Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

Penerapan Metode GA-CBU Pada Algoritma Logistic Regression Untuk Mengatasi Class Imbalance Data Beasiswa KIP-Kuliah

Ahmad Nugraha Poernamawan, Taghfirul Azhima Yoga Siswa*, Rudiman

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
Email: [1]2011102441178@umkt.ac.id, [2,*]tay758@umkt.ac.id, [3]rud959@umkt.ac.id,
Email Penulis Korespondensi: tay758@umkt.ac.id
Submitted: **15/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Masalah ketidakseimbangan kelas (class imbalance) sering menjadi tantangan dalam analisis data, di mana jumlah data pada kelas mayoritas jauh lebih banyak dibandingkan dengan kelas minoritas. Hal ini dapat menyebabkan model klasifikasi lebih cenderung memprediksi kelas mayoritas, sehingga akurasi dalam mengidentifikasi kelas minoritas menjadi rendah. Penelitian ini bertujuan untuk mengimplementasikan algoritma Logistic Regression (LR) yang dikombinasikan dengan metode Clustering Based Undersampling (CBU) sebagai teknik *undersampling*, seleksi fitur, serta optimasi menggunakan Genetic Algorithm (GA) dalam mengklasifikasi data beasiswa KIP-Kuliah di Universitas Muhammadiyah Kalimantan Timur. Selain itu, penelitian ini juga mengevaluasi kinerja model dengan teknik *10-Fold Cross Validation* dan *Confusion Matrix* sebagai metrik akurasi dan bertujuan untuk mengatasi masalah ketidakseimbangan kelas dalam data penerima beasiswa (KIP) di Universitas Muhammadiyah Kalimantan Timur. Data yang digunakan terdiri dari 1075 record dengan 37 fitur yang berkaitan dengan faktor sosial-ekonomi penerima beasiswa. Hasil dari penerapan metode CBU menunjukkan peningkatan akurasi model Logistic Regression dari 62.51% menjadi 67.68%. Selain itu, kombinasi antara GA dan CBU terbukti efektif dalam meningkatkan performa model, memberikan hasil yang lebih stabil dalam mengklasifikasikan kelas minoritas. Diharapkan, penelitian ini dapat memberikan kontribusi yang berarti dalam pengembangan sistem seleksi penerimaan beasiswa yang lebih luas dan efisien, serta menjadi referensi bagi penelitian selanjutnya di bidang data mining dan machine learning.

Kata Kunci: Klasifikasi; Class Imbalance; LR; GA; CBU

Abstract−The issue of class imbalance often poses a challenge in data analysis, where the number of instances in the majority class is significantly higher than that in the minority class. This can lead classification models to be biased towards predicting the majority class, resulting in low accuracy in identifying the minority class. This research aims to implement the Logistic Regression (LR) algorithm combined with the Clustering Based Undersampling (CBU) method as an undersampling technique, feature selection, and optimization using Genetic Algorithm (GA) in classifying KIP-College scholarship data at Muhammadiyah University of East Kalimantan. In addition, this research also evaluates the performance of the model with 10-Fold Cross Validation and Confusion Matrix techniques as accuracy metrics and aims to overcome the problem of class imbalance in the data of scholarship recipients (KIP) at Muhammadiyah University of East Kalimantan. The data used consists of 1075 records with 37 features related to the socio-economic factors of scholarship recipients. The results from the application of the CBU method indicate an increase in the accuracy of the Logistic Regression model from 62.51% to 67.68%. Furthermore, the combination of GA and CBU has providing more stable results in classifying minority classes. It is hoped that this research can make a significant contribution to the development of a more accurate and efficient scholarship recipient selection system, as well as serve as a reference for future studies in the fields of data mining and machine learning.

Keywords: Classification; Class Imbalance; LR; GA; CBU

1. PENDAHULUAN

Pemerintah Indonesia menghadirkan berbagai program bantuan pendidikan, salah satunya Kartu Indonesia Pintar (KIP), yang bertujuan memberikan dukungan pendidikan kepada keluarga kurang mampu. Program ini memainkan peran penting dalam memperluas akses dan pemerataan pendidikan di Indonesia [1]. Jumlah pendaftar yang tinggi menciptakan tantangan bagi tim seleksi untuk menentukan penerima beasiswa secara cepat dan akurat. Ketidakseimbangan data antara pelamar yang memenuhi syarat dan yang tidak juga menjadi hambatan signifikan dalam proses ini. Oleh karena itu, diperlukan sistem analisis yang mampu mengidentifikasi karakteristik penerima beasiswa secara efisien dan efektif [2].

Untuk mengatasi tantangan dalam proses seleksi penerima Beasiswa Kartu Indonesia Pintar (KIP), beberapa peneliti telah memanfaatkan konsep data mining. Penelitian yang dilakukan oleh [2] misalnya, menggunakan metode C4.5 untuk mengklasifikasi penerima beasiswa. Dalam pengujian algoritma C4.5, penanganan data yang mengandung missing value dilakukan dengan metode Listwise Deletion, menghasilkan akurasi sebesar 92%. Pembagian data dilakukan dengan proporsi 80% untuk pelatihan (100 data) dan 20% untuk pengujian (25 data). Namun, penelitian ini memiliki perhatian karena belum menyelesaikan masalah class imbalance pada data dan hanya menggunakan metode pemisahan data berdasarkan jumlah tertentu untuk pelatihan dan pengujian. Oleh karena itu, penelitian ini mengusulkan penerapan metode cluster based undersampling yang dikombinasikan dengan genetic algorithm pada algoritma logistic regression untuk mengatasi ketidakseimbangan kelas/class imbalance serta meningkatkan performa klasifikasi.

Metode klasifikasi, seperti logistic regression, sering digunakan dalam seleksi penerima beasiswa ataupun objek lainnya untuk mengelompokkan data berdasarkan karakteristik tertentu [3]. Logistic regression menghubungkan variabel respon dengan variabel prediktor, menghasilkan dua kategori, yaitu 0 dan 1. Penelitian sebelumnya menunjukkan bahwa logistic regression dapat mencapai akurasi 84% dalam menentukan penerima Beasiswa Bank

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

Indonesia [4]. Namun, penelitian tersebut menghadapi kendala ketidakseimbangan kelas (class imbalance), yang dapat mengurangi performa model akibat tidak diterapkannya teknik data balancing.

