Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

Penerapan Metode GA-NM Pada Algoritma SVM Untuk Mengatasi Class Imbalance Data Beasiswa KIP-Kuliah

Irfan Fiqry Abror, Taghfirul Azhima Yoga Siswa*, Rudiman

Fakultas Sains dan Teknologi, Prodi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
Email: [1]2011102441082@umkt.ac.id, [2,*]tay758@umkt.ac.id, [3]rud959@umkt.ac.id,
Email Penulis Korespondensi: tay758@umkt.ac.id
Submitted: **16/01/2025**; Accepted: **26/02/2025**; Published**: 01/03/2025**

Abstrak−Masalah ketidakseimbangan kelas (class imbalance) sering menjadi tantangan dalam analisis data, terutama ketika jumlah data pada kelas mayoritas jauh lebih banyak dibandingkan dengan kelas minoritas. Hal ini dapat menyebabkan model klasifikasi lebih condong memprediksi kelas mayoritas, sehingga akurasi dalam mengidentifikasi kelas minoritas menjadi rendah. Dalam penelitian ini, metode Support Vector Machine (SVM) dipadukan dengan Near Miss dan Genetic Algorithm (GA) digunakan untuk mengatasi masalah ketidakseimbangan kelas pada data penerima beasiswa Kartu Indonesia Pintar (KIP) di Universitas Muhammadiyah Kalimantan Timur. Data yang digunakan terdiri dari 1075 record dengan 27 fitur yang merepresentasikan faktor sosial-ekonomi penerima beasiswa. Near Miss diterapkan untuk melakukan undersampling terhadap data mayoritas, sehingga menghasilkan distribusi data yang lebih seimbang. Selanjutnya, algoritma SVM digunakan sebagai model klasifikasi utama, dengan seleksi fitur dan parameter yang dioptimalkan menggunakan GA. Hasil penelitian menunjukkan bahwa kombinasi SVM, Near Miss, dan GA dapat meningkatkan performa klasifikasi dalam mengidentifikasi kelas minoritas. Akurasi awal yang didapat tanpa metode tersebut sebesar 60,55% dan setelah diterapkan meningkat jadi 76,88%. Pendekatan ini tidak hanya meningkatkan akurasi model secara keseluruhan tetapi juga menghasilkan performa yang lebih stabil, terutama pada kelas minoritas. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem seleksi penerima beasiswa yang lebih akurat dan efisien, serta menjadi referensi bagi penelitian di bidang data mining dan machine learning.

Kata Kunci: Klasifikasi Beasiswa; Class Imbalance; SVM; GA

Abstract−Class imbalance is a common challenge in data analysis, especially when the number of instances in the majority class significantly exceeds that in the minority class. This imbalance can cause classification models to favor the majority class, resulting in low accuracy in identifying the minority class. In this study, the Support Vector Machine (SVM) method combined with Near Miss and Genetic Algorithm (GA) is used to address the class imbalance problem in the scholarship recipient data of the Kartu Indonesia Pintar (KIP) program at Universitas Muhammadiyah Kalimantan Timur. The dataset consists of 1,075 records with 27 features representing the socio-economic factors of the scholarship recipients. Near Miss was applied to undersample the majority class, producing a more balanced data distribution. Subsequently, the SVM algorithm was utilized as the primary classification model, with feature selection and parameter optimization conducted using GA. The results indicate that the combination of SVM, Near Miss, and GA improved classification performance in identifying the minority class. The initial accuracy obtained without the method was 60.55% and after implementation it increased to 76.88%. This approach not only enhances the overall accuracy of the model but also ensures more stable performance, particularly for the minority class. Therefore, this study is expected to provide a significant contribution to the development of a more accurate and efficient scholarship selection system, as well as serve as a reference for future research in data mining and machine learning.

