Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

Penerapan Metode *GA-TL* Pada Algoritma *Naive Bayes* Untuk Mengatasi *Class Imbalance* Data Beasiswa KIP-Kuliah

Dessy Widyastuti, Taghfirul Azhima Yoga Siswa*, Rudiman

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
Email: [1]2111102441122@umkt.ac.id, [2,*]tay758@umkt.ac.id, [3]rud959@umkt.ac.id
Email Penulis Korespondensi: tay758@umkt.ac.id
Submitted: **14/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Program Kartu Indonesia Pintar (KIP) Kuliah bertujuan membantu mahasiswa dari keluarga kurang mampu untuk melanjutkan pendidikan, namun distribusi data penerima sering mengalami *class imbalance*. Masalah *class imbalance* dalam data penerima Program Kartu Indonesia Pintar (KIP) Kuliah dapat menyebabkan ketidaktepatan sasaran dalam distribusi beasiswa. Ketidakseimbangan jumlah data antara kelompok penerima dan non-penerima dapat memengaruhi kinerja model klasifikasi, sehingga model cenderung lebih akurat dalam memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Hal ini dapat berakibat pada terlewatnya calon penerima beasiswa yang seharusnya memenuhi kriteria, sehingga tujuan utama program untuk membantu mahasiswa dari keluarga kurang mampu menjadi tidak optimal. Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi data penerima beasiswa KIP Kuliah di Universitas Muhammadiyah Kalimantan Timur dengan mengatasi masalah tersebut. Pendekatan yang digunakan adalah kombinasi metode *Genetic Algorithm* untuk seleksi fitur dan optimasi, serta *Tomek Links-Random Undersampling* untuk *balancing* data. Proses penelitian mencakup data *pre-processing*, penerapan 10-*fold cross-validation*, dan evaluasi performa model menggunakan *confusion matrix*. Hasil penelitian menunjukkan bahwa tanpa *Tomek Links-Random Undersampling*, akurasi *Naïve Bayes* meningkat dari 65,2% menjadi 66,0% setelah seleksi fitur dan optimasi menggunakan *Genetic Algorithm*. Sementara itu, dengan penerapan *Tomek Links-Random Undersampling*, akurasi *Naïve Bayes* meningkat dari 56% menjadi 63%. Selain itu, metode ini juga meningkatkan keadilan dalam pengklasifikasian data penerima beasiswa sehingga distribusi bantuan menjadi lebih merata. Peningkatan akurasi pada model ini memberikan manfaat yang signifikan dalam proses seleksi penerima beasiswa di masa depan. Dengan demikian, integrasi pendekatan *machine learning* yang efisien berkontribusi pada optimalisasi program KIP Kuliah, memastikan penerima manfaat tepat sasaran sesuai kriteria yang telah ditentukan.

Kata Kunci: Class Imbalance; Genetic Algorithm; Naive Bayes; Random Undersampling; Tomek Links

Abstract−The Indonesia Smart Card (KIP) Scholarship Program aims to support students from underprivileged families in pursuing higher education, yet the distribution of recipient data often experiences class imbalance, leading to inaccuracies in scholarship allocation. This imbalance, characterized by disproportionate data between recipient and non-recipient groups, affects classification model performance, causing models to favor the majority class and overlook the minority class, potentially excluding eligible recipients. To address this issue, this study combines the Genetic Algorithm for feature selection and optimization with Tomek Links-Random Undersampling for data balancing. The research process includes data preprocessing, 10-fold cross-validation, and performance evaluation using a confusion matrix. Results indicate that without Tomek Links-Random Undersampling, Naïve Bayes accuracy increased from 65.2% to 66.0% after feature selection and optimization using the Genetic Algorithm, while applying Tomek Links-Random Undersampling improved accuracy from 56% to 63%. This method also enhanced fairness in recipient classification, promoting a more equitable distribution of benefits. The improved model accuracy significantly aids future scholarship selection processes, demonstrating that integrating efficient machine learning approaches optimizes the KIP Scholarship Program by ensuring beneficiaries are appropriately targeted based on predetermined criteria.

Keywords: Class Imbalance; Genetic Algorithm; Naive Bayes; Random Undersampling; Tomek Links

1. PENDAHULUAN

Kartu Indonesia Pintar (KIP) merupakan program pemerintah yang bertujuan untuk memberikan bantuan pendidikan kepada anak-anak dari keluarga kurang mampu di Indonesia. Program ini dirancang untuk memastikan bahwa setiap anak, terutama yang berasal dari keluarga miskin dan rentan, memiliki akses untuk melanjutkan pendidikan hingga tingkat yang lebih tinggi. KIP tidak hanya memberikan bantuan finansial, tetapi juga berfungsi sebagai alat untuk mendorong partisipasi pendidikan dan mengurangi angka putus sekolah di Indonesia [1]. Dengan adanya KIP, diharapkan anak-anak dapat menyelesaikan pendidikan mereka tanpa terhambat oleh masalah ekonomi [2].

Program Kartu Indonesia Pintar (KIP) menghadapi masalah ketidakmerataan akses pendidikan akibat ketidakakuratan data penerima. Hal ini menyebabkan siswa yang layak tidak terdaftar, sementara yang tidak memenuhi syarat justru mendapat bantuan, sehingga distribusi sumber daya menjadi tidak adil [3]. *Balancing* data diperlukan untuk memastikan penyesuaian jumlah data antara kelompok penerima yang memenuhi syarat dan tidak, agar program lebih efektif [4].

Penerapan teknik data *mining* dalam berbagai bidang telah menunjukkan potensi yang signifikan untuk meningkatkan efisiensi dan efektivitas proses pengambilan keputusan. Penerapan data *mining* juga dapat membantu dalam mengidentifikasi mahasiswa yang berisiko tidak lulus atau membutuhkan dukungan tambahan. Dengan menggunakan teknik analisis data, institusi dapat memantau kinerja akademik mahasiswa dan memberikan intervensi yang diperlukan untuk mendukung mereka dalam mencapai tujuan akademis mereka [5]. Penelitian ini memprediksi hasil studi mahasiswa melalui pendekatan data *analytic* yaitu data *mining*, dengan menggunakan salah satu teknik data *mining*, yaitu metode klasifikasi berupa algoritma *Naïve Bayes* [6]. Penerapan data *mining* dan machine *learning*,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

seperti algoritma *Naïve Bayes*, efektif dalam menganalisis data beasiswa Kartu Indonesia Pintar (KIP). Teknik ini membantu mengidentifikasi pola kelayakan penerima secara objektif dan efisien [7]. Selain itu, metode ini mampu mengatasi masalah data tidak seimbang, yang umum dalam dataset beasiswa [8].