Masalah class imbalance dapat memengaruhi kinerja model klasifikasi, terutama karena model cenderung bias terhadap kelas mayoritas, sehingga akurasi pada kelas minoritas lebih rendah [5]. Tantangan ini terjadi pada algoritma seperti decision tree, neural network, dan support vector machine yang mengasumsikan distribusi kelas seimbang. Salah satu solusi adalah metode data sampling, seperti undersampling, yang bertujuan menyeimbangkan distribusi data dengan mengurangi instance kelas mayoritas [6]. Namun, teknik ini berpotensi menghilangkan data penting sehingga akurasi hasil klasifikasi sering lebih rendah dibanding metode lainnya, yaitu hanya mencapai 78% [7].

Penelitian ini akan menerapkan metode Clustering Based Undersampling (CBU) untuk mengatasi masalah class imbalance. CBU merupakan teknik yang bertujuan mengurangi jumlah sampel kelas mayoritas secara terarah, sehingga informasi penting tetap terjaga. Teknik ini menggunakan algoritma klasterisasi, seperti K-Means, untuk mengelompokkan data dari kelas mayoritas, kemudian memilih beberapa sampel representatif dari setiap klaster. Pendekatan ini tidak hanya menyeimbangkan distribusi data, tetapi juga menjaga keberagaman sampel. Penelitian oleh [8] menunjukkan bahwa CBU efektif meningkatkan akurasi klasifikasi pada dataset tidak seimbang, serta menghasilkan hasil yang lebih stabil dibandingkan metode undersampling konvensional lainnya.

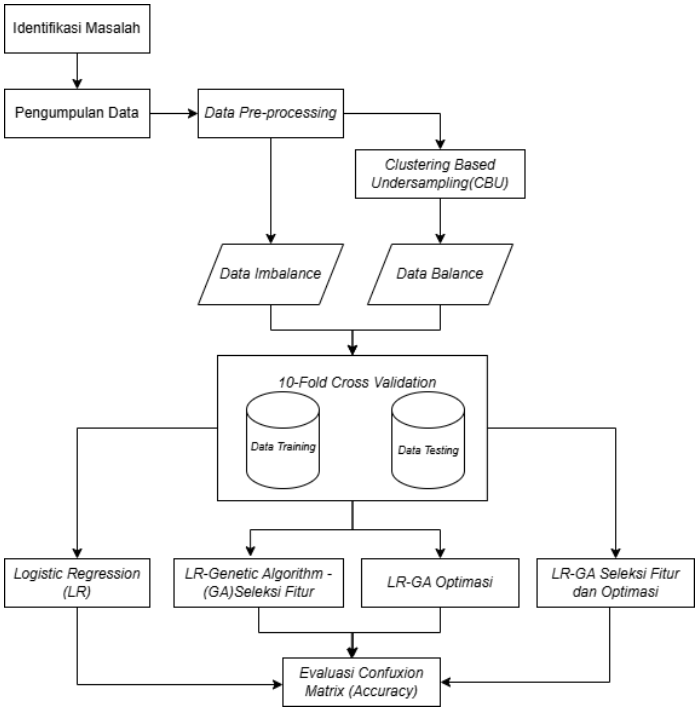
Selain menggunakan clustering-based undersampling untuk menyeimbangkan data, seleksi fitur dan optimasi diperlukan untuk meningkatkan akurasi model machine learning. Genetic Algorithm (GA) adalah metode optimasi berbasis seleksi alam yang mampu menemukan solusi optimal secara efisien [9]. Penelitian [10] menunjukkan bahwa penggunaan GA untuk mengoptimalkan bobot koneksi pada Backpropagation Neural Network (BPN) dalam klasifikasi penerimaan beasiswa Bidikmisi menghasilkan akurasi 90,47%. Studi tersebut menggunakan 12 variabel dengan validasi silang 10-fold, menegaskan efektivitas GA dalam meningkatkan performa model pada data yang kompleks.

Berdasarkan pembahasan di atas, penelitian ini bertujuan untuk mengatasi permasalahan *class imbalance* pada data penerima beasiswa KIP-Kuliah dengan menggabungkan metode *Genetic Algorithm (GA)* dan *Cluster-Based Undersampling (CBU)* yang diterapkan pada algoritma *Logistic Regression*. Kombinasi metode ini diharapkan tidak hanya meningkatkan akurasi dalam klasifikasi data, tetapi juga memberikan hasil yang lebih stabil dan representatif terhadap kelas minoritas. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem untuk seleksi penerimaan beasiswa.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini memiliki langkah-langkah yang akan dilakukan untuk mencapai tujuan dari penelitian. Adapun langkah-langkah penelitian yang akan dilakukan dimulai dari identifikasi masalah sampai dengan tahap akhirnya yaitu evaluasi hasil. Berikut adalah langkah-langkah yang akan dilakukan:



Gambar 1. Tahapan Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

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 Clustering Based Undersampling (CBU) 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 Logistic Regression (LR). Berikutnya permodelan kedua diterapkan dengan algoritma LR ditambah dengan Genetic Algorithm (GA) untuk seleksi fitur. Lalu permodelan ketiga menerapkan LR ditambah dengan GA untuk Optimasi, dan yang terakhir Permodelan menggunakan LR ditambah GA untuk seleksi fitur dan optimasi. Setelah itu didapatkanlah hasil akurasi confuxion matrix dari keempat permodelan tersebut.

2.2 Identifikasi Masalah

Penelitian ini berfokus pada analisis pengaruh kondisi ekonomi dan sosial keluarga terhadap keberhasilan akademis penerima beasiswa Kartu Indonesia Pintar (KIP) di Universitas Muhammadiyah Kalimantan Timur (UMKT). Studi pustaka dilakukan untuk mengidentifikasi kesenjangan penelitian terkait faktor-faktor tersebut, dengan penerapan metode Genetic Algorithm - Cluster Based Undersampling dan Logistic Regression untuk meningkatkan akurasi pada data tidak seimbang.