Keywords: Classification Scholarship; Class Imbalance; SVM; GA

1. PENDAHULUAN

Beasiswa Kartu Indonesia Pintar (KIP) adalah program bantuan pendidikan dari pemerintah untuk pelajar berprestasi dari keluarga kurang mampu, mencakup berbagai jenjang pendidikan. Universitas Muhammadiyah Kalimantan Timur (UMKT) adalah salah satu perguruan tinggi di kota Samarinda yang menjalankan program beasiswa KIP-Kuliah. Berdasarkan wawancara dengan bagian kemahasiswaan, proses seleksi beasiswa KIP-Kuliah dilakukan dengan cara manual dengan mempertimbangkan beberapa kriteria seperti kondisi ekonomi calon mahasiswa. Banyaknya calon pendaftar beasiswa juga menjadi tantangan dalam melakukan seleksi secara cepat dan akurat. Oleh karena itu dibutuhkan model yang tepat dengan mempertimbangkan fitur-fitur representatif agar mahasiswa penerima KIP-Kuliah tepat sasaran [1].

Penerapan machine learning dan data mining dalam bidang pendidikan telah menunjukkan potensi besar untuk mengoptimalkan proses seleksi penerima beasiswa. Penelitian yang dilakukan oleh Bachtiar dan Suyono pada tahun 2021, menggunakan algoritma K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) untuk memprediksi penerima beasiswa keringanan UKT. Hasilnya menunjukkan bahwa metode KNN memiliki akurasi 92,92%, sedangkan metode SVM mencapai akurasi 85,84% [2]. Penelitian serupa yang dilakukan oleh Nata dan Royal pada tahun 2023, membandingkan kinerja berbagai algoritma seperti Naïve Bayes, KNN, SVM, Boosting, dan Decision Tree Random Forest untuk seleksi beasiswa. Hasil terbaik dicapai oleh Naïve Bayes dengan akurasi sebesar 93% [3]. Namun, dalam kedua penelitian tersebut, dataset yang digunakan memiliki masalah ketidakseimbangan data (class imbalance), yang dapat menyebabkan algoritma cenderung memberikan prediksi yang bias terhadap kelas mayoritas. SVM adalah salah satu algoritma pembelajaran mesin yang efektif untuk klasifikasi dan regresi. SVM bekerja dengan mencari hyperplane optimal yang dapat memisahkan kelas-kelas yang berbeda dalam dataset. Algoritma ini memanfaatkan kernel trick untuk memetakan data dari ruang input ke ruang fitur berdimensi tinggi, sehingga dapat

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

menangani masalah klasifikasi yang bersifat non-linier. Kelebihan SVM meliputi kemampuannya dalam bekerja dengan baik pada data berdimensi tinggi, fleksibilitas penggunaannya untuk tugas klasifikasi dan regresi, serta ketahanannya terhadap overfitting dengan penyetelan parameter yang tepat [4]

Class imbalance sering menjadi masalah umum dalam machine learning ketika jumlah kelas mayoritas jauh lebih banyak dibandingkan kelas minoritas. Hal ini dapat menurunkan sensitivitas model terhadap kelas minoritas [5]. Teknik resampling seperti undersampling dan oversampling adalah metode yang efektif untuk mengatasi masalah ini [6]. Dalam dataset KIP-Kuliah UMKT yang digunakan dalam penelitian ini, terdapat ketidakseimbangan kelas yang signifikan, di mana jumlah data kelas mayoritas mencapai 80% dari total data, sementara kelas minoritas hanya sebesar 20%. Ketidakseimbangan ini menciptakan tantangan besar dalam membangun model klasifikasi yang mampu secara akurat memprediksi pendaftar dari kelas minoritas, Teknik Near Miss, misalnya, merupakan salah satu teknik undersampling yang efektif untuk menangani masalah class imbalance pada data klasifikasi. Teknik ini memilih sampel dari kelas mayoritas yang paling dekat dengan kelas minoritas, sehingga model dapat lebih fokus pada prediksi kelas minoritas. Penelitian yang dilakukan oleh Werner de vargas dan Schneider Aranda pada tahun 2023, menunjukkan bahwa Near Miss mampu meningkatkan performa algoritma machine learning dalam memprediksi kelas minoritas dibandingkan dengan metode balancing data lainnya seperti random oversampling dan SMOTE [7]. Selain itu, penelitian serupa yang dilakukan oleh Liashenko dan Kravets pada tahun 2023, menyebutkan bahwa Near Miss memberikan keunggulan dalam mengurangi bias model terhadap kelas mayoritas, yang sering kali menjadi kendala pada data tidak seimbang. Dalam konteks seleksi penerima beasiswa, Near Miss dapat membantu memastikan bahwa model tidak hanya memprioritaskan pendaftar dari kelas mayoritas, tetapi juga mampu memberikan prediksi akurat untuk pendaftar dari kelompok minoritas [8].

