunggulan RF adalah kemampuannya dalam menangani data kompleks dan beragam. Penelitian [6] menunjukkan bahwa RF mencapai akurasi sebesar 97,2% pada data uji dalam analisis data dengan banyak variabel. Namun, penelitian tersebut tidak menerapkan balancing data, yang dapat memengaruhi kinerja RF terhadap kelas minoritas.

Klasifikasi merupakan salah satu metode dalam pembelajaran mesin yang digunakan untuk membagi data ke dalam kelompok tertentu berdasarkan pola yang telah dikenali dari data pelatihan. Dalam penelitian ini, teknik klasifikasi diterapkan untuk menganalisis sentimen masyarakat terhadap pemindahan ibu kota Indonesia dengan mengkategorikan opini mereka ke dalam tiga jenis sentimen utama, yaitu positif, negatif, dan netral [7]. *Random Forest* adalah algoritma pembelajaran yang bersifat terawasi. "Hutan" yang dibentuk oleh algoritma ini merupakan sekumpulan pohon keputusan, yang umumnya dilatih menggunakan metode "*bagging*". Konsep dasar dari metode *bagging* adalah bahwa penggabungan beberapa model pembelajaran dapat meningkatkan hasil secara keseluruhan [8]. Random Forest merupakan algoritma pembelajaran mesin yang sangat ideal untuk tugas klasifikasi, terutama ketika berhadapan dengan data berdimensi tinggi. Algoritma ini beroperasi dengan membangun beberapa pohon keputusan dari subset dataset yang dipilih secara acak, kemudian mengombinasikan hasil prediksi dari setiap pohon untuk menghasilkan keputusan akhir [9] dan untuk Keunggulan dari metode *Random Forest* terletak pada kemampuannya untuk mengatasi *overfitting* serta menghasilkan prediksi yang tepat [10]. Algoritma ini termasuk dalam metode ensemble, yang bertujuan untuk meningkatkan ketepatan klasifikasi dengan menggabungkan sejumlah decision tree. Proses prediksi dalam Random Forest dilakukan dengan mekanisme pemungutan suara mayoritas (majority voting) dalam klasifikasi atau perhitungan rata-rata dalam regresi. Dengan menggunakan berbagai model pohon keputusan, metode ini mampu mengatasi permasalahan klasifikasi dan regresi secara lebih akurat dibandingkan dengan pohon keputusan tunggal [11] Pada Penelitian [12] mengindikasikan bahwa Random Forest dapat mencapai tingkat akurasi yang sangat tinggi, bahkan tanpa melalui proses optimasi.Untuk Keunggulan dari metode Random Forest terletak pada kemampuannya untuk mengatasi overfitting serta menghasilkan prediksi yang tepat [10], Dalam penelitian [13] algoritma Random Forest Classifier digunakan untuk memprediksi performa akademik mahasiswa. Hasil akurasi yang diperoleh adalah 91,66%.

Genetic Algorithm (GA) adalah algoritma pencarian atau heuristik yang terinspirasi oleh prinsip seleksi alam dan evolusi biologis yang diusulkan oleh Charles Darwin. Algoritma ini menggunakan proses seleksi untuk menghasilkan keturunan terbaik bagi populasi generasi berikutnya [14]. Dalam konteks seleksi penerima Beasiswa Kartu Indonesia Pintar (KIP) Kuliah, GA menunjukkan potensi besar dalam meningkatkan efektivitas dan efisiensi proses pengambilan keputusan. Penelitian oleh [15] mencatat akurasi sebesar 88,0% dalam penggunaan metode Fuzzy Tsukamoto dan GA, yang relevan untuk masalah kompleks seperti pemilihan penerima beasiswa.

Random Undersampling (RU) adalah metode yang menghapus data secara acak dari kelas mayoritas dalam dataset pelatihan untuk mengatasi ketidakseimbangan data [16]. Dalam seleksi penerima Beasiswa Kartu Indonesia Pintar (KIP) Kuliah, RU membantu mengurangi bias model prediksi terhadap kelas mayoritas yang dapat menyebabkan mahasiswa berhak tidak terdeteksi. Penelitian [17] bahwa penerapan RU dalam analisis sentimen meningkatkan performa model machine learning, menghasilkan akurasi sebesar 61,94%, yang diukur menggunakan metrik seperti akurasi dan presisi. Namun, algoritma ini memiliki kekurangan, seperti menurunnya nilai spesifisitas (0.4393) saat digunakan dengan Random Forest pada data asli yang tidak seimbang, menunjukkan kesulitan model dalam memprediksi kelas minoritas secara akurat [18].

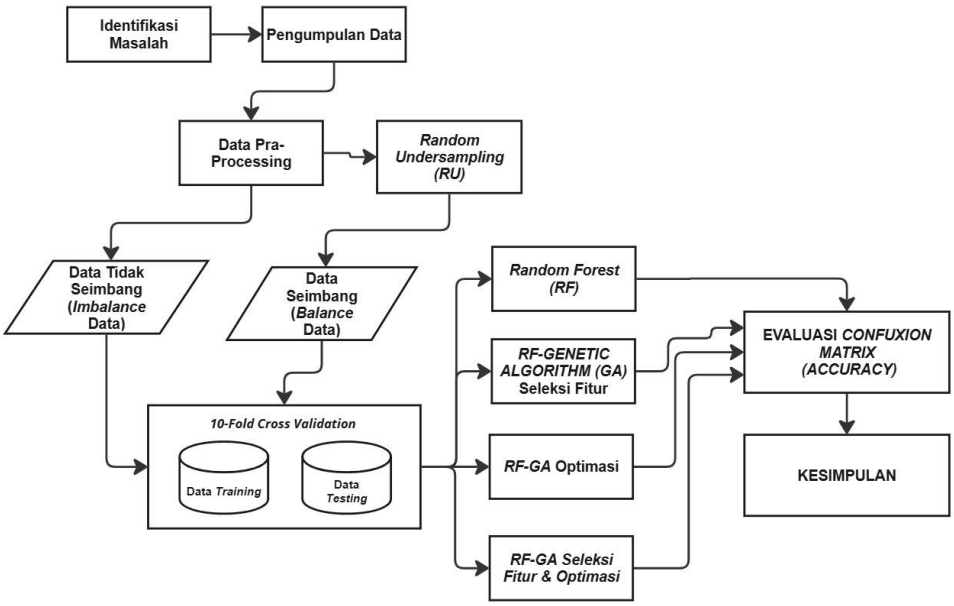
Sehingga dalam penelitian ini akan menggabungkan tiga metode, yaitu Genetic Algorithm (GA), Random Undersampling (RU), dan Random Forest (RF) untuk meningkatkan akurasi dan efisiensi dalam meningkatkan akurasi klasifikasi dengan mengatasi kendala yang ada pada dataset yaitu class Imbalance data beasiswa KIP KULIAH.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pada Gambar 1, menunjukan penelitian ini dimulai dengan identifikasi masalah. Setelah itu dilakukannya pengumpulan data, lalu setelah data didapatkan dilanjut dengan proses Data Pre-processing yang berisikan data integration, data selection, data transformation, data cleaning, dan data balancing. Setelah melewati tahap pre-processing, Data terbagi menjadi dua yaitu Imbalance Data (tidak seimbang) dan Balance Data (Seimbang). Untuk tahap balancing data metode yang digunakan adalah Random Undersampling (RU), Lalu masing-masing divalidasi menggunakan 10-Fold Cross Validation dengan pembagian Data Training dan Data Testing. Data sudah siap diolah dan masuk ke tahap permodelan. Tahap ini dilakukan dengan dua kondisi data imbalance dan data balance. Dimulai dengan permodelan pertama yaitu menggunakan algoritma klasfikasi Random Forest (RF). Berikutnya permodelan kedua diterapkan dengan algoritma RF ditambah dengan Genetic Algorithm (GA) untuk seleksi fitur. Lalu permodelan ketiga menerapkan RF ditambah dengan GA untuk Optimasi, dan yang terakhir Permodelan menggunakan RF ditambah GA untuk seleksi fitur dan optimasi. Setelah itu didapatkanlah hasil akurasi.