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 [11]. Data beasiswa KIP-Kuliah dari Bagian Kemahasiswaan UMKT memerlukan pengolahan seperti Untuk menghindari pengolahan data yang tidak diperlukan dilakukan data pre-pocessing dengan beberapa tahapan yang harus dilalui seperti data integration, data selection, data transformation, data cleaning, dan data [12] sebagai berikut:

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 [13].
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 [14].
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 [15].
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 [16].
e. Data Balancing
   Imbalance data kerap kali menjadi salah satu masalah dalam proses klasifikasi. Distribusi kelas yang tidak seimbang dapat memperburuk performa klasifikasi seperti overfitting, bias terhadap kelas mayoritas [17], Data balancing adalah proses menyeimbangkan distribusi kelas dalam dataset untuk menghindari bias pada algoritma klasifikasi akibat ketidakseimbangan jumlah sampel antar kelas. Dalam penelitian ini, terdapat masalah imbalanced data di mana jumlah instance dalam kelas "diterima beasiswa" jauh lebih sedikit dibandingkan dengan kelas "ditolak beasiswa". Dataset yang tidak seimbang terjadi ketika kelas-kelas di dalam data tidak terwakili secara merata, yang dapat menyebabkan model bias yang berkinerja buruk pada kelas minoritas [18] CBU juga melibatkan metode gabungan dengan K-Means yang memiliki dua jenis data clustering yang banyak digunakan dalam membuat pengelompokan data yaitu Hirarki dan Non Hirarki [19].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

2.5 Clustering Based Undersampling

Clustering based undersampling adalah teknik yang digunakan untuk mengatasi masalah class imbalance dalam dataset, di mana jumlah sampel dari kelas mayoritas jauh lebih banyak dibandingkan dengan kelas minoritas. Teknik ini melibatkan penggunaan algoritma clustering untuk mengelompokkan data dari kelas mayoritas, dan kemudian mengurangi jumlah sampel dengan memilih representatif dari setiap cluster [20]. Adapun rumus CBU menurut [21] sebagai berikut:

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - u_k|| \quad (1)$$

Pada Rumus (1), $J(c_k)$ merupakan fungsi objektif dalam metode Cluster Based Undersampling (CBU) yang digunakan untuk menyeimbangkan distribusi kelas dalam data. Fungsi ini menghitung jumlah jarak antara setiap titik data x_i dalam suatu kluster c_k dengan pusat kluster u_k, menggunakan norma Euclidean. Semakin kecil nilai $J(c_k)$, semakin dekat titik data dalam kluster terhadap pusatnya, sehingga pemilihan sampel dalam proses undersampling menjadi lebih representatif dan mempertahankan karakteristik utama dari kelas minoritas.

2.6 Pembagian Data Training dan Data Testing

Tahapan pembagian data dilakukan dengan membagi dataset menjadi 2 bagian yaitu data training dan data testing. Data training bertugas untuk melatih model dalam mempelajari pola dan hubungan antara fitur-fitur dalam data, sedangkan data testing bertugas untuk menguji kinerja model setelah proses training selesai. Teknik 10-Fold Cross-Validation akan diterapkan untuk mengevaluasi kinerja model machine learning, adapun nilai k yang berperan sebagai iterasi akan bernilai 10. Teknik ini membagi dataset menjadi 10 bagian yang sama besar, di mana setiap bagian secara bergantian akan digunakan sebagai data testing, sedangkan sisanya digunakan sebagai data training. Dengan mengambil nilai rata-rata dari 10 percobaan yang berbeda, 10-Fold Cross-Validation bertujuan untuk memberikan penilaian yang lebih akurat dan dapat diandalkan terhadap kinerja model [22].

2.7 Permodelan

Penelitian ini menggunakan model algoritma klasifikasi Logistic Regression (LR) dengan Clustering Based Undersampling (CBU) untuk menyeimbangkan data dan Genetic Algorithm (GA) untuk seleksi fitur dan optimasi akurasi. Model diterapkan pada kondisi Class Imbalance dan Balance, dengan perbandingan kinerja sebelum dan sesudah penggunaan CBU serta seleksi fitur dan optimasi menggunakan GA. Untuk bahasa pemrograman yang digunakan dalam menjalankan permodelan ini adalah bahasa python [23].

2.8 Algoritma Logistic Regression

Logistic Regression adalah algoritma klasifikasi machine learning yang digunakan untuk memakurasi probabilitas variabel dependen kategoris. Dalam Logistic Regression, variabel yang terikat adalah variable biner yang berisi data berkode 1 (Ya) atau 0 (Tidak). Metode ini merupakan metode regresi linier umum untuk mempelajari pemetaan dari sejumlah variabel numerik ke variabel biner atau probabilistic [24]. Adapun rumus LR menurut [25] sebagai berikut:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X \quad (2)$$

Pada Rumus (2), fungsi logistic regression merepresentasikan hubungan antara probabilitas kejadian suatu kelas dengan variabel prediktor X. Log-odds atau logit dari probabilitas p dihitung menggunakan persamaan $\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$, di mana b_0 adalah intercept dan b_1 adalah koefisien regresi. Persamaan ini memungkinkan pemetaan nilai input X ke dalam probabilitas biner, sehingga dapat digunakan dalam klasifikasi untuk menentukan apakah suatu data masuk ke dalam kelas tertentu atau tidak.

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 [20]. Adapun rumus menurut [9] sebagai berikut:

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

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.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

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 [26].

Tabel 1. Confuxion Matrix

		True Values	
		True	Values
Prediction	True	TP Correct Result	FP Unexpected Result
	False	FN Missing Result	TN Correct Absence Of Result

Pada Tabel 1, evaluasi yang digunakan penelitian ini adalah performa accuracy (akurasi). Akurasi merupakan tolak ukur yang digunakan dalam mengetahui seberapa tepat suatu pola klasifikasi memakurasi kelas data dari data yang akan datang. Dalam praktek data mining ataupun machine learning.
Adapun rumus akurasi menurut [1] sebagai berikut:

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

Pada Rumus (4), akurasi merupakan metrik evaluasi yang mengukur sejauh mana model klasifikasi dapat mengidentifikasi kelas dengan benar. Persamaan $\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ X } 100\%$ menunjukkan proporsi prediksi benar terhadap keseluruhan data. Di sini, TP (True Positive) dan TN (True Negative) mewakili jumlah prediksi yang benar untuk masing-masing kelas, sementara FP (False Positive) dan FN (False Negative) adalah jumlah prediksi yang salah. Lalu di kali kan dengan 100% untuk menampilkan hasil akurasi dalam bentuk percentage/persen.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data beasiswa kip-umkt. Data beasiswa yang diperoleh merupakan kumpulan data peserta beasiswa dari rentang waktu tahun 2021 – 2023. Data tersebut dikumpulkan dari bagian kemahasiswaan meliputi beberapa faktor sosial-ekonomi dalam penerimaan seleksi beasiswa. Data yang didapatkan memiliki total fitur sebanyak 37 fitur dengan 1075 Record. Adapun class atau label pada data ini menjadi dua yaitu class diterima sebanyak 424 dan class ditolak sebanyak 654 Record. Secara detail, hasil dari pengumpulan data terhadap 37 fitur yang didapatkan dari bagian kemahasiswaan. Adapun fitur-fitur data yang berisikan no.pendaftaran, nama siswa, nik, no. kartu keluarga, nik keluarga, nisn, status dtks, status p3ke, no. kip, no. kks, asal sekolah, kab/kota sekolah, provinsi sekolah, tempat lahir, tanggal lahir, jenis kelamin, alamat tinggal, no. 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, mck, jarak pusat kota, program studi, akreditas prodi, dan status pengajuan.