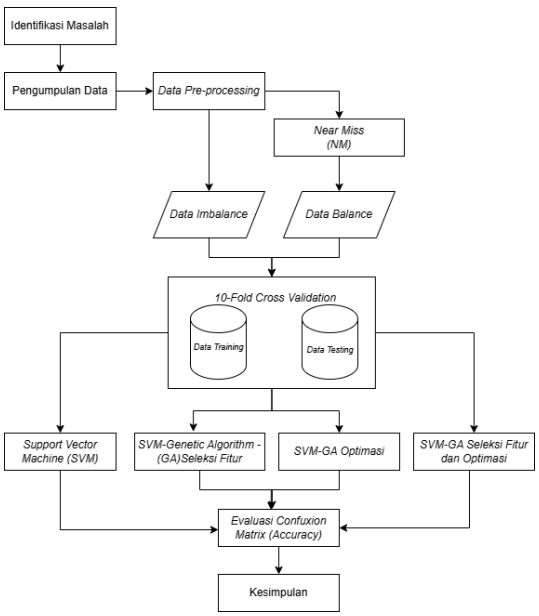
Genetic Algorithm (GA) sering digunakan untuk pemilihan fitur yang relevan, sehingga mengurangi beban komputasi dan meningkatkan performa model. Penelitian yang dilakukan oleh Kocyigit dan Korkmaz pada tahun 2024, menunjukkan bahwa penerapan GA pada deteksi phishing berhasil meningkatkan akurasi model dengan fokus pada fitur paling relevan [9]. Sementara itu, penelitian serupa yang dilakukan oleh Werner de Vargas dan Schneider Aranda pada tahun 2023, mengungkapkan bahwa teknik undersampling Near Miss dapat meningkatkan efektivitas algoritma machine learning pada dataset tidak seimbang, menghasilkan performa yang lebih baik dibandingkan metode random oversampling dan SMOTE [7]. Dalam kombinasi dengan GA, Near Miss memiliki potensi besar untuk meningkatkan akurasi model dengan menyelesaikan masalah class imbalance [10].

Berdasarkan studi literatur yang dilakukan, belum ada peneliti yang mencoba kombinasi algoritma SVM dengan Teknik undersampling Near Miss dan metode optimasi GA untuk seleksi fitur dan optimasi algoritma SVM dalam menangani masalah klasifikasi class imbalance data beasiswa KIP-Kuliah UMKT. Oleh karena itu, penelitian ini bersifat baru diharapkan dapat memberikan kontribusi dalam meningkatkan akurasi klasifikasi dengan mengatasi kendala yang ada pada dataset yaitu class imbalance data beasiswa KIP-Kuliah UMKT.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Setiap penelitian memiliki beberapa tahapan dalam pelakasanaanya. Adapun tahapan yang akan dilakukan seperti pada Gambar 1 Prosedur Penelitian, setiap tahapan akan dijelaskan pada subbab berikutnya.



Gambar 1. Prosedur Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

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 Near Miss (NM). 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 Support Vector Machine (SVM). Berikutnya permodelan kedua diterapkan dengan algoritma SVM ditambah dengan Genetic Algorithm (GA) untuk seleksi fitur. Lalu permodelan ketiga menerapkan NM ditambah dengan GA untuk Optimasi, dan yang terakhir Permodelan menggunakan NM ditambah GA untuk seleksi fitur dan optimasi. Setelah itu didapatkanlah hasil akurasi confuxion matrix dari keempat permodelan tersebut.