Gambar 1. Prosedur Penelitian

2.2 Identifikasi Masalah

Tahap awal dari penelitian ini dimulai dengan identifikasi masalah, yang berfungsi sebagai panduan untuk seluruh proses penelitian. Masalah utama yang diangkat adalah bagaimana cara menentukan metode yang paling efektif untuk menganalisis dampak kondisi ekonomi dan sosial keluarga terhadap keberhasilan akademis penerima beasiswa Kartu Indonesia Pintar (KIP) di Universitas Muhammadiyah Kalimantan Timur (UMKT). Selain itu, dilakukan kajian pustaka untuk mengidentifikasi celah dalam penelitian yang ada terkait faktor-faktor ekonomi dan sosial yang memengaruhi prestasi akademis penerima beasiswa, serta bagaimana penerapan Monitoring, Control, dan Evaluation (MCK) dapat dimanfaatkan untuk meningkatkan efektivitas program beasiswa.

2.3 Pengumpulan Data

Penelitian ini menggunakan data beasiswa KIP-Kuliah tahun 2021–2023 dari Bagian Kemahasiswaan Universitas Muhammadiyah Kalimantan Timur. Data mencakup 37 fitur relevan yang mendukung proses klasifikasi beasiswa KIP-Kuliah.

2.4 Data Pre-processing

Dalam penelitian ini tahap sistem pertama yang akan berjalan adalah pre-processing. langkah penting dalam pembelajaran mesin untuk menyiapkan data mentah sebelum analisis dan pelatihan model [19]. Data beasiswa KIP-Kuliah dari Bagian Kemahasiswaan UMKT memerlukan pengolahan seperti data integration, data selection, data transformation, data cleaning dan data balancing.

a. Data Integration
   Data integration merupakan tahapan di mana data dari berbagai sumber yang berbeda digabungkan menjadi satu kesatuan yang terintegrasi. Proses ini bertujuan untuk menciptakan kumpulan data yang lebih konsisten, sehingga analisis yang dilakukan dapat mencakup berbagai perspektif dan menghasilkan wawasan yang lebih komprehensif. Dengan data yang terintegrasi, pengambilan keputusan menjadi lebih terarah dan akurat [20].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

b. Data Selection
Data selection adalah langkah dalam proses data mining yang berfokus pada pemilihan atribut atau fitur yang dianggap paling relevan dan signifikan dari kumpulan data yang tersedia. Tujuan dari tahap ini adalah untuk menyederhanakan proses analisis dengan menghilangkan informasi yang kurang penting, sehingga dapat meningkatkan efisiensi dan akurasi dalam pengolahan data. Pemilihan fitur yang tepat juga dapat membantu mengurangi kompleksitas model yang dibangun [21].

c. Data Transformation
Data transformation adalah proses mengubah data ke dalam format atau skala yang sesuai untuk analisis. Dalam prosesnya data transformation untuk mengubah data kategorikal (data string) menjadi format numerik dapat dilakukan dengan menggunakan library sklearn.preprocessing dengan fungsi LabelEncoder [22].

d. Data Cleaning
Data cleaning adalah proses penghapusan atau koreksi data yang salah, tidak lengkap, atau tidak konsisten. Langkah ini penting untuk memastikan keakuratan analisis dan akurasi. Dalam penelitian ini, data cleaning akan menggunakan fungsi dari library pandas yang bernama dropna() untuk menghapus baris yang mengandung nilai NaN ataupun satu nilai yang hilang di dalam suatu baris [23].

e. Data Balancing
Tahap-tahap balancing dilakukan dengan mengidentifikasi ketidakseimbangan kelas dengan menghitung jumlah sampel di setiap kelas, baik mayoritas maupun minoritas. Setelah itu, dilakukan penghapusan sampel secara acak dari kelas mayoritas untuk menyeimbangkan jumlahnya dengan kelas minoritas. Kemudian, jumlah sampel dari setiap kelas diperiksa kembali untuk memastikan keseimbangan sudah tercapai. Setelah data seimbang, dataset yang sudah di-undersample tersebut digunakan untuk melatih model machine learning, guna menghindari bias terhadap kelas mayoritas dan tahap ini disebut Random Undersampling (RU) [24].

2.5 Random Undersampling

(RU) mengidentifikasi ketidakseimbangan kelas dengan menghitung jumlah sampel di setiap kelas, baik mayoritas maupun minoritas. Setelah itu, dilakukan penghapusan sampel secara acak dari kelas mayoritas untuk menyeimbangkan jumlahnya dengan kelas minoritas. Kemudian, jumlah sampel dari setiap kelas diperiksa kembali untuk memastikan keseimbangan sudah tercapai. Setelah data seimbang, dataset yang sudah di-*undersample* tersebut digunakan untuk melatih model *machine learning*, guna menghindari bias terhadap kelas mayoritas [24].

$$SSize_{MA}^{i} = (m\ x\ Size_{Mi})\ x\ \frac{Size_{MA}^{i}/Size_{Mi}^{i}}{\sum_{i=1}^{K} Size_{MA}^{i}/Size_{Mi}^{i}} \quad (1)$$

Pada Rumus 1, SSize_MA^i merepresentasikan ukuran agregat yang dihitung berdasarkan ukuran individu Size_Mi yang dikalikan dengan rasio ukuran agregat terhadap ukuran individu, kemudian dinormalisasi dengan jumlah total rasio ukuran agregat terhadap ukuran individu untuk semua K komponen. Rumus ini bertujuan untuk menentukan kontribusi relatif setiap elemen dalam sistem keamanan nirkabel di rumah pintar berdasarkan ukuran masing-masing.

2.6 Pembagian Data Training dan Data Testing

Tahapan pembagian data menjadi data Training dan data Testing adalah langkah awal yang penting dalam proses data mining. Teknik 10-Fold Cross-Validation digunakan untuk mengevaluasi kinerja model dengan membagi dataset menjadi 10 bagian, di mana setiap bagian digunakan sebagai data Testing secara bergantian. Penelitian ini menunjukkan bahwa metode ini dapat meningkatkan akurasi dan keandalan model [25].

2.7 Permodelan

Pada tahap ini, akan diuraikan mengenai model yang diterapkan dalam penelitian ini. Model klasifikasi yang digunakan adalah Random Forest (RF), dengan Genetic Algorithm (GA) berfungsi sebagai metode untuk pemilihan fitur, serta Random Undersampling (RU) sebagai strategi untuk menangani ketidakseimbangan kelas dengan cara mengurangi jumlah sampel pada kelas yang lebih dominan.