Penelitian sebelumnya telah mengeksplorasi penerapan algoritma *Naïve Bayes* dalam konteks klasifikasi, termasuk dalam pengelolaan data beasiswa, telah menunjukkan efektivitas yang signifikan. Algoritma ini beroperasi dengan menghitung probabilitas dari setiap kelas berdasarkan fitur yang ada, yang memungkinkan prediksi yang akurat mengenai kelayakan penerima beasiswa [9]. Beberapa penelitian sebelumnya telah mengeksplorasi penerapan algoritma *Naïve Bayes* dalam konteks beasiswa dan program pendidikan. Misalnya, penelitian oleh Febri dan Sari [10] menunjukkan bahwa *Naïve Bayes* dapat digunakan dalam klasifikasi penerima beasiswa dengan akurasi yang tinggi dengan hasil akurasi sebesar 81% menunjukkan bahwa *Naïve Bayes* dapat memberikan hasil yang baik. Selain itu, penelitian penelitian oleh Wahid [11] menunjukkan bahwa penggunaan algoritma *Naïve Bayes* dalam mengklasifikasikan penerima bantuan sosial menghasilkan akurasi sebesar 82%. Namun, dalam penelitian diatas tidak menggunakan teknik *data balancing* yang membuat data penelitian tersebut menjadi *class imbalance.*

Class imbalance sering terjadi dalam dataset beasiswa, dengan lebih banyak data calon penerima yang tidak layak dibandingkan yang layak, yang menyebabkan model klasifikasi cenderung mengklasifikasikan kelas mayoritas [8]. Untuk mengatasi ini, teknik seperti gabungan metode *Tomek Links* dan *random undersampling* digunakan. *Tomek Links* menghapus data kelas mayoritas yang berdekatan dengan kelas minoritas, sementara *random undersampling* mengurangi jumlah data kelas mayoritas untuk mencapai keseimbangan [12], [13].

Teknik *undersampling* merupakan salah satu metode yang digunakan untuk mengatasi masalah ketidakseimbangan data dalam pengolahan data dan *machine learning.* Dengan menggabungan metode *Tomek Links* dan *Random Undersampling* efektif mengatasi ketidakseimbangan data dalam *machine learning.* Ketidakseimbangan dapat menyebabkan bias terhadap kelas mayoritas, mengurangi kemampuan model dalam mengidentifikasi kelas minoritas. Metode *Tomek Links* menghapus contoh kelas mayoritas yang berdekatan dengan kelas minoritas, memperbaiki batas keputusan dan meningkatkan akurasi model [14].

Selain menggunakan *Tomek Links* dan *ramdom undersampling* untuk menyeimbangkan data, *genetic algorithm (GA)* juga diperlukan untuk mengoptimasi tingkat akurasi dari model *machine learning*. *Genetic Algorithm (GA)* merupakan metode pencarian serta optimasi yang terinspirasi pada proses seleksi alam dan genetika [15]. Dengan menggabungkan prinsip-prinsip seleksi alam dan genetika, *Genetic Algorithm* mampu menemukan solusi yang optimal dalam waktu yang relatif singkat [16]. *Genetic Algorithm* juga sering dipakai untuk menemukan solusi optimal terhadap masalah yang kompleks. Penelitian oleh Religia dan Maulana menunjukkan bahwa *Naïve Bayes* yang dikombinasikan dengan *genetic algorithm* dapat digunakan untuk mengklasifikasikan data yang tidak seimbang, sehingga cocok untuk diterapkan dalam konteks Beasiswa KIP sehingga menghasilkan akurasi sebesar 93% [17]. Pendekatan serupa pada Khotimah [18] mencatat bahwa dengan menggunakan *Genetic Algorithm* untuk pemilihan fitur, akurasi *Naïve Bayes* dapat mencapai 90%. Penelitian ini menunjukkan bahwa penggunaan metode *Genetic Algorithm* dan *Naïve Bayes* dapat meningkatkan hasil klasifikasi secara signifikan. Dengan memanfaatkan *Genetic Algorithm*, kita dapat melakukan seleksi fitur yang lebih baik, sehingga model *Naïve Bayes* dapat lebih fokus pada fitur-fitur yang relevan dan signifikan dalam menentukan kelayakan penerima beasiswa.

Penerapan metode *Genetic Algorithm-Tomek Links* yang dipadukan dengan *Random Undersampling* pada algoritma *Naïve Bayes* untuk mengatasi masalah *class imbalance* dalam data Beasiswa KIP-Kuliah merupakan langkah inovatif yang bertujuan untuk meningkatkan akurasi klasifikasi tanpa melakukan manipulasi data. *Class imbalance* sering kali menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas, sehingga mengurangi kemampuan model dalam mengenali kelas minoritas yang penting. Dalam konteks ini, penggunaan metode *Tomek Links* sebagai teknik *undersampling*, yang dikombinasikan dengan *Random Undersampling*, dapat membantu meningkatkan kualitas data dengan menghapus contoh yang tidak relevan dari kelas mayoritas dan secara acak mengurangi jumlah contoh dari kelas mayoritas untuk mencapai keseimbangan yang lebih baik antara kelas [19].

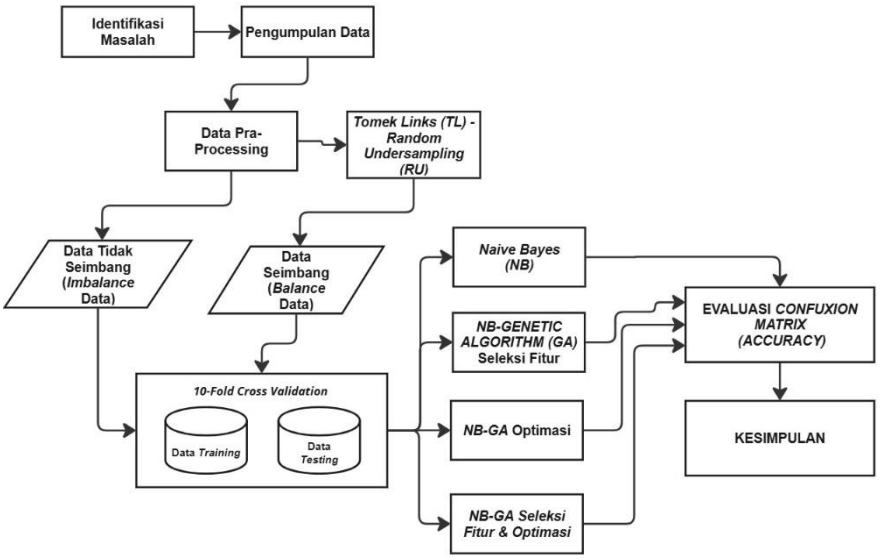
Dengan mempertimbangkan penerapan metode *Genetic Algorithm* dengan gabungan *Tomek Links* dan *random undersampling* pada algoritma *Naïve Bayes* pada data beasiswa Kartu Indonesia Pintar sangat diperlukan. Metode ini tidak hanya dapat membantu dalam mengatasi masalah *class imbalance*, tetapi juga meningkatkan akurasi dan keadilan dalam proses seleksi penerima beasiswa. Dengan pendekatan ini, diharapkan dapat memberikan kontribusi yang signifikan dalam meningkatkan akses pendidikan bagi anak-anak dari keluarga kurang mampu di Indonesia.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Langkah-langkah penelitian disusun secara sistematis untuk memastikan alur penelitian dapat dipahami dengan baik, mulai dari pengumpulan data, praproses data, *balancing* data, hingga tahap evaluasi model. Setiap tahap dijelaskan secara rinci untuk memberikan gambaran yang jelas tentang proses penelitian yang dilakukan. Penelitian dimulai dengan identifikasi masalah dan pengumpulan data, diikuti oleh proses *balancing* menggunakan *Tomek Links* dan *Random Undersampling* untuk mengatasi ketidakseimbangan data. Gambar 1 berikut adalah langkah-langkah yang akan dilakukan:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737