3.1 Hasil Pre-processing

Pada tahap ini, akan menampilkan hasil proses dari pra-processing yang telah dilakukan terhadap data, dan diterapkan untuk dua kondisi yang pertama untuk data imbalance dan yang kedua untuk data balance. Hasil akan ditampikan sebagai berikut:

a. Data Integration
   Data beasiswa yang dikumpulkan dari Kemahasiswaan Universitas Muhammadiyah Kalimantan Timur (UMKT) telah digabungkan dari 2 sumber yang relevan dengan proses penerimaan beasiswa. Proses penggabungan ini bertujuan untuk mempermudah pengolahan data. Langkah ini menjadi penting karena setiap sumber data memberikan kontribusi informasi yang berbeda, sehingga menciptakan dataset yang lebih kaya dan komprehensif untuk analisis selanjutnya. Berikut tabelnya akan ditampilkan pada Tabel 2.

Tabel 2. 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

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

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

b. Data Selection

Data selection adalah proses pemilihan fitur relevan untuk analisis lebih lanjut. Dalam penelitian ini, fitur dipilih berdasarkan relevansinya dalam memprediksi akurasi penerima beasiswa, sedangkan fitur yang dianggap kurang relevan tidak digunakan. Dari 37 fitur yang tersedia dalam dataset Beasiswa Kip-Kuliah UMKT, beberapa fitur yang dinilai kurang relevan untuk klasifikasi beasiswa akan dihilangkan dan menjadi total 23 dataset, dan hasil seleksi data akan dipaparkan dalam Tabel 3.

Tabel 3. 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

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

Pada Tabel 3, 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

Data transformation adalah proses mengubah data ke format atau skala yang sesuai untuk analisis. Pada penelitian ini, data kategorikal (string) akan diubah menjadi numerik menggunakan fungsi LabelEncoder dari library sklearn.preprocessing [15]. Transformasi ini dilakukan untuk mempermudah algoritma klasifikasi (LR) dalam perhitungan. Fitur data yang akan ditransformasi akan ditampilkan dalam Tabel 4.

Tabel 4. Hasil Data Transformation

No	Status Pengajuan (Sebelum data Transformation)	Status Pengajuan (Sesudah data Transformation)
1	Diterima	0
2	Ditolak	1

Pada Tabel 4, 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.