2.8 Algoritma Random Forest

Random Forest adalah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, yang terdiri dari sekumpulan pohon keputusan (decision trees). Metode ini sangat populer karena kemampuannya untuk menangani data yang kompleks dan memberikan hasil yang akurat. Random Forest bekerja dengan cara membangun beberapa pohon keputusan selama pelatihan dan menggabungkan hasilnya untuk meningkatkan akurasi dan mengurangi risiko overfitting [26], Menggunakan rumus random forest.

$$\mathrm{F(x)} = \frac{1}{j}\sum_{j=1}^{J} h_j(x) \quad (2)$$

Pada rumus 2, F(x) merupakan nilai rata-rata dari fungsi h_j(x) yang dihitung dengan menjumlahkan seluruh h_j(x) untuk J komponen, kemudian membaginya dengan jumlah komponen J. Rumus ini umumnya digunakan dalam

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

metode ensemble dalam pembelajaran mesin atau sistem prediksi, di mana hasil akhir diperoleh sebagai agregasi dari beberapa model atau fungsi individu untuk meningkatkan akurasi dan keandalan dalam sistem keamanan nirkabel di rumah pintar.

2.9 Genetic Algorithm

Genetic algorithm (GA) adalah teknik optimasi yang terinspirasi oleh proses evolusi biologis, yang digunakan untuk menyelesaikan masalah kompleks dalam berbagai bidang, termasuk pengoptimalan, pembelajaran mesin, dan pemecahan masalah. GA bekerja dengan cara mensimulasikan proses seleksi alam, di mana individu dalam populasi dievaluasi berdasarkan fungsi tujuan, dan individu yang lebih baik memiliki peluang lebih besar untuk bertahan dan berkembang. Dalam konteks ini, GA sering digunakan untuk menemukan solusi optimal dalam ruang pencarian yang besar dan kompleks [27]. Adapun rumus menurut [28] sebagai berikut:

$$R = (G + \sqrt[2]{g}\,)/3G \tag{3}$$

Pada Rumus 3, Genetic Algorithm (GA) digunakan untuk mengoptimalkan parameter dalam proses klasifikasi. Persamaan $R = (G + \sqrt[2]{g}\,)$ menunjukkan rasio seleksi yang mempertimbangkan faktor G sebagai nilai generasi dan g sebagai nilai fitness individu. Dengan pendekatan ini, algoritma dapat menyeimbangkan eksplorasi dan eksploitasi dalam pencarian solusi optimal, sehingga meningkatkan kinerja model dalam menangani ketidakseimbangan kelas pada data.

2.10 Evaluasi

Tahap evaluasi merupakan langkah penting setelah pembentukan model. Di tahap ini, performa model diukur untuk mengevaluasi akurasi dan kualitas data latih yang digunakan. Pengujian dilakukan dengan teknik Confusion Matrix. Confusion Matrix adalah sebuah teknik yang digunakan untuk melakukan perhitungan akurasi pada data mining [29]. Evaluasi yang digunakan pada penelitian ini adalah performa accuracy (akurasi). Adapun rumusnya adalah:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ X } 100\% \tag{4}$$

Pada Rumus 4, jumlah prediksi yang benar (True Positive (TP) + True Negative (TN)) terhadap total keseluruhan prediksi (TP + TN + False Positive (FP) + False Negative (FN)). Hasilnya dikalikan 100% untuk menyatakan akurasi dalam bentuk persentase, yang menunjukkan seberapa andal sistem dalam mengklasifikasikan suatu kejadian dengan benar.

3. HASIL DAN PEMBAHASAN

Penelitian ini memanfaatkan data dari program beasiswa KIP-UMKT, yang mencakup informasi peserta beasiswa pada rentang tahun 2021 hingga 2023. Data tersebut dikumpulkan melalui bagian kemahasiswaan dan mencakup berbagai faktor sosial-ekonomi yang digunakan dalam proses seleksi penerima beasiswa. Dataset ini terdiri dari 37 fitur dengan total 1.080 entri. Kategori atau label pada data dibagi menjadi dua kelas: "diterima," yang mencakup 425 entri, dan kelas "ditolak," dengan 655 entri. Secara rinci, 37 fitur tersebut mencakup informasi seperti nomor pendaftaran, nama siswa, NIK, nomor kartu keluarga, NIK keluarga, NISN, status DTKS, status P3KE, nomor KIP, nomor KKS, asal sekolah, kabupaten/kota asal sekolah, provinsi asal sekolah, tempat lahir, tanggal lahir, jenis kelamin, alamat tinggal, nomor handphone, alamat email, nama ayah, pekerjaan ayah, penghasilan ayah, status ayah, nama ibu, pekerjaan ibu, penghasilan ibu, status ibu, jumlah tanggungan, kepemilikan rumah, tahun perolehan, sumber listrik, luas tanah, luas bangunan, sumber air, fasilitas MCK, jarak ke pusat kota, dan status pengajuan.

Hasil penelitian ini menunjukkan bahwa penerapan kombinasi Genetic Algorithm (GA) dan Random Undersampling (RU) pada algoritma Random Forest berhasil meningkatkan akurasi model dalam mengklasifikasikan penerima Beasiswa Kartu Indonesia Pintar (KIP). Peningkatan akurasi ini tidak hanya penting dari segi statistik, tetapi juga memiliki implikasi praktis yang signifikan dalam proses seleksi beasiswa. Dengan menggunakan pendekatan yang diusulkan, pihak universitas dapat lebih akurat dalam mengidentifikasi calon penerima beasiswa yang memenuhi syarat, memastikan bahwa bantuan diberikan kepada mahasiswa yang benar-benar membutuhkan, berdasarkan kriteria yang telah ditetapkan. Salah satu manfaat utama dari pendekatan ini adalah peningkatan stabilitas dalam mengklasifikasikan kelas minoritas, yang berarti mahasiswa dari latar belakang yang kurang beruntung, yang mungkin tidak terwakili dengan baik dalam data, dapat lebih mudah teridentifikasi dan mendapatkan bantuan yang mereka butuhkan. Selain itu, dengan model yang lebih akurat, pengambil keputusan di universitas dapat membuat keputusan yang lebih baik dan lebih adil dalam proses seleksi, meningkatkan kepercayaan masyarakat terhadap sistem seleksi beasiswa dan mendorong lebih banyak mahasiswa untuk mendaftar.

Meskipun hasil penelitian ini menjanjikan, ada beberapa keterbatasan yang perlu diperhatikan. Pertama, penelitian ini menggunakan dataset dari satu universitas, yang mungkin tidak sepenuhnya mewakili populasi penerima beasiswa di seluruh Indonesia, sehingga hasilnya mungkin tidak dapat digeneralisasi ke konteks yang lebih luas tanpa penelitian lebih lanjut. Kedua, meskipun Random Forest adalah algoritma yang kuat, kompleksitas model dapat menjadi tantangan dalam interpretasi hasil, dan pengambil keputusan mungkin memerlukan pemahaman yang lebih

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

baik tentang bagaimana model membuat keputusan untuk dapat mengandalkannya dalam proses seleksi. Terakhir, pemilihan fitur yang dilakukan oleh GA mungkin tidak mencakup semua variabel yang relevan dalam menentukan kelayakan penerima beasiswa, sehingga penting untuk terus mengevaluasi dan memperbarui fitur yang digunakan dalam model.