Gambar 1. Prosedur Penelitian

Penelitian ini dimulai dengan identifikasi masalah dan pengumpulan data beasiswa KIP UMKT, di mana data yang terkumpul diproses melalui metode *Tomek Links (TL)* dan *Random Undersampling (RU)* untuk mengatasi ketidakseimbangan data. Data kemudian diklasifikasikan menjadi data seimbang dan tidak seimbang, yang masing-masing divalidasi menggunakan *10-Fold Cross Validation* dengan pembagian *data training* dan *data testing*. Beberapa pendekatan digunakan untuk klasifikasi, yaitu algoritma *Naive Bayes (NB)*, *NB* dengan seleksi fitur menggunakan *Genetic Algorithm (NB-GA)*, *NB* yang dioptimalkan dengan *GA*, serta kombinasi *NB-GA* untuk seleksi fitur dan optimasi. Hasil dari setiap metode dievaluasi menggunakan matriks *Confusion Matrix* untuk menentukan akurasi, yang kemudian menjadi dasar dalam penarikan kesimpulan.

2.2 Identifikasi Masalah

Langkah awal dalam penelitian ini dimulai dengan mengidentifikasi masalah, yang akan menjadi panduan bagi 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, penelitian ini juga melibatkan studi pustaka untuk menemukan kesenjangan dalam penelitian yang telah ada mengenai 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

Data yang digunakan dalam penelitian ini merupakan data penerima beasiswa Kartu Indonesia Pintar (KIP) yang diperoleh dari Universitas Muhammadiyah Kalimantan Timur (UMKT) untuk periode tahun 2020-2023. Hasil pengumpulan data mencakup berbagai variabel yang berkaitan dengan informasi keluarga penerima beasiswa, seperti pekerjaan dan penghasilan ayah serta ibu, status hidup orang tua, dan jenis beasiswa yang diterima. Terdapat 37 fitur utama yang digunakan dalam analisis ini, di mana setiap fitur berperan dalam memahami dampak kondisi ekonomi dan sosial keluarga terhadap keberhasilan akademis penerima beasiswa.

2.4 Data *Pre-Processing*

Tahap ini meliputi data *integretion*, data *selection*, data *transformation*, data *cleaning*, dan data *balancing* beasiswa KIP dari UMKT. Pada tahap *balancing* data, metode *Tomek Links (TL)* dan *Random Undersampling (RU)* diterapkan untuk mengurangi ketidakseimbangan kelas dalam dataset. Selanjutnya, algoritma *Naïve Bayes (NB)* digunakan untuk klasifikasi, baik tanpa optimasi maupun dengan seleksi fitur menggunakan *Genetic Algorithm (GA)*. Metode ini juga dibandingkan dengan kombinasi *NB-GA* yang mencakup optimasi dan seleksi fitur untuk meningkatkan performa klasifikasi. Data yang digunakan diperoleh dari informasi mengenai beasiswa KIP di Kota Universitas Muhammadiyah Kalimantan Timur (UMKT).

2.5 Pembagian Data *Training* dan Data *Testing*

Tahapan pembagian data dilakukan dengan membagi dataset menjadi dua bagian, yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih model dalam mempelajari pola dan hubungan antara fitur-fitur dalam data, sementara *data testing* bertugas untuk menguji kinerja model setelah proses pelatihan selesai. Dalam penelitian ini, teknik *K-Fold Cross-Validation* akan diterapkan untuk *mengevaluasi* kinerja model algoritma *Naive Bayes*, dengan nilai *k* yang ditetapkan sebesar 10. Teknik ini membagi dataset beasiswa KIP menjadi 10 bagian yang sama besar.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

Setiap bagian secara bergantian akan digunakan sebagai *data testing*, sementara bagian lainnya digunakan sebagai data *training*. Dengan melakukan penelitian yang melibatkan penggunaan *K-fold cross-validation* dengan nilai *k 10* untuk mengevaluasi kinerja model *naïve bayes* dalam memprediksi *attrisi* karyawan. Referensi ini memberikan wawasan tentang penerapan *K-fold cross-validation* dalam konteks dunia nyata [20].

2.6 Permodelan

Metode *Tomek Links* digunakan untuk mengurangi tumpang tindih antara kelas mayoritas dan minoritas pada data yang tidak seimbang, sehingga meningkatkan kualitas data untuk proses klasifikasi. *Random Undersampling* adalah teknik sederhana yang digunakan untuk mengatasi ketidakseimbangan kelas dengan mengurangi jumlah sampel dari kelas mayoritas. Meskipun dapat menyebabkan hilangnya informasi, metode ini sering dikombinasikan dengan teknik lain untuk meningkatkan performa model klasifikasi [21]. *Naïve Bayes* dipilih karena kesederhanaan dan kecepatan dalam klasifikasi data, sementara *Genetic Algorithm (GA)* diterapkan untuk seleksi fitur karena kemampuannya menemukan kombinasi fitur optimal yang meningkatkan akurasi model. Penelitian ini juga akan membandingkan hasil model sebelum dan sesudah penerapan *Tomek Links*. Secara garis besar, langkah-langkah pemodelan yang dilakukan dapat dirangkum di hasil dan pembahasan

2.7 Evaluasi

Confusion matrix adalah alat evaluasi yang penting dalam machine learning, khususnya dalam konteks klasifikasi. *Matrix* ini memberikan gambaran yang jelas tentang kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan label sebenarnya dari data. Dalam *confusion matrix*, terdapat empat komponen utama: *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Masing-masing komponen ini menggambarkan jumlah prediksi yang benar dan salah yang dilakukan oleh model terhadap kelas positif dan negatif [22].