2.2 Identifikasi Masalah

Identifikasi masalah merupakan langkah pertama dalam penelitian karena ini akan menjadi dasar bagi seluruh proses penelitian. Masalah utama yang diangkat dalam studi ini berkaitan dengan objek penelitian, yaitu menentukan metode paling efektif untuk mengatasi class imbalance pada klasifikasi data beasiswa KIP-Kuliah UMKT. Selain itu, penelitian ini juga melakukan kajian pustaka untuk mengidentifikasi celah dalam penelitian yang ada mengenai klasifikasi Beasiswa.

2.3 Pengumpulan Data

Penelitian ini menggunakan data beasiswa KIP-Kuliah tahun 2021 - 2023 yang diperoleh dari Unit Kemahasiswaan UMKT yang berlokasi di Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75124. Data yang terkumpul mencakup 37 fitur yang relevan dan dapat berkontribusi dalam pengklasifikasian beasiswa KIP-Kuliah.

2.4 Data Pre-processing

Data beasiswa KIP-Kuliah yang diperoleh dari Unit Kemahasiswaan UMKT harus melalui proses pengolahan lanjutan sebelum memasuki tahap pemodelan, agar bagian data yang tidak relevan dapat dihilangkan. Tahapan persiapan pengolahan data ini sangat penting untuk menjamin kualitas hasil dalam proses data mining. Tahapan persiapan tersebut meliputi integrasi data, pemilihan data, transformasi data (menggunakan label encoder), pembersihan data, dan penyeimbangan data (menggunakan teknik undersampling Near Miss) 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 dan kohesif, 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 [11].
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 [12].
c. Data Transformation
   Transformasi data adalah proses mengubah data ke dalam format atau skala yang sesuai untuk dianalisis. Salah satu langkah dalam transformasi data adalah mengonversi data kategorikal (berbentuk string) menjadi data numerik, yang dapat dilakukan menggunakan library sklearn.preprocessing dengan fungsi LabelEncoder [11]. Dalam penelitian ini, setiap data non-numerik (berbentuk string) akan diubah menjadi format numerik menggunakan fungsi LabelEncoder dari library sklearn.preprocessing pada Python.
d. Data Cleaning
   Data cleaning adalah proses untuk memperbaiki kualitas data dengan menghapus atau memperbaiki data yang salah, tidak lengkap, atau tidak konsisten. Tahapan ini sangat penting untuk memastikan data yang digunakan dalam analisis memiliki tingkat keakuratan yang tinggi, sehingga hasil yang diperoleh dapat dipercaya. Dalam penelitian ini, proses data cleaning dilakukan menggunakan fungsi *dropna()* dari library Pandas. Fungsi ini berfungsi untuk menghapus baris-baris dalam dataset yang mengandung nilai *NaN* atau memiliki nilai yang hilang dalam salah satu atributnya, sehingga data yang diolah menjadi lebih bersih dan siap untuk analisis lebih lanjut [13].
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 di mana jumlah instance dalam kelas "diterima beasiswa" jauh lebih sedikit dibandingkan dengan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

kelas "ditolak beasiswa". Hal ini dapat menyebabkan model machine learning lebih cenderung bias terhadap kelas mayoritas dibandingkan kelas minoritas [14].

2.5 Near Miss

Near Miss adalah teknik yang digunakan untuk mengatasi masalah class imbalance dalam dataset, di mana jumlah contoh dari satu kelas (biasanya kelas mayoritas) jauh lebih banyak dibandingkan dengan kelas lainnya (kelas minoritas). Teknik ini bertujuan untuk mengurangi jumlah contoh dari kelas mayoritas dengan cara memilih contoh yang paling representatif dari kelas tersebut, sehingga meningkatkan kinerja model dalam mengklasifikasikan kelas minoritas. Metode *Near Miss* berfokus pada pemilihan contoh yang berada di sekitar batas antara kelas mayoritas dan minoritas, sehingga contoh yang terpilih adalah yang paling dekat dengan contoh dari kelas minoritas [21]. Adapun rumus Near Miss sebagai berikut:

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{n}(x_{ik}, y_{jk})^2} \quad (1)$$