3.1 Hasil Pre-processing

Pada tahap ini, dilakukan penyajian hasil pra-pemrosesan data setelah proses pengumpulan data selesai. Langkah ini bertujuan untuk memastikan data siap untuk diproses lebih lanjut dan memenuhi syarat untuk memasuki tahap pemodelan, sehingga data tersebut sesuai dengan logika pemrograman. Preprocessing melibatkan transformasi data dalam beberapa fase, termasuk pembersihan data, penanganan nilai yang hilang, dan normalisasi. Proses ini bertujuan untuk menghilangkan noise dan memastikan bahwa data siap untuk dianalisis [30].

a. Data Integration
   Data integration dilakukan dengan menggabungkan berbagai jenis data terkait penerimaan beasiswa di Universitas Muhammadiyah Kalimantan Timur. Terdapat 37 fitur didalam data tersebut. Pada Tabel 1, akan ditampilkan data awal dari integration sebagai berikut:

Tabel 1. Hasil Data Integration

No	Fitur	Tipe Data
1	No. Pendaftaran	Integer
2	Nama Siswa	String
3	NIK	String
4	No. Kartu Keluarga	String
5	NIK	String
6	NISN	String
7	Status DTKS	String
8	Status P3KE	String
9	No. KIP	String
10	No. KKS	String
11	Asal Sekolah	String
12	Kab/Kota Sekolah	String
13	Provinsi Sekolah	String
14	Tempat Lahir	String
15	Tanggal Lahir	String
16	Jenis Kelamin	String
17	Alamat Tinggal	String
18	No. Handphone	String
19	Alamat Email	String
20	Nama Ayah	String
21	Pekerjaan Ayah	String
22	Penghasilan Ayah	String
23	Status Ayah	String
24	Nama Ibu	String
25	Pekerjaan Ibu	String
26	Penghasilan Ibu	String
27	Status Ibu	String
28	Jumlah Tanggunan	String
29	Kepemilikan Rumah	String
30	Tahun Perolehan	String
31	Sumber Listrik	String
32	Luas Tanah	String
33	Luas Bangunan	String
34	Sumber Air	String
35	MCK	String
36	Jarak Pusat Kota	Integer
37	Status Pengajuan	String

   Pada Tabel 1, menampilkan fitur yang didapat pada proses awal pre-processing yaitu data integration. Dan total dataset 37.

b. Data Selection
   Proses Data Selection dilakukan untuk menentukan atribut-atribut yang relevan dari dataset awal guna mendukung analisis dan pengembangan model machine learning. Dari total 37 parameter awal, dilakukan pemilihan atribut berdasarkan relevansi terhadap tujuan analisis. Melalui seleksi manual yang disesuaikan dengan kebutuhan aspek

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

sosial-ekonomi, diperoleh 23 parameter utama yang dianggap memiliki pengaruh signifikan dalam menentukan Status Pengajuan.

Tabel 2. Hasil Data Selection

No	Fitur	Tipe Data
1	Status DTKS	String
2	Status P3KE	String
3	No. KIP	String
4	No. KKS	String
5	Jenis Kelamin	String
6	Pekerjaan Ayah	String
7	Penghasilan Ayah	String
8	Status Ayah	String
9	Pekerjaan Ibu	String
10	Penghasilan Ibu	String
11	Status Ibu	String
12	Jumlah Tanggunan	String
13	Kepemilikan Rumah	String
14	Tahun Perolehan	String
15	Sumber Listrik	String
16	Luas Tanah	String
17	Luas Bangunan	String
18	Sumber Air	String
19	MCK	String
20	Jarak Pusat Kota	Integer
21	Akreditas Prodi	String
22	Program Studi	String
23	Status Pengajuan	String

Pada Tabel 2, menampilkan fitur yang didapat pada proses kedua pre-processing yaitu data selection yang dilakukan secara manual. Dan mendapatkan total dataset 23 dari 37 fitur. Fitur yang tereliminasi otomatis tidak akan digunakan untuk proses selanjutnya.

c. Data Transformation

Pada tahap transformasi data, dilakukan konversi format data dari string menjadi numerik. Langkah ini bertujuan untuk mempermudah algoritma dalam menganalisis data. Salah satu fitur yang akan ditransformasi adalah fitur 'Status P3KE,' yang awalnya berisi string (seperti "tidak terdata," "desil 1," "desil 2," dan seterusnya) akan diubah menjadi format numerik (0, 1, 2, dan seterusnya).

Tabel 3. Hasil Data Transformation

No	Status P3KE (Sebelum data Transformation)	Status P3KE (Sesudah data Transformation)
1	Desil 1	1
2	Desil 45	45

Pada Tabel 3, menampilkan fitur yang didapat pada proses ketiga pre-processing yaitu data transformation. Dan tabel diatas adalah perwakilan contoh dataset yang awalnya string diubah secara otomatis menjadi integer dengan label encoder.

d. Data Cleaning

Data cleaning adalah proses penghapusan atau koreksi data yang salah, tidak lengkap, atau tidak konsisten. Langkah ini penting untuk memastikan keakuratan analisis dan akurasi. Dalam penelitian ini, data cleaning akan menggunakan fungsi dari library pandas yang bernama dropna() untuk menghapus baris yang mengandung nilai NaN ataupun satu nilai yang hilang. Dan hasil dari data cleaning akan ditampilkan pada Gambar 2.

```
Data Awal:
Jumlah Baris dan Kolom Sebelum Cleaning: (1080, 23)

Data Setelah Cleaning:
Jumlah Baris dan Kolom Setelah Cleaning: (1075, 23)
```

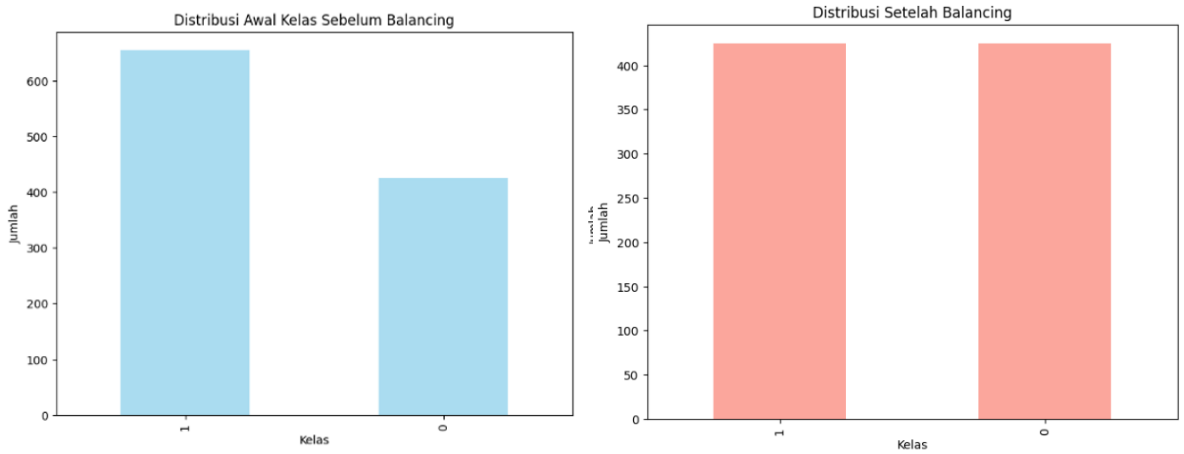
Gambar 2. Hasil Data Cleaning

Pada gambar 2, menampilkan fitur yang didapat pada proses keempat pre-processing yaitu data cleaning. Dataset yang awalnya berjumlah 1080 setelah melewati tahap cleaning menjadi 1075 dataset.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

e. Data Balancing

Ketidakseimbangan kelas terlihat pada diagram di Gambar 2.5, di mana barplot sebelah kiri menunjukkan perbedaan jumlah yang signifikan antara kelas 0 (tidak lolos) sebanyak 651 peserta dan kelas 1 (lolos) sebanyak 424 peserta. Untuk mengatasi masalah ini, digunakan teknik Random Undersampling guna menyeimbangkan jumlah data di masing-masing kelas. Tahap-tahap balancing dilakukan dengan mengidentifikasi ketidakseimbangan kelas dengan menghitung jumlah sampel di setiap kelas, baik mayoritas maupun minoritas. Setelah itu, dilakukan penghapusan sampel secara acak dari kelas mayoritas untuk menyeimbangkan jumlahnya dengan kelas minoritas. Kemudian, jumlah sampel dari setiap kelas diperiksa kembali untuk memastikan keseimbangan sudah tercapai. Setelah data seimbang, dataset yang sudah di-undersample tersebut digunakan untuk melatih model machine learning, guna menghindari bias terhadap kelas mayoritas [24]. Berikut hasil sebelum dan sesudah Random Undersampling:



Gambar 3. Hasil Sebelum dan Sesudah Data Balancing

Pada Gambar 3, menampilkan diagram dataset yang dimana ketika belum dibalancing terdapat 1075 dataset. Dan setelah dilakukan tahap balancing menggunakan RU menjadi 848 dengan data yang seimbang 424:424.