Tabel 1. *Confusion Matrix*

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

Tabel 1 menjelaskan empat komponen utama dalam *confusion matrix*: *True Positive (TP)* dan *True Negative (TN)* menunjukkan prediksi yang benar untuk kelas positif dan negatif, sementara *False Positive (FP)* dan *False Negative (FN)* mencerminkan kesalahan prediksi pada kelas positif dan negatif. Komponen ini digunakan untuk mengevaluasi kinerja model klasifikasi.

Tahap evaluasi adalah langkah krusial yang dilakukan setelah model dibentuk. Pada tahap ini, kinerja model diukur untuk menilai akurasi dan kualitas data pelatihan yang digunakan. Pengujian dilakukan dengan menggunakan teknik *Confusion Matrix. Confusion Matrix* adalah metode yang digunakan untuk menghitung akurasi dalam data *mining* [23]. Dalam penelitian ini, evaluasi yang diterapkan adalah kinerja akurasi *(accuracy).* Berikut adalah rumusnya:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; X \; 100\% \quad (1)$$

3. HASIL DAN PEMBAHASAN

3.1 Hasil Data *Pre-Processing*

Pada tahap ini, dilakukan penyajian hasil *pre-processing* data setelah melalui tahap pengumpulan sebelumnya. Proses ini bertujuan untuk memastikan data siap digunakan dalam tahap pemodelan dengan menghilangkan bagian-bagian yang tidak relevan, sehingga data menjadi lebih bersih dan terstruktur.

a. Data *Integretion*

Data *integration* dilakukan dengan menggabungkan berbagai jenis data terkait penerimaan beasiswa di Universitas Muhammadiyah Kalimantan Timur. Terdapat 37 fitur didalam data tersebut. Pada Tabel 2, akan ditampilkan data awal dari integration sebagai berikut:

Tabel 2. *Integretion* Beasiswa Kip-Kuliah

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*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

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 2 mencakup fitur data Beasiswa KIP, seperti identitas siswa, riwayat sekolah, kondisi sosial-ekonomi, serta informasi orang tua, termasuk penghasilan, pekerjaan, dan fasilitas tempat tinggal. Status pengajuan digunakan untuk menentukan kelayakan penerima beasiswa.

b. Data *Selection*

Tahap Data *Selection* memilih 23 dari 37 fitur awal berdasarkan relevansi terhadap aspek sosial-ekonomi untuk mendukung analisis dan model *machine learning.*

Tabel 3. *Selection* Beasiswa Kip-Kuliah

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*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

No	Fitur	Tipe Data
22	Program Studi	*String*
23	Status Pengajuan	*String*

Pada tabel 3 menyajikan fitur-fitur terkait data Beasiswa KIP-Kuliah, yang mencakup informasi mengenai status, pekerjaan, penghasilan, serta kondisi rumah dan akses layanan yang mempengaruhi kelayakan penerimaan beasiswa.

c. Data *Transformation*

Pada tahap transformasi data, fitur 'Status P3KE' yang semula berformat string (misalnya Diterima, Ditolak) akan dikonversi menjadi nilai numerik (misalnya 0, 1, 2) untuk mempermudah pemrosesan algoritma. Berikut adalah data string yang telah diubah menjadi nilai numerik.

Tabel 4. *Transformation* Beasiswa Kip-Kuliah

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

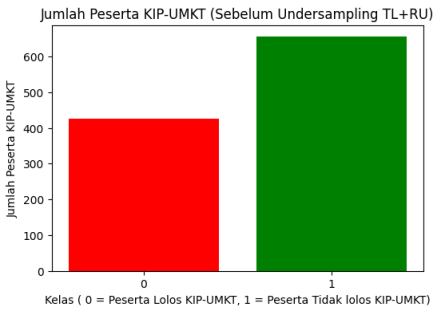
Pada tabel 4 menunjukkan transformasi fitur 'Status Pengajuan' dari format string (Diterima, Ditolak) menjadi nilai numerik (0, 1).

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. Dari 1080 data menjadi 1075 setelah dilakukan data *cleaning.*

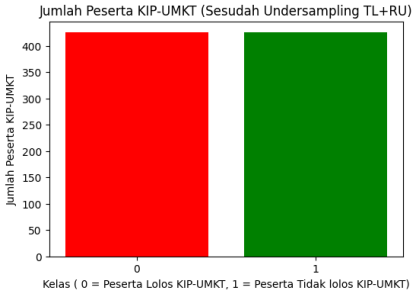
e. Data *Balancing*

Ketidakseimbangan jumlah kelas dapat dilihat dengan jelas pada grafik Gambar 2 , yang menunjukkan perbedaan signifikan antara kelas 0 (peserta tidak lolos) sebanyak 651 dan kelas 1 (peserta lolos) sebanyak 424. Untuk mengatasi ketimpangan ini, digunakan teknik *undersampling* dengan kombinasi metode *Tomek Links* dan Random *Undersampling* untuk menyeimbangkan distribusi jumlah kelas. Berikut adalah gambar 2 sebelum *balancing* dan sesudah dilakukan *balancing.*



Gambar 2. hasil data sebelum *balancing*

Pada gambar 2 menunjukkan distribusi jumlah peserta beasiswa KIP-UMKT sebelum dilakukan proses undersampling menggunakan metode *Tomek Links* dan *Random Undersampling (TL+RU).* Terlihat adanya ketidakseimbangan jumlah antara kelas 0 (peserta lolos KIP-UMKT) yang ditandai dengan warna merah dan berjumlah sekitar 424, dengan kelas 1 (peserta tidak lolos KIP-UMKT) yang ditandai dengan warna hijau dan memiliki jumlah sekitar 651. Ketimpangan ini menekankan perlunya teknik *balancing* data untuk menyamakan jumlah data di kedua kelas agar proses analisis dan pengembangan model *machine learning* lebih optimal.



Gambar 3. hasil data sesudah *balancing*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

Pada gambar 3 menunjukkan bahwa teknik *undersampling Tomek Links-Random Undersampling* berhasil menyeimbangkan distribusi data antara kelas 0 dan kelas 1, sehingga kedua kelas memiliki jumlah yang setara. Sebelum proses *balancing*, kelas 0 memiliki 425 data sementara kelas 1 berjumlah 651. Namun, setelah proses *undersampling* diterapkan, jumlah data pada kedua kelas menjadi sama, masing-masing sebanyak 848. Pada penelitian ini, meskipun data telah diseimbangkan, analisis tetap akan dilakukan pada data yang belum melalui proses *balancing (imbalance* data).