Pada Rumus 1, $d(x_i, y_j)$ menggambarkan jarak antara dua titik data $x_i$dan y_j yang dihitung menggunakan formula jarak Euclidean. Dalam rumus ini, k merepresentasikan fitur-fitur dari data, sedangkan n adalah jumlah total fitur. Proses ini melibatkan penghitungan selisih kuadrat antara nilai setiap fitur dari dua data tersebut, kemudian menjumlahkan seluruh selisih kuadrat tersebut, dan akhirnya mengambil akar kuadrat hasil penjumlahan untuk mendapatkan jarak akhir. Perhitungan ini digunakan dalam metode Near Miss untuk memilih sampel kelas mayoritas yang paling dekat dengan kelas minoritas, guna menangani masalah ketidakseimbangan data.

2.6 Pembagian Data Training dan Data Testing

Proses pembagian data dilakukan dengan memisahkan dataset menjadi dua bagian, yaitu data training dan data testing. Data training berfungsi untuk melatih model agar dapat memahami pola serta hubungan antar fitur dalam data, sementara data testing digunakan untuk mengevaluasi kinerja model setelah proses training selesai. Teknik K-Fold Cross-Validation akan digunakan untuk menilai performa model machine learning, dengan nilai k yang ditetapkan sebanyak 10. Teknik ini membagi dataset menjadi 10 bagian yang sama besar, di mana setiap bagian secara bergantian dijadikan sebagai data testing, sementara bagian lainnya digunakan sebagai data training. Dengan menghitung rata-rata dari 10 percobaan yang berbeda, 10-Fold Cross-Validation bertujuan untuk memberikan evaluasi kinerja model yang lebih akurat dan terpercaya [15].

2.7 Permodelan

Tahap ini menjelaskan mengenai model yang akan digunakan. Penelitian ini menggunakan model klasifikasi Support Vector Machine (SVM) dengan melakukan balancing data menggunakan Near Miss Undersampling, pembagian data training dan testing menggunakan 10-Fold Cross-Validatin, seleksi fitur dan optimasi algoritma menggunakan menggunakan Genetic Algorithm (GA). Akhir penelitian ini akan membadingkan tanpa dan dengan penggunaan Teknik Undersampling Near Miss. Dan bahasa pemrograman akan digunakan dalam menjalankan permodelan ini adalah bahasa python [16] dengan ekstensi google collab.

2.8 Support Vector Machine

Support Vector Machine (SVM) adalah algoritma machine learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini bertujuan untuk menemukan hyperplane optimal yang memisahkan data berdasarkan kelasnya. Hyperplane ini bertindak sebagai batas pemisah di antara data yang berbeda kelas. SVM mengandalkan dua jenis hyperplane: hyperplane linear dan hyperplane non-linear. Jika data dapat dipisahkan sempurna dengan hyperplane linear, maka disebut sebagai SVM linear. Namun, jika data tidak bisa dipisahkan secara linear, maka SVM menggunakan teknik transformasi kernel untuk mentransfer data ke dimensi fitur yang lebih tinggi sehingga hyperplane linear dapat dibentuk. Melalui berbagai fungsi kernel seperti kernel linear, polinomial, atau Gauss, SVM mampu menangani data dengan karakteristik non-linear secara efektif. Kernel ini berfungsi untuk memproyeksikan data ke ruang berdimensi lebih tinggi guna meningkatkan struktur data, yang mempermudah pemisahannya. Menurut [17] rumus umum SVM linear dapat ditulis sebagai berikut :

$$f(x) = sign(w \cdot x + b) \quad (2)$$

Pada Rumus 2, $f(x) = sign(w \cdot x + b)$ adalah fungsi keputusan yang digunakan dalam Support Vector Machine (SVM) untuk mengklasifikasikan data ke dalam dua kelas. Fungsi ini bekerja dengan menghitung nilai $w \cdot x + b$, di mana w adalah vektor bobot yang menentukan arah hyperplane, x adalah vektor fitur data, dan b adalah bias yang menggeser posisi hyperplane. Hasil fungsi tanda $sign$ akan bernilai +1 jika data berada di satu sisi hyperplane dan -1 jika berada di sisi lainnya. Fungsi ini memastikan bahwa data diklasifikasikan sesuai dengan sisi hyperplane tempatnya berada.