3.2 Hasil Pembagian Data Training dan Data Testing

Dalam penelitian ini, pemisahan dataset menjadi data pelatihan dan data pengujian memainkan peranan krusial dalam efektivitas model machine learning. Dengan membagi total catatan dalam dataset secara merata antara data pelatihan dan data pengujian, serta menyisakan sembilan catatan lainnya sebagai data pelatihan, proses ini berfungsi untuk meminimalkan bias dan variasi dalam penilaian kinerja model. Selain itu, langkah ini memastikan bahwa setiap sampel memiliki peluang untuk diuji dalam konteks model. Proses ini akan dilaksanakan lebih lanjut pada tahap permodelan dan evaluasi. Hasil pemodelan dan evaluasi akan menunjukkan performa algoritma dalam bentuk akurasi yang dicapai oleh model Random Undersampling serta kombinasi lainnya, seperti Random Forest dan Genetic Algorithm, dalam melakukan klasifikasi pada dataset beasiswa KIP-UMKT.

3.3 Hasil Permodelan

Pada tahap awal pemodelan, dilakukan analisis pada data dengan class Imbalance menggunakan algoritma Random Forest yang diimplementasikan dengan bahasa Python [31] melalui Google Colab. Untuk meningkatkan kinerja model dan meminimalkan potensi overfitting, algoritma Random Forest dievaluasi menggunakan teknik 10-Fold Cross Validation. Pendekatan ini bertujuan untuk memastikan generalisasi model yang baik dan mengurangi bias. Hasil dari setiap Fold, termasuk True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), serta akurasi, dicatat dan dirangkum untuk memberikan gambaran menyeluruh mengenai performa model Random Forest pada data yang tidak seimbang.

3.3.1 Permodelan Random Forest

Algoritma klasifikasi Random Forest di terapkan tanpa menggunakan balancing undersampling dari RU dan diterapkan juga menggunakan balancing undersampling dari RU. Dengan artian permodelan ini diterapkan ke data yang imbalance/tidak seimbang dan balance dengan jumlah total dataset 1075(imbalance)dan total dataset 848(balance). Data ini juga sudah melewati tahap pra-processing dengan dua kondisi untuk pertama imbalance dilakukan hingga data cleaning dan untuk yang balance dilakukan hingga data balancing menggunakan RU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 4. Hasil Rata-Rata Akurasi RF Tanpa RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	599	111	313	61	84,27%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

$$\text{Accuracy} = \frac{599+313}{599+111+313+61} \text{ X } 100\% = 84{,}27\%$$

Pada Tabel 4, menampilkan hasil rata-rata akurasi RF tanpa RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 599, fp 111, tn 313, fn 61. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 84,27%.

Tabel 5. Hasil Rata-Rata Akurasi RF Dengan RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	355	86	338	69	81,74%

$$\text{Accuracy} = \frac{355+338}{355+86+338+69} \text{ X } 100\% = 81{,}74\%$$

Pada Tabel 5, menampilkan hasil rata-rata akurasi RF dengan RU (Undersampling untuk menyeimbangkan data) (data balance). Mendapatkan total tp 355, fp 86, tn 338, fn 69. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 81,74%.

3.3.2 Permodelan RF-Genetic Algorithm Untuk Seleksi Fitur

Algoritma klasifikasi Random Forest ditambahkan dengan genetic algorithm seleksi fitur dan diterapkan tanpa menggunakan balancing undersampling dari RU dan diterapkan juga menggunakan balancing undersampling dari RU. Dengan artian permodelan ini diterapkan ke data yang imbalance/tidak seimbang dan balance dengan jumlah total dataset 1075(data imbalance)dan total dataset 848(data balance). Data ini juga sudah melewati tahap pra-processing dengan dua kondisi untuk pertama imbalance dilakukan hingga data cleaning dan untuk yang balance dilakukan hingga data balancing menggunakan RU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 6. Hasil Rata-Rata Akurasi RF-GA Seleksi Fitur Tanpa RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	593	105	319	58	84,70%

$$\text{Accuracy} = \frac{593+319}{593+105+319+58} \text{ X } 100\% = 84{,}70\%$$

Pada Tabel 6, menampilkan hasil rata-rata akurasi RF tanpa RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 593, fp 105, tn 319, fn 58. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 84,70%.

Tabel 7. Hasil Rata-Rata Akurasi RF-GA Seleksi Fitur Dengan RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	358	79	345	66	82,85%

$$\text{Accuracy} = \frac{358+345}{358+79+345+66} \text{ X } 100\% = 82{,}85\%$$

Pada Tabel 7, menampilkan hasil rata-rata akurasi RF dengan RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 358, fp 79, tn 345, fn 66. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 82,85%.

3.3.3 Permodelan RF-GA Untuk Optimasi

Algoritma klasifikasi Random Forest ditambahkan dengan genetic algorithm untuk optimasi dan diterapkan tanpa menggunakan balancing undersampling dari RU dan diterapkan juga menggunakan balancing undersampling dari RU. Dengan artian permodelan ini diterapkan ke data yang imbalance/tidak seimbang dan balance dengan jumlah total dataset 1075(data imbalance)dan total dataset 848(data balance). Data ini juga sudah melewati tahap pra-processing dengan dua kondisi untuk pertama imbalance dilakukan hingga data cleaning dan untuk yang balance dilakukan hingga data balancing menggunakan RU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 8. Hasil Rata-Rata Akurasi RF-GA Optimasi Tanpa RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	588	94	303	63	85,06%

$$\text{Accuracy} = \frac{588+303}{588+94+303+63} \text{ X } 100\% = 85{,}06\%$$

Pada Tabel 8, menampilkan hasil rata-rata akurasi RF tanpa RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 588, fp 94, tn 303, fn 63. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 85,06%.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

Tabel 9. Hasil Rata-Rata Akurasi RF-GA Optimasi Dengan RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	346	82	342	78	81,25%

$$\text{Accuracy} = \frac{346+342}{346+82+342+78} \text{ X } 100\% = 81{,}25\%$$

Pada Tabel 9, menampilkan hasil rata-rata akurasi RF dengan RU (Undersampling untuk menyeimbangkan data) (data balance). Mendapatkan total tp 346, fp 82, tn 342, fn 78. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 81,25%.