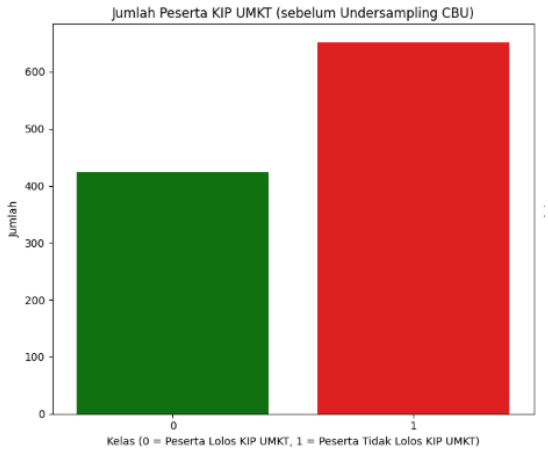
```
Data Awal:
Jumlah Baris dan Kolom Sebelum Cleaning: (1080, 23)

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.

e. Data Balancing

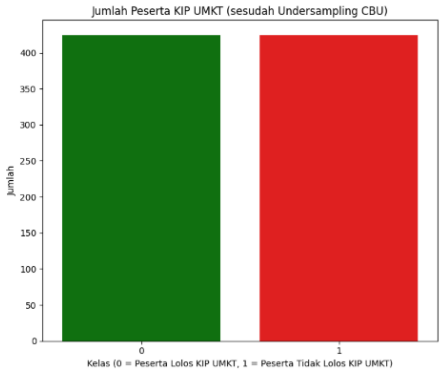
Data balancing adalah proses menyeimbangkan distribusi kelas dalam dataset untuk menghindari bias pada algoritma klasifikasi akibat ketidakseimbangan jumlah sampel antar kelas. Dalam penelitian ini, terdapat masalah imbalanced data antara kelas "diterima”beasiswa dan "ditolak” beasiswa, yang dapat menyebabkan model lebih bias terhadap kelas mayoritas (John Baer, 2023). Untuk mengatasi hal ini, teknik Clustering Based Undersampling (CBU) digunakan, di mana algoritma clustering mengelompokkan data kelas mayoritas dan mengurangi jumlah sampel dengan memilih representatif dari setiap cluster [20]. Hasil perbandingan bisa dilihat pada gambar sebagai berikut



Gambar 3. Hasil Sebelum Clustering Based Undersampling

Pada Gambar 3, menampilkan diagram dataset sebelum melakukan data balancing menggunakan metode clustering based undersampling. Hasil yang didapat ketika belum dibalancing terdapat 1075 dataset, dengan ketentuan class 0(lolos) 654 record dan class 1(tidak lolos) 424 record.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747



Gambar 4. Hasil Sesudah Clustering Based Undersampling

Pada Gambar 4, menampilkan diagram dataset setelah melakukan data balancing menggunakan metode clustering based undersampling. Hasil yang didapat ketika setelah dibalancing terdapat 848 dataset, dengan ketentuan class 0(lolos) 424 record dan class 1(tidak lolos) 424 record.

3.2 Hasil Pembagian Data Training dan Data Testing

Pada tahap ini, pembagian dataset dibagi menjadi dua yaitu data training dan data testing untuk menambah kinerja model machine learninng. Menggunakan 10-Fold Cross Validation, dataset dibagi menjadi sepuluh bagian sama besar. Setiap iterasi, satu bagian digunakan sebagai data testing dan sembilan sebgai data training. Demi mengurangi bias dan variasi dalam estimasi kinerja model serta memastikan setiap sampel diuji.

3.3 Hasil Permodelan

Pada tahap ini, akan menampilkan hasil pembelajaran algoritma dalam bentuk akurasi yang didapatkan oleh model mulai dari sekedar menggunakan algoritma klasifikasi Logistic Regression (LR) hingga ditambahkan model-model lainnya seperti Cluster Based Undersampling (CBU) dan Genetic Algorithm (GA) seleksi fitur&optimasi yang akan dijelaskan sebelumnya terhadap klasifikasi data beasiswa kip-kuliah UMKT, beserta beberapa hasil dari penerapan algoritma seleksi fitur GA terhadap dataset yang memberikan kombinasi fitur-fitur terbaik [27]. Semua pengerjaan ini akan dilakukan pada tahap permodelan sebagai berikut:

3.3.1 Permodelan Logistic Regression

Algoritma klasifikasi Logistic Regression di terapkan tanpa menggunakan balancing undersampling dari CBU dan diterapkan juga menggunakan balancing undersampling dari CBU. 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 CBU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 5. Hasil Rata-Rata Akurasi LR Tanpa CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	535	287	137	116	62.51%

$$\text{Accuracy} = \frac{535+137}{535+287+137+116} \text{ X } 100\% = 62.51\%$$

Pada Tabel 5, menampilkan hasil rata-rata akurasi LR tanpa CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 535, fp 287, tn 137, fn 116. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 62.51%.

Tabel 6. Hasil Rata-Rata Akurasi LR Dengan CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	289	165	259	135	64.63%

$$\text{Accuracy} = \frac{289+259}{289+165+259+135} \text{ X } 100\% = 64.63\%$$

Pada Tabel 6, menampilkan hasil rata-rata akurasi LR dengan CBU(Undersampling untuk menyeimbangkan data)(balance). Mendapatkan total tp 289, fp 165, tn 259, fn 135. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 64.63%.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

3.3.2 Permodelan LR-Genetic Algorithm Untuk Seleksi Fitur

Algoritma klasifikasi Logistic Regression ditambah dengan Genetic Algorithm untuk Seleksi Fitur di terapkan tanpa menggunakan balancing undersampling dari CBU dan diterapkan juga menggunakan balancing undersampling dari CBU. 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 CBU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 7. Hasil Seleksi Fitur GA Tanpa CBU

No	Nama Fitur	Fitur Ke-
1	No. KKS	3
2	Jenis Kelamin	4
3	Pekerjaan Ayah	5
4	Penghasilan Ayah	6
5	Penghasilan Ibu	9
6	Status Ibu	10
7	Kepemilikan Rumah	12
8	Luas Tanah	15
9	Sumber Air	17
10	MCK	18
11	Jarak Pusat Kota (KM)	19
12	Akreditasi Prodi	21

Pada Tabel 7, menampilkan hasil seleksi fitur yang dipilih Genetic Algorithm dalam menjalankan permodelannya. Dari 23 fitur utama terdapat 12 yang digunakan GA sebagai fitur yaitu fitur ke-3,4,5,6,9,10,12,15,17,18,19,21. Dan permodelan ini dijalankan tanpa CBU (Undersampling untuk menyeimbangkan data) jadi data yang digunakan adalah data imbalance. Hasil rata-rata dan perhitungan akurasinya sebagai berikut:

Tabel 8. Hasil Rata-Rata Akurasi LR-GA Seleksi Fitur Tanpa CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	564	284	140	87	65.48%

$$\text{Accuracy} = \frac{564+140}{564+284+140+87} \text{ X } 100\% = 65.48\%$$

Pada Tabel 8, menampilkan hasil rata-rata akurasi LR tanpa CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 564, fp 284, tn 140, fn 87. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 65.48%.

Tabel 9. Hasil Seleksi Fitur GA Seleksi Fitur Dengan CBU

No	Nama Fitur	Fitur Ke-
1	Status DTKS	0
2	Status P3KE	1
3	No. KIP	2
4	No. KKS	3
5	Jenis Kelamin	4
6	Pekerjaan Ayah	5
7	Status Ibu	10
8	Jumlah Tanggungan	11
9	Tahun Perolehan	13
10	Sumber Listrik	14
11	Luas Tanah	15
12	MCK	18