2.9 Genetic Algorithm

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

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 [18]. Adapun rumus menurut [19] sebagai berikut:

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

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

2.10 Evaluasi

Tahap evaluasi merupakan langkah penting setelah pembentukan model. Di tahap ini, performa model diukur untuk mengevaluasi akurasi dan kualitas data latih yang digunakan. Pengujian dilakukan dengan teknik Confusion Matrix. Confusion Matrix adalah sebuah teknik yang digunakan untuk melakukan perhitungan akurasi pada data mining [20]

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, menampilkan performa model diukur guna mengevaluasi tingkat akurasi serta kualitas data latih yang digunakan. Pengujian dilakukan menggunakan teknik Confusion Matrix, yang berfungsi sebagai metode perhitungan akurasi dalam data mining [21] .Evaluasi yang digunakan dalam penelitian ini adalah akurasi model. Rumus yang digunakan sebagai berikut:

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

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. Dan X 100% untuk menjadikan hasil akurasi tersebut menjadi tipe percentage/persen.

3. HASIL DAN PEMBAHASAN

Penelitian ini memanfaatkan dataset beasiswa KIP-Kuliah UMKT yang mencakup data peserta beasiswa antara tahun 2021 hingga 2023. Data tersebut diperoleh dari bagian kemahasiswaan dan mencakup berbagai faktor sosial-ekonomi yang berpengaruh dalam seleksi penerimaan beasiswa. Dari 37 fitur yang diminta sebelumnya, fitur yang berhasil didapat berjumlah 27 fitur dan mencatatkan 1080 entri data. Terdapat dua kelas dalam data ini, yakni kelas diterima dengan 425 entri dan kelas ditolak yang mencatatkan 655 entri. Secara rinci, data yang dikumpulkan mencakup 27 fitur yang diperoleh dari bagian kemahasiswaan. Fitur-fitur tersebut antara lain adalah No. Pendaftaran, Nama Siswa, Status DTKS, Status P3KE, No. KIP, No. KKS, Jenis Kelamin, 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 (KM), Program Studi, Akreditasi Prodi, Status Pengajuan.

3.1 Hasil Pre-processing

Pada tahap ini, akan menampilkan hasil proses dari pre-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
   Setelah memperoleh data beasiswa KIP-Kuliah dari Bagian Kemahasiswaan Universitas Muhammadiyah Kalimantan Timur (UMKT). Untuk mempermudah proses pengolahan, berbagai jenis data beasiswa yang berkaitan dengan penerimaan digabungkan menjadi satu. Informasi penerima KIP tersedia dalam format digital, sedangkan data peserta yang tidak lolos diperoleh manual dari dokumen fisik. Langkah ini dilakukan untuk

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

mengintegrasikan informasi dari berbagai sumber, sehingga menghasilkan dataset yang lebih lengkap dan mendukung analisis secara lebih efektif. Hasilnya akan ditampilkan pada Tabel 2 sebagai berikut:

Tabel 2. Data Integration

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

Pada Tabel 2 data integration, menampilkan fitur yang didapat pada proses awal pre-processing yaitu data integration. Dan total dataset 27.

b. Data Selection

Proses Data Selection bertujuan untuk memilih atribut yang relevan dari dataset awal guna mendukung analisis dan pengembangan model machine learning. Dari total 27 parameter awal, dilakukan seleksi atribut berdasarkan relevansi terhadap tujuan analisis. Melalui seleksi manual yang disesuaikan dengan aspek sosial-ekonomi, menghilangkan data pribadi yang dianggap sensitif dan tidak berpengaruh terhadap hasil klasifikasi. Terpilih 23 parameter utama yang dianggap memiliki pengaruh signifikan terhadap penentuan status pengajuan. Hasilnya akan ditampilkan pada Tabel 3 sebagai berikut:

Tabel 3. 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

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

18	Sumber Air	String
19	MCK	String
20	Jarak Pusat Kota	Integer
21	Akreditas Prodi	String
22	Program Studi	String
23	Status Pengajuan	String

Pada Tabel 3 data selection, 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 [11]. Transformasi ini dilakukan untuk mempermudah algoritma klasifikasi (SVM) dalam perhitungan. Dan semua fitur yang bertipe string akan otomatis di rubah menjadi integer oleh label encoder. Perwakilan fitur data yang akan ditransformasi akan ditampilkan dalam Tabel 4.