3.2 Hasil Pembagian Data *Training* dan Data *Testing*

Pada penelitian ini, dataset dibagi menggunakan metode 10-*Fold Cross-Validation*, di mana dataset dibagi menjadi sepuluh bagian dengan satu bagian sebagai data testing dan sembilan bagian sebagai data training. Pembagian dilakukan secara acak pada setiap iterasi, memastikan setiap sampel diuji secara merata. Teknik ini membantu mengurangi bias dan variasi dalam evaluasi kinerja model.

3.3 Hasil Permodelan

Hasil pemodelan dan evaluasi akan menyajikan performa algoritma pembelajaran dalam bentuk tingkat akurasi yang dicapai. Analisis ini mencakup model *Naive Bayes* serta model hasil kombinasi lainnya, seperti *Tomek Links-Random Undersampling* dan *Genetic Algorithm*, sebagaimana telah dijelaskan sebelumnya, untuk mengklasifikasikan data beasiswa KIP UMKT.

3.3.1 Permodelan *Naive Bayes*

Permodelan *Naïve Bayes* adalah teknik klasifikasi dalam *machine learning* yang menggunakan *Teorema Bayes* untuk mengukur probabilitas kelas berdasarkan fitur. Metode ini mengasumsikan independensi antar fitur, yang sering memberikan hasil baik, terutama dalam pengolahan bahasa alami dan analisis sentimen.

Tabel 5. hasil rata rata akurasi dari *NB* tanpa *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	630	352	72	21	65.2%

Pada tabel 5 hasil evaluasi permodelan *Naïve Bayes* tanpa penerapan teknik *Tomek Links-Random Undersampling* pada data beasiswa KIP UMKT, diperoleh total nilai evaluasi dari 10-*fold cross-validation*. Total *True Positive (TP)* yang dihasilkan adalah 630, sedangkan *False Positive (FP)* mencapai 352, *True Negative (TN)* sebanyak 72, dan *False Negative (FN)* tercatat 21. Rata-rata akurasi yang diperoleh dari keseluruhan fold adalah sebesar 65.2%.

$$Accuracy = \frac{630+72}{630+352+72+21} \text{ X } 100\% = 65.2\%$$

Hasil ini menunjukkan bahwa akurasi rata-rata model pada data tidak seimbang adalah 65.2%, yang merefleksikan performa model secara keseluruhan dalam mengklasifikasikan data.

Tabel 6. hasil rata rata akurasi *NB* dengan *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	407	352	72	17	56%

Pada tabel 6 hasil evaluasi permodelan *Naïve Bayes* menggunakan *Tomek Links-Random Undersampling (TL-RU)*, total nilai dari proses *10-fold cross-validation* menunjukkan *True Positive (TP)* sebesar 407, *False Positive (FP)* sebesar 352, *True Negative (TN)* sebanyak 72, dan *False Negative (FN)* sebesar 17. Dengan demikian, rata-rata akurasi yang diperoleh adalah 56%.

$$Accuracy = \frac{407+72}{407+352+72+17} \text{ X } 100\% = 56\%$$

Hasil perhitungan akurasi pada permodelan *Naïve Bayes* menggunakan *Tomek Links-Random Undersampling (TL-RU)* menunjukkan bahwa akurasi yang diperoleh adalah 56%.

3.3.2 Permodelan *NB-GA* Seleksi Fitur

Pemodelan *Naive Bayes* yang dipadukan dengan algoritma genetika *(GA)* untuk seleksi fitur efektif dalam menangani dataset tidak seimbang. *GA* mengoptimalkan seleksi fitur, meningkatkan akurasi klasifikasi dengan *Naive Bayes* yang cepat namun rentan terhadap fitur tidak relevan.

Tabel 7. hasil rata rata akurasi dari *NB-GA* seleksi fitur tanpa *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	643	363	61	8	65.4%

Pada tabel 7 hasil evaluasi permodelan *Naïve Bayes-GA* untuk seleksi fitur tanpa penerapan teknik *Tomek Links-Random Undersampling* pada data beasiswa KIP UMKT, diperoleh total nilai evaluasi dari 10-*fold cross-validation*. Total *True Positive (TP)* yang dihasilkan adalah 643, *False Positive (FP)* mencapai 363, *True Negative*

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

(TN) sebanyak 61, dan *False Negative (FN)* tercatat 8. Dengan demikian, rata-rata akurasi yang diperoleh dari keseluruhan *fold* adalah sebesar 65.4%.

$$Accuracy = \frac{643+61}{643+363+61+8} \text{ X } 100\% = 65.4\%$$

Hasil ini menunjukkan bahwa rata-rata akurasi setelah penerapan *Genetic Algorithm* pada seleksi fitur adalah 65.4%, yang mencerminkan peningkatan stabilitas performa model meskipun data tetap tidak seimbang.

Tabel 8. hasil rata rata akurasi *NB-GA* seleksi fitur dengan *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	258	163	261	166	61%

Pada tabel 8 hasil evaluasi permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk seleksi fitur menggunakan *Tomek Links-Random Undersampling (TL-RU)*, total nilai dari proses *10-fold cross-validation* menunjukkan *True Positive (TP)* sebesar 258, *False Positive (FP)* sebesar 163, *True Negative (TN)* sebanyak 261, dan *False Negative (FN)* sebesar 166.

$$Accuracy = \frac{258+261}{258+163+261+166} \text{ X } 100\% = 61\%$$

Hasil perhitungan akurasi pada permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk seleksi fitur menggunakan *Tomek Links-Random Undersampling* menunjukkan bahwa akurasi yang diperoleh adalah 61%.

3.3.3 Permodelan *NB-GA* Optimasi

Pemodelan *Naive Bayes* yang dipadukan dengan *Genetic Algorithm (GA)* efektif untuk meningkatkan performa klasifikasi pada dataset tidak seimbang, dengan *GA* mengoptimalkan seleksi fitur untuk meningkatkan akurasi model.

Tabel 9. hasil rata rata akurasi *NB-GA* optimasi tanpa *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	639	362	62	12	65.2%

Pada tabel 9 hasil evaluasi permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk optimasi algoritma tanpa penerapan teknik *Tomek Links-Random Undersampling* pada data beasiswa KIP UMKT, diperoleh total nilai evaluasi dari 10-*fold cross-validation*. Total *True Positive (TP)* yang dihasilkan adalah 639, *False Positive (FP)* mencapai 362, *True Negative (TN)* sebanyak 62, dan *False Negative (FN)* tercatat 12. Dengan demikian, rata-rata akurasi yang diperoleh dari keseluruhan *fold* adalah sebesar 65.2%.

$$Accuracy = \frac{639+62}{639+363+62+12} \text{ X } 100\% = 65.2\%$$

Hasil perhitungan akurasi pada permodelan *Naïve Bayes-Genetic Algorithm* untuk optimasi algoritma tanpa penerapan teknik *Tomek Links-Random Undersampling* menunjukkan bahwa akurasi yang diperoleh adalah 65.2%.