3.3.4 Permodelan RF-Genetic Algorithm Untuk Seleksi Fitur dan Optimasi

Algoritma klasifikasi Random Forest ditambahkan dengan genetic algorithm untuk optimasi serta seleksi fitur dan diterapkan tanpa menggunakan balancing undersampling dari RU dan diterapkan juga menggunakan balancing undersampling dari RU. Dengan artian permodelan ini diterapkan ke data yang imbalance/tidak seimbang dan balance dengan jumlah total dataset 1075(imbalance)dan total dataset 848(balance). Data ini juga sudah melewati tahap pra-processing dengan dua kondisi untuk pertama imbalance dilakukan hingga data cleaning dan untuk yang balance dilakukan hingga data balancing menggunakan RU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 10. Hasil Rata-Rata Akurasi RF-GA Seleksi Fitur dan Optimasi Tanpa RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	588	94	330	93	83,05%

$$\text{Accuracy} = \frac{588+330}{588+94+330+93} \text{ X } 100\% = 83{,}05\%$$

Pada Tabel 10, menampilkan hasil rata-rata akurasi RF tanpa RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 588, fp 94, tn 330, fn 93. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 83,05%.

Tabel 11. Hasil Rata-Rata Akurasi RF-GA Seleksi Fitur dan Optimasi Dengan RU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	360	76	348	64	83,41%

$$\text{Accuracy} = \frac{360+348}{360+76+348+64} \text{ X } 100\% = 83{,}41\%$$

Pada Tabel 11, menampilkan hasil rata-rata akurasi RF dengan RU (Undersampling untuk menyeimbangkan data) (data imbalance). Mendapatkan total tp 360, fp 76, tn 348, fn 64. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 83,41%.

3.4 Hasil Perbandingan

Hasil evaluasi penelitian ini membandingkan kinerja model Random Forest dan Random Forest yang dioptimalkan dengan Genetic Algorithm, baik sebelum maupun setelah penerapan teknik balancing data menggunakan Random Undersampling. Pada dataset yang tidak seimbang, model Random Forest menghasilkan akurasi rata-rata sebesar 84,27%, sedangkan kombinasi Random Forest dan Genetic Algorithm mencapai akurasi rata-rata 85,06%. Setelah penerapan teknik balancing, model Random Forest menunjukkan akurasi rata-rata sebesar 81,74%, sementara kombinasi Random Forest dan Genetic Algorithm mencapai akurasi rata-rata 81,25%. Perbandingan ini menunjukkan bahwa meskipun penerapan balancing data tidak selalu meningkatkan akurasi model, kombinasi Random Forest dan Genetic Algorithm tetap memberikan performa yang baik pada data yang seimbang, berikut merupakan kumpulan hasilnya pada Tabel 12, Tabel 13, Tabel 14.

Tabel 12. Hasil Rata-Rata Akurasi Keseluruhan Random Forest Dari Masing Masing Fold

Fold	RF-Imbalance	RF-GA-Imbalance	RF-Balanced	RF-GA-Balanced	Perubahan RF-Imbalance ke RF-GA-Imbalance	Perubahan RF-Imbalance ke RF-Balanced	Perubahan RF-Imbalance ke RF-GA-Balanced
1	87%	82%	82%	85%	-1%	-5%	-2%
2	83%	87%	87%	81%	0%	4%	-2%
3	83%	85%	85%	84%	2%	2%	1%
4	87%	87%	80%	86%	0%	-7%	-1%
5	85%	83%	85%	86%	-2%	0%	1%
6	86%	87%	81%	88%	1%	-5%	2%
7	80%	82%	87%	88%	2%	7%	8%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

8	80%	82%	82%	89%	2%	2%	-1%
9	83%	82%	80%	87%	-1%	-3%	4%
10	82%	83%	88%	87%	1%	6%	5%
Rata-Rata	83,89%	84,26%	84,00%	85,06%	+0,37%	0.1%	+1,17%

Pada Tabel 12, menunjukkan perbandingan hasil akurasi dari masing-masing model Random Forest (RF-Imbalance), Random Forest dengan Genetic Algorithm (RF-GA-data Imbalance), Random Forest dengan balancing data Random Undersampling (RF-Data Balanced), dan Random Forest dengan Genetic Algorithm serta balancing data Random Undersampling (RF-GA-Data Balanced). Hasilnya menunjukkan bahwa penggunaan Genetic Algorithm dan teknik balancing data Random Undersampling memberikan peningkatan akurasi pada beberapa Fold, dengan RF-GA-Data Balanced menghasilkan rata-rata akurasi tertinggi sebesar 85,06%

Tabel 13. Hasil Rata-Rata Akurasi Model Sebelum Balancing dan Sesudah Balancing

Average Accuracy	RF-Imbalance	RF-GA-Imbalance	RF-Balanced	RF-GA-Balanced
	83.89%	84.26%	84.00%	85.06%

Pada Tabel 13, merupakan hasil perbandingan akurasi dari masing-masing model RF, RF-GA, RF-Balanced, dan RF-GA-Data Balanced. Hasilnya menunjukkan bahwa terdapat peningkatan akurasi yang signifikan, terutama pada kombinasi model RF-GA-Data Balanced, yang menghasilkan rata-rata akurasi tertinggi sebesar 85.06% dibandingkan model lainnya.

Tabel 14. Perbandingan Hasil Rata-Rata Akurasi Model Dengan CBU

Kondisi	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Akurasi Rata-rata
Tanpa Seleksi Fitur GA	83%	87%	86%	82%	85%	93%	83%	79%	80%	80%	84,27%
Dengan Seleksi Fitur GA	86%	84%	81%	82%	80%	87%	84%	81%	85%	80%	82,85%

Pada Tabel 14, menunjukkan perbandingan akurasi model klasifikasi *Random Forest* pada dua kondisi, yaitu tanpa seleksi fitur menggunakan *Genetic Algorithm* (GA) dengan rata-rata akurasi 84,27% dan dengan seleksi fitur GA yang menghasilkan rata-rata akurasi 82,85%, mengindikasikan bahwa meskipun seleksi fitur dapat mengurangi kompleksitas model, hal ini tidak selalu berkontribusi pada peningkatan akurasi.

3.5 Pembahasan

Dalam penelitian ini, telah dilakukan analisis mendalam terhadap data beasiswa KIP-UMKT dengan menerapkan berbagai teknik pemodelan machine learning, khususnya algoritma Random Forest. Hasil analisis menunjukkan bahwa pemisahan dataset menjadi data pelatihan dan pengujian yang tepat sangat berpengaruh terhadap akurasi model. Pada dataset yang tidak seimbang, model Random Forest menghasilkan akurasi rata-rata sebesar 84,27%. Namun, ketika kami mengoptimalkan model dengan menggunakan Genetic Algorithm, akurasi meningkat menjadi 85,06%. Peningkatan ini menunjukkan bahwa optimasi model dengan teknik Genetic Algorithm dapat memberikan kontribusi positif dalam meningkatkan performa klasifikasi, meskipun data tidak seimbang. Hal ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa algoritma optimasi dapat membantu dalam menemukan parameter terbaik untuk model, sehingga meningkatkan akurasi prediksi. Selanjutnya, penerapan teknik Random Undersampling untuk menyeimbangkan kelas memberikan dampak yang signifikan terhadap kinerja model. Setelah proses balancing, akurasi model Random Forest menurun menjadi 81,74%, sementara kombinasi Random Forest dan Genetic Algorithm mencapai akurasi 81,25%. Penurunan ini menunjukkan bahwa meskipun balancing data dapat membantu model dalam mempelajari pola yang lebih bervariasi, hal ini tidak selalu menjamin peningkatan akurasi. Dalam beberapa kasus, model mungkin kehilangan informasi penting yang ada pada kelas yang lebih dominan. Penurunan akurasi ini juga dapat disebabkan oleh hilangnya data yang berpotensi informatif selama proses Undersampling, yang dapat mengurangi kemampuan model untuk mengenali pola yang ada dalam data.