Tabel 4. Data Transformation

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

Pada Tabel 4 data transformation, menampilkan fitur yang didapat pada proses ketiga pre-processing yaitu data transformation. Dan tabel diatas adalah perwakilan fitur dataset yang sebelumnya bertipe string lalu diubah secara otomatis menjadi integer dengan label encoder.

d. Data Cleaning

Data cleaning merujuk pada proses untuk menghapus atau memperbaiki data yang salah, tidak lengkap, atau tidak konsisten. Tahapan ini sangat penting untuk memastikan bahwa analisis dan hasil yang diperoleh akurat. Dalam penelitian ini, proses data cleaning dilakukan menggunakan fungsi dropna() dari pustaka pandas, yang berfungsi untuk menghapus baris yang memiliki nilai NaN atau nilai yang hilang pada satu atau lebih kolom. Setelah proses ini, data beasiswa KIP-Kuliah mengalami perubahan jumlah. Data berkurang dari 1080 menjadi 1075. Hasil nya akan ditampilkan pada Gambar 2 sebagai berikut:

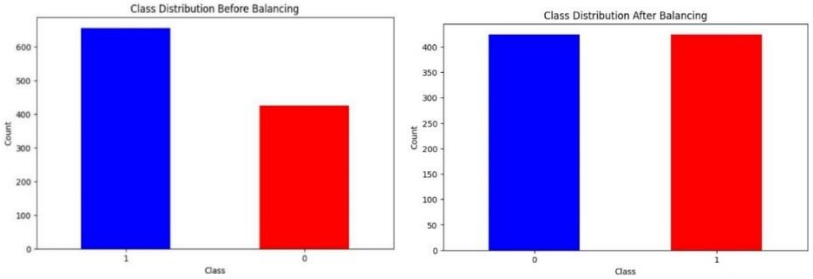
```
Data Awal:
Jumlah Baris dan Kolom Sebelum Cleaning: (1080, 23)

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

Gambar 2. Hasil Data Cleaning

Pada Gambar 2 data cleaning, 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

Ketidakseimbangan kelas antara kelas-kelas dapat dilihat pada diagram Gambar 3, di mana barplot di sebelah kiri menunjukkan ketidakseimbangan jumlah antara kelas 1 (tidak lolos) yang berjumlah 654 dan kelas 0 (lolos) yang berjumlah 424. Untuk mengatasi masalah ini, teknik undersampling dengan metode Near Miss digunakan untuk mengatasi masalah class imbalance dalam dataset, di mana jumlah contoh dari satu kelas (biasanya kelas mayoritas) jauh lebih banyak dibandingkan dengan kelas lainnya (kelas minoritas). Metode Near Miss berfokus pada pemilihan contoh yang berada di sekitar batas antara kelas mayoritas dan minoritas, sehingga contoh yang terpilih adalah yang paling dekat dengan contoh dari kelas minoritas [22] Dan tampilan data sesudah di seimbangkan menggunakan NM akan ditampilkan pada gambar 4. Dimana dataset berubah menjadi total 848 dengan jumlah antara kelas 1 (tidak lolos) yang berjumlah 424 dan kelas 0 (lolos) yang berjumlah 424.



Gambar 3. Hasil Sebelum dan Sesudah Data Balancing

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

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

3.2 Hasil Pembagian Data Training dan Data Testing

Dalam penelitian ini, pembagian *dataset* menjadi *data training* dan *data testing* memiliki peran penting dalam menentukan kinerja model *machine learning*. Seluruh *record* pada dataset dibagi secara proporsional menggunakan metode *10-Fold Cross-Validation*. Dalam proses ini, dataset dibagi menjadi sepuluh bagian yang sama besar, di mana satu bagian digunakan sebagai data testing dan sembilan bagian lainnya sebagai data *training*. Pada setiap *fold*, pembagian data dilakukan secara acak, sehingga setiap bagian memiliki peluang yang sama untuk menjadi data *training* maupun *testing*. Pendekatan ini membantu mengurangi bias dan variasi dalam estimasi kinerja model, serta memastikan bahwa setiap sampel diuji pada model. Proses ini akan dijelaskan lebih rinci pada tahap selanjutnya, yaitu hasil pemodelan dan evaluasi.

3.3 Hasil Permodelan