Tabel 10. hasil rata rata akurasi *NB-GA* optimasi dengan *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	274	166	258	150	62%

Pada tabel 10 hasil evaluasi permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk optimasi algoritma menggunakan *Tomek Links-Random Undersampling (TL-RU)*, total nilai dari proses *10-fold cross-validation* menunjukkan *True Positive (TP)* sebesar 274, *False Positive (FP)* sebesar 166, *True Negative (TN)* sebanyak 258, dan *False Negative (FN)* sebesar 150.

$$Accuracy = \frac{674+166}{274+166+258+150} \text{ X } 100\% = 62\%$$

Hasil perhitungan akurasi pada permodelan *Naïve Bayes-Genetic Algorithm* untuk optimasi algoritma menggunakan *Tomek Links-Random Undersampling* menunjukkan bahwa akurasi yang diperoleh adalah 62%.

3.3.4 Permodelan *NB-GA* Seleksi Fitur dan Optimasi

Pemodelan *Naive Bayes* yang dipadukan dengan *Genetic Algorithm (GA)* untuk seleksi fitur dan optimasi efektif meningkatkan akurasi klasifikasi, terutama pada dataset tidak seimbang, dengan *GA* mengoptimalkan fitur relevan untuk memperbaiki performa model.

Tabel 11. hasil rata rata akurasi *NB-GA* seleksi & optimasi tanpa *TL-RU*

Total Value Each Fold	*TP*	*FP*	*TN*	*FN*	*Average Accuracy*
	644	361	63	7	66%

Pada Tabel 11 hasil evaluasi permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk seleksi fitur dan optimasi algoritma tanpa penerapan teknik *Tomek Links-Random Undersampling (TL-RU)*, diperoleh total nilai

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

evaluasi dari 10-fold cross-validation. Total *True Positive (TP)* adalah 644, *False Positive (FP)* sebesar 361, *True Negative (TN)* sebanyak 63, dan *False Negative (FN)* tercatat 7. Dengan demikian, rata-rata akurasi yang diperoleh dari permodelan ini adalah 66%.

$$Accuracy = \frac{674+166}{274+166+258+150} \text{ X } 100\% = 66\%$$

Akurasi model *Naïve Bayes-Genetic Algorithm (NB-GA)* tanpa *Tomek Links-Random Undersampling (TL-RU)* mencapai 66%.

Tabel 12. hasil rata rata akurasi *NB-GA* seleksi & optimasi dengan *TL-RU*

Total Value Each Fold	***TP***	***FP***	***TN***	***FN***	***Average Accuracy***
	281	165	259	143	63%

Pada Tabel 12 hasil evaluasi permodelan *Naïve Bayes-Genetic Algorithm (NB-GA)* untuk seleksi fitur dan optimasi algoritma dengan *Tomek Links-Random Undersampling (TL-RU),* total nilai dari proses *10-fold cross-validation* menunjukkan *True Positive (TP)* sebesar 281, *False Positive (FP)* sebesar 165, *True Negative (TN)* sebanyak 259, dan *False Negative (FN)* sebesar 143 sehingga rata-rata akurasi yang diperoleh adalah 63%.

$$Accuracy = \frac{281+259}{281+165+259+143} \text{ X } 100\% = 63\%$$

Hasil perhitungan akurasi pada permodelan *Naïve Bayes-Genetic Algorithm* untuk seleksi fitur dan optimasi algoritma dengan *Tomek Links-Random Undersampling* menunjukkan bahwa akurasi yang diperoleh adalah 62%.

3.4 Hasil Perbandingan

Penelitian ini membandingkan hasil evaluasi model *Naïve Bayes (NB)* yang dikombinasikan dengan *Genetic Algorithm (GA)* sebelum dan sesudah penerapan *undersampling* menggunakan *Tomek Links* dan *Random Undersampling.* Fokus utama adalah perbandingan akurasi dan hasil seleksi fitur dari kombinasi *NB*, *GA*, dan optimasi.

Tabel 13. Perbandingan hasil rata rata akurasi *NB* sebelum penggunaan *TL-RU*

NB	***NB-GA* Seleksi Fitur**	***NB-GA* Optimasi**	***NB-GA* Seleksi Fitur & Optimasi**	**Perubahan *NB* ke *NB-GA* Seleksi Fitur**	**Perubahan *NB* ke *NB-GA* Optimasi**	**Perubahan *NB* ke *NB-GA* Seleksi Fitur & Optimasi**
65.2 %	65.4%	65.2%	66%	+0.2	0	+0.8

Tabel 13 menunjukkan kombinasi seleksi fitur dan optimasi *GA* meningkatkan akurasi rata-rata menjadi 66,0%, naik 0,8% dari model *NB* awal, menunjukkan dampak signifikan dibanding penerapan metode secara terpisah.

Tabel 14. Perbandingan hasil rata rata akurasi *NB* sesudah penggunaan *TL-RU*

NB	***NB-GA* Seleksi Fitur**	***NB-GA* Optimasi**	***NB-GA* Seleksi Fitur & Optimasi**	**Perubahan *NB* ke *NB-GA* Seleksi Fitur**	**Perubahan *NB* ke *NB-GA* Optimasi**	**Perubahan *NB* ke *NB-GA* Seleksi Fitur & Optimasi**
56 %	61%	62%	63%	+5	+6	+7

Berdasarkan Tabel 14, rata-rata akurasi model *Naive Bayes (NB)* awal adalah 56%. Seleksi fitur dengan *Genetic Algorithm (GA)* meningkatkan akurasi menjadi 61% (+5%), sementara optimasi *GA* secara terpisah menghasilkan 62% (+6%). Kombinasi seleksi fitur dan optimasi *GA* mencapai akurasi tertinggi 63% (+7%), menunjukkan dampak signifikan GA setelah penyeimbangan data dengan *Tomek Links-Random Undersampling.*

3.5 Pembahasan

Penelitian ini berfokus pada klasifikasi data penerima beasiswa KIP Kuliah di Universitas Muhammadiyah Kalimantan Timur, dengan menggunakan data yang diperoleh dari Kemahasiswaan UMKT selama periode 2021-2023. Tahapan penelitian mencakup identifikasi masalah, pengumpulan data, pra-pemrosesan data, hingga penerapan model *machine learning*. Dalam prosesnya, penelitian ini mengimplementasikan dua jenis model: model pertama menggunakan teknik *undersampling* dengan metode *Tomek Links-Random Undersampling*, sementara model kedua tanpa menggunakan teknik tersebut. Kedua model ini dibandingkan berdasarkan hasil evaluasinya untuk menentukan pendekatan yang paling efektif dalam klasifikasi data beasiswa KIP.