Pada Tabel 9, menampilkan hasil seleksi fitur yang dipilih Genetic Algorithm dalam menjalankan permodelannya. Dari 23 fitur utama terdapat 12 yang digunakan GA sebagai fitur yaitu fitur ke-0, 1, 2,3,4,5,10,11,13,14,15,18. Dan permodelan ini dijalankan dengan CBU (Undersampling untuk menyeimbangkan data) jadi data yang digunakan adalah data balance. Hasil rata-rata dan perhitungan akurasinya sebagai berikut:

Tabel 10. Hasil Rata-Rata Akurasi LR-GA Seleksi Fitur Dengan CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	301	151	273	123	67,68%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

$$\text{Accuracy} = \frac{301+274}{301+151+273+123} \text{ X } 100\% = 67{,}68\%$$

Pada Tabel 10, menampilkan hasil rata-rata akurasi LR dengan CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 301, fp 151, tn 273, fn 123. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 67.68%.

3.3.3 Permodelan LR-GA Untuk Optimasi

Algoritma klasifikasi Logistic Regression ditambah dengan Genetic Algorithm untuk peningkatan optimasi di terapkan tanpa menggunakan balancing undersampling dari CBU dan diterapkan juga menggunakan balancing undersampling dari CBU. 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 CBU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 11. Hasil Rata-Rata Akurasi LR-GA Optimasi Tanpa CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	563	291	133	88	64.74%

$$\text{Accuracy} = \frac{563+133}{563+291+133+88} \text{ X } 100\% = 64.74\%$$

Pada Tabel 11, menampilkan hasil rata-rata akurasi LR tanpa CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 563, fp 291, tn 133, fn 88. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 64.74%.

Tabel 12. Hasil Rata-Rata Akurasi LR-GA Optimasi Dengan CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	309	177	247	115	65.55%

$$\text{Accuracy} = \frac{309+247}{309+177+247+115} \text{ X } 100\% = 65.55\%$$

Pada Tabel 12, menampilkan hasil rata-rata akurasi LR dengan CBU(Undersampling untuk menyeimbangkan data)(balance). Mendapatkan total tp 309, fp 177, tn 247, fn 115. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 65.55%.

3.3.4 Permodelan LR-Genetic Algorithm Untuk Seleksi Fitur dan Optimasi

Algoritma klasifikasi Logistic Regression ditambah dengan Genetic Algorithm untuk Seleksi Fitur dan Optimasi di terapkan tanpa menggunakan balancing undersampling dari CBU dan diterapkan juga menggunakan balancing undersampling dari CBU. 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 CBU. Hasilnya akan ditampilkan sebagai berikut:

Tabel 13. Hasil Seleksi Fitur GA Seleksi Fitur dan Optimasi Tanpa CBU

No	Nama Fitur	Fitur Ke-
1	Status DTKS	0
2	No. KIP	2
3	No. KKS	3
4	Pekerjaan Ayah	5
5	Penghasilan Ibu	9
6	Jumlah Tanggungan	11
7	Sumber Listrik	14
8	Luas Tanah	15
9	Luas Bangunan	16
10	MCK	18
11	Jarak Pusat Kota (KM)	19
12	Akreditasi Prodi	21

Pada Tabel 13, menampilkan hasil seleksi fitur yang dipilih Genetic Algorithm dalam menjalankan permodelannya. Dari 23 fitur utama terdapat 12 yang digunakan GA sebagai fitur yaitu fitur ke-0, 2, 3, 5, 9, 11, 14, 15, 16, 18, 19, 21. Dan permodelan ini dijalankan tanpa CBU(Undersampling untuk menyeimbangkan data) jadi data yang digunakan adalah data imbalance. Hasil rata-rata dan perhitungan akurasinya sebagai berikut:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

Tabel 14. Hasil Rata-Rata Akurasi LR-GA Seleksi Fitur dan Optimasi Tanpa CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	553	259	165	98	66.78%

$$\text{Accuracy} = \frac{553+259}{553+259+165+98} \text{ X } 100\% = 66.78\%$$

Pada Tabel 14, menampilkan hasil rata-rata akurasi LR tanpa CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 553, fp 259, tn 165, fn 98. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 66.78%.

Tabel 15. Hasil Seleksi Fitur GA Seleksi Fitur dan Optimasi Dengan CBU

No	Nama Fitur	Fitur Ke-
1	Status DTKS	0
2	Status P3KE	1
3	No. KKS	3
4	Pekerjaan Ayah	5
5	Status Ayah	6
6	Status Ibu	7
7	Jumlah Tanggungan	10
8	Sumber Listrik	11
9	Luas Tanah	14
10	Sumber Air	15
11	MCK	18
12	Program Studi	20

Pada Tabel 15, menampilkan hasil seleksi fitur yang dipilih Genetic Algorithm dalam menjalankan permodelannya. Dari 23 fitur utama terdapat 12 yang digunakan GA sebagai fitur yaitu fitur ke-0, 1, 3, 5, 6, 7, 10, 11, 14, 15, 18, 20. Dan permodelan ini dijalankan dengan CBU (Undersampling untuk menyeimbangkan data) jadi data yang digunakan adalah data balance. Hasil rata-rata dan perhitungan akurasinya sebagai berikut:

Tabel 16. Hasil Rata-Rata Akurasi LR-GA Seleksi Fitur dan Optimasi Dengan CBU

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	300	154	270	124	67.21%

$$\text{Accuracy} = \frac{300+154}{300+154+274+124} \text{ X } 100\% = 67.21\%$$

Pada Tabel 16, menampilkan hasil rata-rata akurasi LR dengan CBU(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 300, fp 154, tn 274, fn 124. Dan dilanjutkan dengan rumus akurasi total hasilnya adalah 67.21%.

3.4 Hasil Perbandingan

Pada tahap ini, akan menampilkan hasil perbandingan dari keseluruhan permodelan diatas,dan akan dijelaskan detailnya di sub bab selanjutnya yaitu tahap hasil peningkatan akurasi, berikut tampilan hasil perbandingan permodelan:

Tabel 17. Perbandingan Hasil Fitur Yang Dipilih GA Seleksi Fitur dan Optimasi

No	Nama Fitur	Sebelum Optimasi	Setelah Optimasi	Setelah CBU	Setelah Optimasi + CBU
1	Status DTKS	×	✓	✓	✓
2	Status P3KE	×	×	✓	✓
3	No. KIP	×	✓	✓	×
4	No. KKS	✓	✓	✓	✓
5	Jenis Kelamin	✓	×	✓	×
6	Pekerjaan Ayah	✓	✓	✓	✓
7	Status Ayah	×	×	×	✓
8	Status Ibu	✓	×	✓	✓
9	Penghasilan Ayah	✓	×	×	×
10	Penghasilan Ibu	✓	✓	×	×
11	Jumlah Tanggungan	×	✓	✓	✓
12	Sumber Listrik	×	✓	✓	✓
13	Luas Tanah	✓	✓	✓	✓
14	Sumber Air	✓	×	×	✓

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

15	MCK	✓	✓	✓	✓
17	Akreditasi Prodi	✓	✓	×	×