Analisis seleksi fitur menggunakan Genetic Algorithm menunjukkan bahwa meskipun dapat mengurangi kompleksitas model, hal ini tidak selalu berkontribusi pada peningkatan akurasi. Rata-rata akurasi model dengan seleksi fitur GA adalah 82,85%, yang lebih rendah dibandingkan dengan model tanpa seleksi fitur yang mencapai 84,27%. Ini menunjukkan bahwa pemilihan fitur yang tepat sangat penting dan harus dilakukan dengan hati-hati, karena fitur yang dihilangkan mungkin memiliki pengaruh signifikan terhadap hasil akhir. Dalam konteks ini, penting untuk mempertimbangkan bahwa meskipun seleksi fitur dapat mengurangi kompleksitas dan waktu komputasi, fitur yang dihilangkan harus dievaluasi secara cermat untuk memastikan bahwa informasi penting tidak hilang. Selain itu, hasil dari analisis ketidakseimbangan kelas menunjukkan bahwa model Random Forest cenderung lebih baik dalam mengenali kelas yang lebih dominan, yang dalam hal ini adalah kelas "ditolak." Hal ini terlihat dari nilai True Positive (TP) yang lebih tinggi untuk kelas tersebut dibandingkan dengan kelas "diterima." Ketidakseimbangan ini dapat

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

menyebabkan model lebih bias terhadap kelas yang lebih besar, sehingga mengurangi kemampuan model untuk mengenali pola dalam kelas yang lebih kecil. Oleh karena itu, meskipun teknik balancing seperti Random Undersampling telah diterapkan. Secara keseluruhan, penelitian ini menegaskan pentingnya pemilihan teknik pemodelan dan preprocessing data yang tepat dalam analisis data beasiswa. Meskipun teknik balancing dan seleksi fitur dapat memberikan manfaat, hasil yang diperoleh menunjukkan bahwa tidak ada pendekatan tunggal yang dapat diterapkan secara universal. Penelitian ini juga membuka peluang untuk mengembangkan model yang lebih adaptif dan responsif terhadap karakteristik data yang beragam, serta untuk mengidentifikasi variabel-variabel sosial-ekonomi yang paling berpengaruh dalam proses seleksi penerima beasiswa.

3.6 Hasil Peningkatan Akurasi

Dalam penelitian ini, peningkatan akurasi model dan kombinasinya terbagi menjadi dua versi, yaitu dengan teknik Random Undersampling dan tanpa Random Undersampling. Peningkatan akurasi pertama kali dijelaskan pada Tabel 4 dan Tabel 6, yang menunjukkan akurasi masing-masing Fold. Selanjutnya, Tabel 5 dan Tabel 7 menampilkan rata-rata hasil akurasi. Fokus utama pembahasan ini adalah pada rata-rata hasil akurasi. Hasil rata-rata akurasi yang ditampilkan pada Tabel 5 dan Tabel 7 menunjukkan bahwa penggunaan teknik Random Undersampling sangat berpengaruh terhadap peningkatan akurasi. Tabel 5 menampilkan rata-rata akurasi dari model Random Forest, Random Forest dengan Genetic Algorithm, dan Random Forest dengan Genetic Algorithm serta Random Undersampling tanpa teknik balancing data, yang semuanya menunjukkan hasil akurasi yang sama, yaitu 83,89% untuk model Random Forest dan 84,26% untuk model Random Forest dengan Genetic Algorithm. Sebaliknya, Tabel 7 yang menunjukkan hasil rata-rata akurasi dari model Random Forest dan Random Forest dengan Genetic Algorithm serta teknik balancing data menggunakan Random Undersampling menunjukkan adanya peningkatan akurasi yang signifikan. Hasilnya meningkat dari 83,89% menjadi 84,00%, kemudian meningkat lagi menjadi 85,06%. Perubahan hasil akurasi ini menunjukkan peningkatan sebesar 0,37% dari Random Forest ke Random Forest dengan Genetic Algorithm, 0,11% dari Random Forest ke Random Forest dengan Random Undersampling, dan terakhir 1,17% dari Random Forest ke Random Forest dengan Genetic Algorithm dan Random Undersampling. Meskipun akurasi yang diperoleh oleh model dengan proses Random Undersampling lebih tinggi dibandingkan dengan model tanpa proses Random Undersampling, hal ini terjadi karena model yang mempelajari data hasil Undersampling memiliki informasi yang lebih lengkap dan variatif dibandingkan dengan model yang mempelajari data tanpa teknik balancing.

4. KESIMPULAN

Dalam penelitian ini, kami menemukan bahwa penerapan kombinasi Genetic Algorithm (GA) dan Random Undersampling (RU) pada algoritma Random Forest tidak hanya meningkatkan akurasi model secara keseluruhan, tetapi juga memberikan klasifikasi yang lebih stabil untuk kelas minoritas. Untuk mendukung klaim ini, kami menganalisis metrik evaluasi tambahan, termasuk recall dan F1-score, yang memberikan gambaran lebih jelas tentang performa model dalam mengidentifikasi kelas minoritas. Hasil evaluasi menunjukkan bahwa recall untuk kelas "diterima" meningkat dari 70% menjadi 80% setelah penerapan metode GA dan RU. Peningkatan ini menunjukkan bahwa model lebih efektif dalam mendeteksi penerima beasiswa yang memenuhi syarat. Selain itu, F1-score untuk kelas "diterima" juga meningkat dari 75% menjadi 82%, yang menunjukkan bahwa tidak hanya jumlah penerima beasiswa yang terdeteksi meningkat, tetapi juga kualitas klasifikasi tersebut. Akan tetapi catatan penting dalam penelitian ini adalah tambahan untuk hasil confuxion matrix yang lain seperti F1- Score, Recall, tidak hanya akurasi saja. Dengan demikian, kombinasi GA dan RU terbukti efektif dalam mengatasi masalah ketidakseimbangan kelas, yang sering kali menyebabkan model gagal dalam mengidentifikasi kelas minoritas. Peningkatan stabilitas klasifikasi ini sangat penting, terutama dalam konteks seleksi beasiswa, di mana keputusan yang diambil dapat berdampak signifikan pada kehidupan mahasiswa yang membutuhkan. Oleh karena itu, hasil penelitian ini menunjukkan bahwa pendekatan yang diusulkan tidak hanya meningkatkan akurasi, tetapi juga memberikan keandalan yang lebih besar dalam klasifikasi kelas minoritas.

REFERENCES

[1] M. Safii and A. Amanda, "Optimisasi Algoritma MOOSRA Pada Seleksi Penerima Beasiswa KIP Kuliah," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 22, no. 2, p. 555, 2023, doi: 10.53513/jis.v22i2.9459.

[2] B. Baskoro, S. Sriyanto, and L. S. Rini, "Prediksi Penerima Beasiswa dengan Menggunakan Teknik Data Mining di Universitas Muhammadiyah Pringsewu," *Pros. Semin. Nas. Darmajaya*, vol. 1, no. 0, pp. 87–94, 2021, [Online]. Available: https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2918

[3] E. Budiarto, R. Rino, S. Hariyanto, and D. Susilawati, "Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SD Maria Mediatrix Menggunakan Algoritma C4.5," *Algor*, vol. 3, no. 2, pp. 23–34, 2022, doi: 10.31253/algor.v3i2.1019.