Pada tahap akhir, hasil yang diperoleh dianalisis dan dikaitkan kembali dengan rumusan masalah yang telah ditentukan. Penelitian ini secara khusus bertujuan untuk menjawab pertanyaan terkait seberapa besar peningkatan akurasi yang dapat dicapai oleh algoritma *Naïve Bayes* dalam mengklasifikasikan data beasiswa KIP dengan penerapan metode gabungan *Tomek Links-Random Undersampling* untuk menangani ketidakseimbangan data, serta *Genetic Algorithm (GA)* untuk seleksi fitur dan optimasi. Pembahasan selanjutnya disusun berdasarkan rumusan masalah yang telah dirumuskan.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

3.6 Hasil Peningkatan Akurasi

Peningkatan akurasi yang terlihat setelah penggunaan seleksi fitur dengan *Genetic Algorithm (GA)* dan optimasi fitur dapat dijelaskan lebih rinci dengan memeriksa perbandingan hasil model *Naive Bayes (NB)* pada dataset yang telah diterapkan dengan dan tanpa teknik *balancing* data *(Tomek Links-Random Undersampling)*. Proses ini menunjukkan bagaimana seleksi fitur dan optimasi berperan dalam meningkatkan akurasi model.

Pada awalnya, model *Naive Bayes (NB)* yang tidak menggunakan teknik seleksi fitur atau optimasi hanya menghasilkan akurasi rata-rata sebesar 65.2%. Model ini beroperasi pada data asli tanpa adanya penanganan masalah ketidakseimbangan kelas *(class imbalance)*, yang kemungkinan mengurangi akurasi pada kelas minoritas. Saat fitur difilter menggunakan *Genetic Algorithm (NB-GA* Seleksi Fitur), akurasi hanya sedikit meningkat menjadi 65.4%, yang menunjukkan peningkatan hanya sebesar 0.2%. Peningkatan ini relatif kecil, karena meskipun *GA* membantu dalam memilih fitur yang lebih relevan, namun masalah ketidakseimbangan kelas yang ada pada data asli belum sepenuhnya diatasi.

Dengan menggunakan *NB-GA* Optimasi, akurasi bertahan pada 65.2%, yang berarti tidak ada peningkatan signifikan. Ini menunjukkan bahwa meskipun optimasi diterapkan, tidak ada perubahan substansial dalam performa model yang diakibatkan oleh ketidakseimbangan data yang cukup signifikan. Namun, ketika seleksi fitur dan optimasi digabungkan (*NB-GA* Seleksi Fitur & Optimasi), akurasi rata-rata meningkat menjadi 66%, memberikan peningkatan sebesar 0.8% dibandingkan dengan model *NB* awal. Meskipun ada peningkatan, hasil ini tidak terlalu mencolok, mengingat permasalahan ketidakseimbangan kelas dalam data yang digunakan.

Setelah penerapan *Tomek Links-Random Undersampling*, yang bertujuan untuk mengurangi data yang tidak relevan serta mengatasi ketidakseimbangan kelas, peningkatan akurasi menjadi lebih signifikan. Sebagai contoh, model *NB-GA* Seleksi Fitur mencapai akurasi rata-rata 61%, memberikan peningkatan sebesar 5% dibandingkan model *NB* yang hanya memperoleh 56% pada *fold* pertama. Pada *fold* lainnya, peningkatan juga signifikan, misalnya pada fold 2, model *NB-GA* Seleksi Fitur memperoleh 68%, yang lebih tinggi 12% dibandingkan dengan model *NB* yang hanya memperoleh 56%.

Perubahan yang lebih besar tercatat ketika optimasi dengan *GA* diterapkan pada model *NB-GA* Optimasi. Pada fold pertama, akurasi model meningkat menjadi 65%, yang menunjukkan peningkatan sebesar 7% dibandingkan model *NB*. Peningkatan akurasi terus berlanjut di berbagai fold, terutama pada fold 5, di mana *NB-GA* Optimasi mencatatkan akurasi 72%, lebih tinggi 16% dibandingkan dengan model *NB*. Hal ini menunjukkan bahwa dengan optimasi, model dapat menyesuaikan lebih baik dengan data yang lebih terimbangi setelah penerapan *Tomek Links-Random Undersampling*.

Ketika seleksi fitur dan optimasi *GA* digabungkan (*NB-GA* Seleksi Fitur & Optimasi), akurasi rata-rata model meningkat menjadi 63%, yang merupakan peningkatan sebesar 7% dibandingkan dengan model *NB* yang hanya mencapai 56%. Hasil terbaik tercatat pada fold 5, dengan akurasi mencapai 72%, yang memberikan peningkatan 16% dibandingkan model dasar *(NB)*. Ini menunjukkan bahwa dengan seleksi fitur yang tepat dan optimasi yang lebih canggih, model menjadi lebih efisien dalam menangani data yang telah disesuaikan (setelah *balancing* data). Secara keseluruhan, setelah penerapan *Tomek Links-Random Undersampling*, peningkatan akurasi model menjadi sangat signifikan. Model yang menggunakan seleksi fitur dengan *GA* menunjukkan peningkatan rata-rata sebesar 5% (dari 56% menjadi 61%), sedangkan model yang menggunakan optimasi *GA* memperoleh peningkatan 6% (dari 56% menjadi 62%). Yang lebih penting lagi, gabungan antara seleksi fitur dan optimasi *GA* menghasilkan peningkatan akurasi sebesar 7% (dari 56% menjadi 63%).

4. KESIMPULAN

Penelitian ini mengeksplorasi penerapan kombinasi metode *Genetic Algorithm (GA)* dan *Tomek Links-Random Undersampling (TL-RU)* pada algoritma *Naïve Bayes* untuk mengatasi masalah *class imbalance* dalam data beasiswa Kartu Indonesia Pintar (KIP). Dengan fokus pada Universitas Muhammadiyah Kalimantan Timur, penelitian ini mencakup proses data *pre-processing* yang melibatkan integrasi data, seleksi fitur, transformasi, pembersihan, serta *balancing* data. Hasil penelitian menunjukkan peningkatan akurasi secara signifikan dalam model *Naïve Bayes*. Tanpa *TL-RU*, akurasi meningkat dari 65,2% menjadi 66% setelah proses seleksi fitur dan optimasi dengan *GA*. Dengan penerapan *TL-RU*, model berhasil meningkatkan akurasi dari 56% menjadi 63%. Pendekatan gabungan ini terbukti mampu meningkatkan keadilan distribusi dalam klasifikasi data, memperjelas batas antara kelas mayoritas dan minoritas, serta mengurangi bias model terhadap kelas dominan. Selain itu, penelitian ini menunjukkan bahwa teknik *Genetic Algorithm* tidak hanya membantu dalam seleksi fitur yang relevan tetapi juga meningkatkan akurasi dengan menemukan kombinasi optimal dari fitur yang memengaruhi hasil klasifikasi. Teknik *10-fold cross-validation* digunakan untuk mengevaluasi kinerja model dengan metode *confusion matrix* untuk pengukuran akurasi. Secara keseluruhan, pendekatan ini memberikan solusi inovatif untuk masalah ketidakseimbangan kelas yang kerap ditemukan dalam analisis data beasiswa, mendukung pengambilan keputusan berbasis data yang lebih akurat, dan berpotensi diterapkan pada data serupa untuk mendukung program-program pendidikan lainnya.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2259−2269
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6737