Pada Tabel 17, menampilkan perubahan fitur yang dipilih pada berbagai tahap proses, yaitu sebelum optimasi, setelah optimasi, setelah penerapan Cluster-Based Undersampling (CBU), dan setelah kombinasi optimasi dan CBU. Beberapa fitur, seperti *Status DTKS* dan *Pekerjaan Ayah*, dipertahankan setelah proses optimasi, sementara fitur seperti *Status Ayah* hanya terpilih pada tahap kombinasi optimasi dan CBU, mencerminkan peningkatan selektivitas dan relevansi fitur dalam mendukung akurasi model.

Tabel 18. Perbandingan Hasil Rata-Rata Akurasi Model Tanpa CBU

LR	LR-GA Seleksi Fitur	LR-GA Optimasi	LR-GA Seleksi Fitur & Optimasi	Perubahan LR ke LR-GA Seleksi Fitur	Perubahan LR ke LR-GA Optimasi	Perubahan LR ke LR-GA Seleksi Fitur & Optimasi
62.51%	65.48%	64.74%	66.78%	+2.97%	+2.22%	+4.27%

Pada Tabel 18 menampilkan perbandingan rata-rata akurasi model Logistic Regression (LR) tanpa CBU pada empat kondisi: baseline, setelah seleksi fitur GA, optimasi GA, dan kombinasi keduanya. Akurasi meningkat dari 62% menjadi 65% dengan seleksi fitur dan optimasi, serta mencapai 67% pada kombinasi keduanya, dengan peningkatan akurasi maksimal sebesar 5% dari baseline.

Tabel 19. Perbandingan Hasil Rata-Rata Akurasi Model Dengan CBU

LR	LR-GA Seleksi Fitur	LR-GA Optimasi	LR-GA Seleksi Fitur & Optimasi	Perubahan LR ke LR-GA Seleksi Fitur	Perubahan LR ke LR-GA Optimasi	Perubahan LR ke LR-GA Seleksi Fitur & Optimasi
64.63%	67.68%	65.55%	67.21%	+3.04%	+0.91%	+2.58%

Pada Tabel 19, menampilkan rata-rata akurasi model Logistic Regression (LR) dengan penerapan CBU. Akurasi meningkat dari 65% menjadi 67% pada seleksi fitur GA dan kombinasi GA, meskipun hanya meningkat 66% pada optimasi GA saja. Peningkatan tertinggi terjadi pada kombinasi seleksi fitur dan optimasi, mencatat kenaikan sebesar 2% dibandingkan baseline.

3.5 Pembahasan

Penelitian ini berfokus pada klasifikasi data beasiswa kip-kuliah Universitas Muhammadiyah Kalimantan Timur, dengan menggunakan data yang diperoleh dari Bagian Kemahasiswaan UMKT selama periode 2021-2023. Proses penelitian ini meliputi beberapa tahapan penting, dimulai dari identifikasi masalah, pengumpulan data, data pre-processing, hingga penerapan model machine learning. Penelitian ini mengimplementasikan dua versi model, pertama menggunakan teknik undersampling CBU dan kedua tanpa teknik tersebut. Kedua model ini kemudian dibandingkan berdasarkan hasil evaluasinya untuk menentukan pendekatan yang paling efektif dalam klasifikasi data beasiswa kip-kuliah UMKT. Setelah melalui berbagai macam proses dan mendapatkan hasil akhir, maka di tahap akhir akan membahas hasil yang didapat dan dikaitkan kembali dengan masalah yang dirumuskan diawal, mulai menjawab masalah mengenai seberapa besar peningkatan accuracy yang didapatkan oleh Logistic Regression (LR) dalam mengklasifikasi data beasiswa kip-kuliah UMKT dengan menerapkan metode Clustering Based Undersampling(CBU) sebagai undersampling dalam menyeimbangkan data class imbalance dan seleksi fitur serta optimasi menggunakan Genetic Algorithm (GA). Berdasarkan rumusan masalah yang ada, berikut adalah pembahasan lanjutan.

3.6 Hasil Peningkatan Akurasi

Hasil penelitian ini, dilakukan evaluasi terhadap model Logistic Regression (LR) yang diterapkan pada data beasiswa KIP Kuliah dengan berbagai pendekatan, termasuk penerapan Genetic Algorithm (GA) untuk seleksi fitur dan optimasi, serta penggunaan teknik Cluster Based Undersampling (CBU) untuk menangani masalah ketidakseimbangan kelas. Hasil dari setiap pendekatan dibandingkan untuk menilai peningkatan akurasi yang dicapai. Sebelum penerapan CBU, rata-rata akurasi LR tanpa modifikasi adalah 62.51%. Setelah penerapan GA untuk seleksi fitur, akurasi meningkat menjadi 65.48%, memberikan perubahan sebesar +2.97%. Penerapan optimasi dengan GA menghasilkan rata-rata akurasi 64.74% (+2.23%), sementara kombinasi seleksi fitur dan optimasi memberikan rata-rata akurasi tertinggi sebesar 66.78% (+4.27%). Setelah penerapan CBU, rata-rata akurasi LR meningkat menjadi 64.63%. Dengan penerapan GA untuk seleksi fitur, rata-rata akurasi meningkat menjadi 67.68% (+3.05%). Metode optimasi dengan GA memberikan rata-rata akurasi 65.55% (+0.92%), sedangkan kombinasi seleksi fitur dan optimasi menghasilkan rata-rata akurasi sebesar 67.21% (+2.58%). Dari hasil yang diperoleh, dapat disimpulkan bahwa penerapan GA untuk seleksi fitur dan optimasi, baik sebelum maupun setelah penerapan CBU, memberikan kontribusi positif terhadap peningkatan akurasi model. Penerapan teknik GA untuk seleksi fitur dan optimasi, baik sebelum maupun setelah penerapan CBU, menunjukkan peningkatan akurasi yang signifikan dalam model Logistic Regression. Kombinasi dari kedua teknik ini memberikan hasil terbaik, dengan akurasi tertinggi dicapai setelah penerapan CBU, yang menunjukkan bahwa pendekatan yang terintegrasi dapat meningkatkan kinerja model dalam mengklasifikasikan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