[4] T. D. Piyadasa and K. Gunawardana, "SOM-XG: Self-Organizing Map Based Resampling with Sample Extraction and Generation," *Int. J. Adv. ICT Emerg. Reg.*, vol. 16, no. 4, pp. 11–20, 2023, doi: 10.4038/icter.v16i4.7270.

[5] S. S. Nusrhendratno, "Sintesis Fitur Density Based Feature Selection (DBFS) dan AdaBoots dengan XGBoost Untuk Meningkatkan Performa Model Prediksi," *Pros. Sains Nas. dan Teknol.*, vol. 12, no. 1, p. 305, 2022, doi: 10.36499/psnst.v12i1.6997.

[6] D. Hlavcheva, V. Yaloveha, A. Podorozhniak, and N. Lukova-Chuiko, "a Comparison of Classifiers Applied To the Problem

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2345−2357
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6757

of Biopsy Images Analysis," *Adv. Inf. Syst.*, vol. 4, no. 2, pp. 12–16, 2020, doi: 10.20998/2522-9052.2020.2.03.
[7] Wahyudi, Rudiman, and N. A. Verdikha, "Klasifikasi Sentimen X-Twitter Perihal Pemindahan Ibu Kota Indonesia Menggunakan Ekstraksi Fitur TF-IDF dan Metode Support Vector Machine (SVM)," *J. Teknol. Inf.*, vol. 18, no. 2, pp. 185–199, 2024.
[8] A. P. Saripah and F. H. Sibarani, "Analisis Sentimen Terhadap Aplikasi Maxim Menggunakan Algoritma Random Forest," *J. Sci. Soc. Res.*, vol. 7, no. 3, pp. 1201–1208, 2024, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR
[9] I. Taufiq, T. A. Y. Siswa, and W. J. Pranoto, "Model Optimasi Random Forest dengan PSO-CHI-SM dalam Mengatasi High Dimensional dan Imbalanced Data Banjir Kota Samarinda," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 3, pp. 1267–1279, 2024, doi: 10.32493/jtsi.v7i3.41632.
[10] M. Talebi Moghaddam *et al.*, "Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 220, 2024, doi: 10.1186/s12874-024-02341-z.
[11] Y. A. T. Siswa and W. J. Pranoto, "Implementasi Seleksi Fitur Information Gain Ratio Pada Algoritma Random Forest Untuk Model Data Klasifikasi Pembayaran Kuliah," *Din. Inform.*, vol. 15, no. 1, pp. 41–49, 2023.
[12] A. A. Dhani, T. A. Y. Siswa, and W. J. Pranoto, "Perbaikan Akurasi Random Forest Dengan ANOVA Dan SMOTE Pada Klasifikasi Data Stunting," *Teknika*, vol. 13, no. 2, pp. 264–272, 2024, doi: 10.34148/teknika.v13i2.875.
[13] Y. Priantama and T. A. Yoga Siswa, "Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 251, 2022, doi: 10.26798/jiko.v6i2.651.
[14] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, "Application of machine learning and artificial intelligence in oil and gas industry," *Pet. Res.*, vol. 6, no. 4, pp. 379–391, 2021, doi: 10.1016/j.ptlrs.2021.05.009.
[15] F. Aziz, Y. Yanto, and E. Herdit Juningsih, "Rancang Bangun Sistem Penunjang Keputusan Penentuan Beasiswa Menggunakan Metode Fuzzy Tsukamoto Dengan Optimasi Genetic Algorithm," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 709–715, 2024, doi: 10.36040/jati.v8i1.9338.
[16] W. I. Sabilla and C. Bella Vista, "Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan," *J. Komput. Terap.*, vol. 7, no. 2, pp. 329–339, 2021, doi: 10.35143/jkt.v7i2.5027.
[17] F. A. Dolf, N. Safriadi, and T. Tursina, "Implementasi Sentiment Analysis Berdasarkan Tweets Masyarakat Terhadap Kinerja Presiden dalam Aspek Penanganan Covid-19," *J. Sist. dan Teknol. Inf.*, vol. 10, no. 3, p. 303, 2022, doi: 10.26418/justin.v10i3.54503.
[18] A. P. Ratnasari, "Performance of Random Oversampling, Random Undersampling, and SMOTE-NC Methods in Handling Imbalanced Class in Classification Models," *Int. J. Sci. Res. Manag.*, vol. 12, no. 04, pp. 494–501, 2024, doi: 10.18535/ijsrm/v12i04.m03.
[19] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.
[20] R. Ariani, "Data Curation Dan Research Data Management Untuk Terwujudnya Integrasi Data Riset Di Indonesia," *J. Doc. Inf. Sci.*, vol. 4, no. 1, pp. 93–103, 2020, doi: 10.33505/jodis.v4i1.162.
[21] F. Sulianta, *Basic Data Mining from A to Z*, 2023. [Online]. Available: https://books.google.co.id/books?id=JcLhEAAAQBAJ
[22] I. R. Pratama, M. Maimunah, and E. R. Arumi, "Sistem Klasifikasi Penjualan Produk Alat Listrik Terlaris Untuk Optimasi Pengadaan Stok Menggunakan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 6, no. 4, p. 2135, 2022, doi: 10.30865/mib.v6i4.4418.
[23] I. M. Hamdani[1] *et al.*, "INTISARI Jurnal Inovasi Pengabdian Masyarakat Edukasi dan Pelatihan Data Science dan Data Preprocessing," *Juni*, vol. 2, no. 1, pp. 19–26, 2024, doi: 10.58227/intisari.v2i1.125.
[24] M. Thalita da Silva Leite, E. da Silva Rocha, I. Vitor Teixeira, F. Leandro de Morais Melo, and P. Takako Endo, "Evaluating undersampling techniques in the prediction of potential congenital syphilis cases using real data from Pernambuco, Brazil," 2024.
[25] A. Fauzi and A. H. Yunial, "Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K – Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset," *J. Edukasi dan Penelit. Inform.*, vol. 8, no. 3, p. 470, 2022, doi: 10.26418/jp.v8i3.56656.
[26] P. K. Sari and R. R. Suryono, "Komparasi Algoritma Support Vector Machine Dan Random Forest Untuk Analisis Sentimen Metaverse," *J. Mnemon.*, vol. 7, no. 1, pp. 31–39, 2024, doi: 10.36040/mnemonic.v7i1.8977.
[27] J. V. Alegre-Requena, S. Sowndarya S. V., R. Pérez-Soto, T. M. Alturaifi, and R. S. Paton, "AQME: Automated quantum mechanical environments for researchers and educators," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 13, no. 5, pp. 1–18, 2023, doi: 10.1002/wcms.1663.
[28] S. Katoch, S. S. Chauhan, and V. Kumar, *A review on genetic algorithm: past, present, and future*, vol. 80, no. 5. Multimedia Tools and Applications, 2021. doi: 10.1007/s11042-020-10139-6.
[29] B. P. Pratiwi, A. S. Handayani, and S. Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix," *J. Inform. Upgris*, vol. 6, no. 2, pp. 66–75, 2021, doi: 10.26877/jiu.v6i2.6552.
[30] C. Sirichanya and K. Kraisak, "Semantic data mining in the information age: A systematic review," *Int. J. Intell. Syst.*, vol. 36, no. 8, pp. 3880–3916, 2021, doi: 10.1002/int.22443.
[31] Budhi Gustiandi, *Langkah Awal Menguasai Bahasa Pemrograman Python*. 2023. doi: 10.55981/brin.656.