REFERENCES

[1] W. D. Yuniarti, L. Z. Damayanti, and S. Nur'aini, "Sistem Pendukung Keputusan Penerima Bantuan Kartu Indonesia Pintar dengan Metode Weighted Product," *J. Transform.*, vol. 20, no. 2, p. 92, 2023, doi: 10.26623/transformatika.v20i2.5877.

[2] P. Sam *et al.*, "Implementasi Pendukung Keputusan Metode Saw Untuk Penerimaan Kip," Dj djtechno, vol. 5, no. 2, pp. 391–401, 2024, doi: 10.46576/djtechno.

[3] B. G. Dimmera and P. D. P. Purnasari, "Permasalahan Dan Solusi Program Indonesia Pintar Dalam Mewujudkan Pemerataan Pendidikan Di Kabupaten Bengkayang," *Sebatik*, vol. 24, no. 2, pp. 307–314, 2020, doi: 10.46984/sebatik.v24i2.1137.

[4] F. A. Nikmah, N. T. Wardani, and N. Matsani, "Apakah Kartu Indonesia Pintar Berhasil Menurunkan Angka Putus Sekolah?," *J. Komun. Pendidik.*, vol. 4, no. 2, p. 72, 2020, doi: 10.32585/jkp.v4i2.581.

[5] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access*, vol. 10, no. June, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[6] A. Karima and T. A. Y. Siswa, "Prediksi Kinerja Mahasiswa Dalam Perkuliahan Berbasis Learning Management System Menggunakan Algoritma Naïve Bayes," *Progresif J. Ilm. Komput.*, vol. 18, no. 2, p. 211, 2022, doi: 10.35889/progresif.v18i2.922.

[7] D. B. Siswanto and D. Normawati, "Sistem Klasifikasi Monitoring dan Evaluasi Kelayakan Penerima Beasiswa UAD Menggunakan Algoritma Naïve Bayes," *J. Saintekom*, vol. 13, no. 2, pp. 161–172, 2023, doi: 10.33020/saintekom.v13i2.428.

[8] T. A. Zuraiyah, M. M. Mulyati, and G. H. F. Harahap, "Perbandingan Metode Naïve Bayes, Support Vector Machine Dan Recurrent Neural Network Pada Analisis Sentimen Ulasan Produk E-Commerce," *Multitek Indones.*, vol. 17, no. 1, pp. 27–43, 2023, doi: 10.24269/mtkind.v17i1.7092.

[9] A. U. Kurnia, A. S. Budi, and P. H. Susilo, "Sistem Pendukung Keputusan Penerimaan Beasiswa Menggunakan Metode Naive Bayes," *Joutica*, vol. 5, no. 2, p. 397, 2020, doi: 10.30736/jti.v5i2.484.

[10] F. M. Febri and D. P. Sari, "Determination of Bank Indonesia Scholarship Recipients Using Naïve Bayes Classifier," *Barekeng J. Ilmu Mat. dan Terap.*, vol. 17, no. 3, pp. 1595–1604, 2023, doi: 10.30598/barekengvol17iss3pp1595-1604.

[11] A. Wahid, F. Azim, and F. Firdausi, "Application of Data Mining to Classify Receiving Social Assistance Using the Naïve Bayes Method," *Insid. - J. Sist. Inform. Cerdas*, vol. 1, no. 2, pp. 62–66, 2023, doi: 10.31967/inside.v1i2.881.

[12] E. A. Alabdulqader *et al.*, "Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted convolutional neural network," *Sci. Rep.*, vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-65315-7.

[13] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop Oversampling for Class Imbalance Learning: A Review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022, doi: 10.1109/Accecs.2022.3169512.

[14] X. Liu, L. Guo, H. Wang, J. Guo, S. Yang, and L. Duan, "Research on imbalance machine learning methods for MR T1 WI soft tissue sarcoma data," *BMC Med. Imaging*, vol. 22, no. 1, pp. 1–13, 2022, doi: 10.1186/s12880-022-00876-5.

[15] H. Ardiansyah and M. B. S. Junianto, "Penerapan Algoritma Genetika untuk Penjadwalan Mata Pelajaran," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 329, 2022, doi: 10.30865/mib.v6i1.3418.

[16] S. Katoch, S. S. Chauhan, and V. Kumar, "A Review on Genetic Algorithm: Past, Present, and Future," Multimed. Tools Appl., vol. 80, 2021, doi: 10. 1007/s11042-020-10139-6.

[17] Y. Religia and D. Maulana, "Genetic Algorithm Optimization on Nave Bayes for Airline Customer Satisfaction Classification," *JISA(Jurnal Inform. dan Sains)*, vol. 4, no. 2, pp. 121–126, 2021, doi: 10.31326/jisa.v4i2.925.

[18] B. K. Khotimah, M. Miswanto, and H. Suprajitno, "Optimization of feature selection using genetic algorithm in naïve Bayes classification for incomplete data," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, pp. 334–343, 2020, doi: 10.22266/ijies2020.0229.31.

[19] I. S. Ramadhan and A. Salam, "Teknik Random Undersampling untuk Mengatasi Ketidakseimbangan Kelas pada CT Scan Kista Ginjal," *Techno.Com*, vol. 23, no. 1, pp. 20–28, 2024, doi: 10.62411/tc.v23i1.9738.

[20] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, "Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.

[21] A. F. Watratan, A. P. B, D. and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," "J. Appl. Comput. Sci. Technol." vol. 1, no. 1, pp. 7–14, 2020.

[22] K. Adil, A. Ahmed, and M. Essaid, "Fire prediction using Machine Learning Algorithms based on the confusion matrix," pp. 1–11, 2023, [Online]. Available: https://doi.org/10.21203/rs.3.rs-3215936/v1

[23] B. P. Pratiwi and A. Silvia, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi WSN Menggunakan Confusion Matrix," vol. 6, no. 2, pp. 66–75, 2020, doi: 10.26877/jiu.v6i2.6552.