data yang tidak seimbang, seperti pada kasus beasiswa KIP Kuliah.Analisis terhadap fitur-fitur yang terpilih menunjukkan bahwa fitur-fitur yang konsisten terpilih, seperti No. KKS, Pekerjaan Ayah, Luas Tanah, dan MCK, memiliki relevansi yang tinggi terhadap target klasifikasi, yang berkontribusi pada peningkatan akurasi model. Selain itu, fitur-fitur baru yang muncul setelah penerapan optimasi dan *CBU*, seperti Status P3KE dan Status Ayah, menunjukkan bahwa proses seleksi dan optimasi berhasil mengidentifikasi atribut yang lebih signifikan untuk klasifikasi.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan mengenai penerapan metode *Genetic Algorithm (GA)* yang dikombinasikan dengan *Clustering Based Undersampling (CBU)* pada algoritma *Logistic Regression* untuk mengatasi *class imbalance* data penerima beasiswa KIP-Kuliah, dapat disimpulkan bahwa penerapan *CBU* berhasil menyeimbangkan distribusi kelas dalam dataset, meningkatkan akurasi model Logistic Regression dari 62.51% menjadi 67.68% setelah penerapan *GA* untuk seleksi fitur dan optimasi. Metode seleksi fitur dan optimasi menggunakan *GA* secara bersamaan memberikan hasil terbaik, menunjukkan kontribusi signifikan terhadap peningkatan akurasi model dibandingkan dengan penerapan salah satu metode secara terpisah. Kombinasi *CBU* dan *GA* pada *Logistic Regression* terbukti efektif dalam menangani masalah *class imbalance*, meningkatkan akurasi keseluruhan dan memberikan hasil yang lebih stabil dalam mengklasifikasikan kelas minoritas. Penelitian selanjutnya disarankan untuk mengeksplorasi metode balancing lainnya, seperti SMOTE (Synthetic Minority Over-sampling Technique) atau hybrid sampling, untuk membandingkan efektivitasnya terhadap model klasifikasi. serta penggunaan algoritma klasifikasi yang lebih kompleks seperti Gradient Boosting, untuk melihat potensi peningkatan akurasi lebih lanjut. Penggunaan algoritma yang lebih besar dapat membantu dalam menangani data yang lebih kompleks dan beragam.

REFERENCES

[1] N. Indriyani, A. Fauzi, and A. B. H. Y. Yanto, “Pemodelan Prediksi Penerima Beasiswa Kip-Kuliah Menggunakan Metode Weight Product,” *IMTechno J. Ind. Manag. Technol.*, vol. 5, no. 1, 2024, doi: 10.31294/imtechno.v5i1.2958.
[2] A. S. Suweleh, D. Susilowati, and Hairani, “Aplikasi Penentuan Penerima Beasiswa Menggunakan Algoritma C4.5,” *J. BITe*, vol. 2, no. 1, pp. 12–21, 2020, doi: 10.30812/bite.v2i1.798.
[3] P. Dewi, R. Nur Aulia, and R. Taufiqillah, “Customer Churn Prediction for Life Insurance Using Binary Logistic Regression,” *Econ. Rev. J.*, vol. 3, no. 3, pp. 2289–2299, 2024, doi: 10.56709/mrj.v3i3.353.
[4] D. Megah Sari, N. Arifin, Nurfitrianingsih, and A. M. Yusuf, “Implementation of Decision Support System for Scholarship Recipients at Bank Indonesia,” *Ceddi J. Educ.*, vol. 1, no. 1, pp. 13–22, 2022, doi: 10.56134/cje.v1i1.10.
[5] J. Prasetya, “Penerapan Klasifikasi Naive Bayes dengan Algoritma Random Oversampling dan Random Undersampling pada Data Tidak Seimbang Cervical Cancer Risk Factors,” *Leibniz J. Mat.*, vol. 2, no. 2, pp. 11–22, 2022, doi: 10.59632/leibniz.v2i2.173.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, no. February, pp. 321–357, 2020, doi: 10.1613/jair.953.
[7] M. Kim and K. B. Hwang, “An empirical evaluation of sampling methods for the classification of imbalanced data,” *PLoS One*, vol. 17, no. 7 July, pp. 1–22, 2022, doi: 10.1371/journal.pone.0271260.
[8] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, “The effect of rebalancing techniques on the classification performance in cyberbullying datasets,” *Neural Comput. Appl.*, vol. 36, no. 3, pp. 1049–1065, 2024, doi: 10.1007/s00521-023-09084-w.
[9] 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.
[10] N. Cahyani, S. S. Pangastuti, K. Fithriasari, I. Irhamah, and N. Iriawan, “Classification of Bidikmisi Scholarship Acceptance using Neural Network Based on Hybrid Method of Genetic Algorithm,” *Indones. J. Stat. Its Appl.*, vol. 5, no. 2, pp. 396–404, 2021, doi: 10.29244/ijsa.v5i2p396-404.
[11] 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.
[12] Y. D. Evitasari, W. J. Pranoto, and N. A. Verdikha, “Evaluasi Support Vector Machine Dengan Optimasi Metode Genetic Algorithm Pada Klasifikasi Banjir Kota Samarinda,” *J. Sains Komput. dan Teknol. Inf.*, vol. 6, no. 1, pp. 49–53, 2023, doi: 10.33084/jsakti.v6i1.5462.
[13] 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.
[14] F. Sulianta, *Basic Data Mining from A to Z*. Feri Sulianta, 2023. [Online]. Available: https://books.google.co.id/books?id=JcLhEAAAQBAJ
[15] 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.
[16] 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.
[17] D. Ariyadi, T. Azhima, and Y. Siswa, “Penerapan Metode PSO-SMOTE Pada Algoritma Random Forest Untuk Mengatasi Class Imbalance Data Bencana Tanah Longsor,” vol. 6, no. 1, pp. 320–329, 2025.
[18] A. Kochkarev, A. Khvostikov, D. Korshunov, A. Krylov, and M. Boguslavskiy, “Data balancing method for training

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2322−2334
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6747

segmentation neural networks," *CEUR Workshop Proc.*, vol. 2744, pp. 1–9, 2020, doi: 10.51130/graphicon-2020-2-4-19.
[19] M. Fajar and Rudiman, "Klasifikasi Jenis Tanah Wakaf Muhammadiyah di Tanjung Redeb dengan Metode K-Means Berbasis Sig," *Borneo Student Res.*, vol. 3, no. 2, p. 2022, 2022, [Online]. Available: https://muhammadsyaf.wordpress.com/2017/03/04/sistem-informasi-geografis-dan-
[20] 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.
[21] J. Zhang, L. Chen, and F. Abid, "Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method," *J. Healthc. Eng.*, vol. 2019, 2020, doi: 10.1155/2019/7294582.
[22] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.
[23] Budhi Gustiandi, *Langkah Awal Menguasai Bahasa Pemrograman Python*. 2023. doi: 10.55981/brin.656.
[24] F. H. Harahap, "IJM : Indonesian Journal of Multidisciplinary Klasifikasi Menggunakan Model Regresi Logistik Multinomial dan Regresi Logistik Multinomial Komponen Utama," vol. 1, pp. 632–642, 2023.
[25] P. Schober and T. R. Vetter, "Statistical Minute," *Int. Anesth. Res. Soc.*, vol. 129, no. 2, p. 2019, 2021.
[26] 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.
[27] R. Syaputra, T. A. Y. Siswa, and W. J. Pranoto, "Model Optimasi SVM Dengan PSO-GA dan SMOTE Dalam Menangani High Dimensional dan Imbalance Data Banjir," *Teknika*, vol. 13, no. 2, pp. 273–282, 2024, doi: 10.34148/teknika.v13i2.876.