Pada tahap ini, akan menampilkan hasil pembelajaran algoritma dalam bentuk akurasi yang didapatkan oleh model mulai dari menggunakan algoritma klasifikasi Support Vector Machine (SVM) hingga ditambahkan model-model lainnya seperti undersampling dari Near Miss (NM) dan Genetic Algorithm (GA) seleksi fitur dan 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. Semua tahap ini akan dilakukan pada bagian permodelan sebagai berikut:

3.3.1 Permodelan Support Vector Machine

Algoritma klasifikasi Support Vector Machine (SVM) di terapkan tanpa menggunakan balancing undersampling dari NM dan diterapkan juga menggunakan balancing undersampling dari NM. Dalam artian permodelan ini diterapkan terhadap 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 NM. Hasilnya akan ditampilkan sebagai berikut:

Tabel 5. Hasil Rata-Rata Akurasi SVM Tanpa NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	650	423	1	1	60,55%

$$\text{Accuracy} = \frac{650+1}{650+423+1+1} \text{ X } 100\% = 60{,}55\%$$

Pada Tabel 5, menampilkan hasil rata-rata akurasi SVM tanpa NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 650, fp 423, tn 1, fn 1. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 60,55%.

Tabel 6. Hasil Rata-Rata Akurasi SVM Dengan NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	288	252	172	136	54,24%

$$\text{Accuracy} = \frac{288+172}{288+252+172+136} \text{ X } 100\% = 54{,}24\%$$

Pada Tabel 6, menampilkan hasil rata-rata akurasi SVM dengan NM(Undersampling untuk menyeimbangkan data)(balance). Mendapatkan total tp 288, fp 252, tn 172, fn 136. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 54,24%.

3.3.2 Permodelan SVM-Genetic Algorithm Untuk Seleksi Fitur

Algoritma klasifikasi Support Vector Machine (SVM) ditambah dengan Genetic Algorithm untuk Seleksi Fitur di terapkan tanpa menggunakan balancing undersampling dari NM dan diterapkan juga menggunakan balancing undersampling dari NM. Genetic Algorithm adalah metaheuristik berbasis populasi yang digunakan untuk menyelesaikan berbagai masalah optimasi. GA terinspirasi dari proses evolusi biologis seperti yang diusulkan oleh Charles Darwin, yang melibatkan konsep seleksi alam, rekombinasi (crossover), dan mutasi. Algoritma ini mencari solusi optimal dengan menggunakan sejumlah calon solusi yang disebut populasi, dan melalui iterasi proses seleksi, crossover, dan mutasi, populasi tersebut akan berkembang menuju solusi optimal [23]. Permodelan ini diterapkan terhadap 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 NM. Hasilnya akan ditampilkan sebagai berikut:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

Tabel 7. Hasil Rata-Rata Akurasi SVM-GA Seleksi Fitur Tanpa NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	600	297	127	51	67,62%

$$\text{Accuracy} = \frac{600+127}{127+297+127+51} \text{ X } 100\% = 67{,}62\%$$

Pada Tabel 7, menampilkan hasil rata-rata akurasi SVM tanpa NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 600, fp 297, tn 127, fn 51. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 67,62%.

Tabel 8. Hasil Rata-Rata Akurasi SVM-GA Seleksi Fitur Dengan NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	282	119	305	142	69,22%

$$\text{Accuracy} = \frac{282+305}{282+119+305+142} \text{ X } 100\% = 69{,}22\%$$

Pada Tabel 8, menampilkan hasil rata-rata akurasi SVM dengan NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 282, fp 119, tn 305, fn 142. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1 , sehingga mendapatkan hasil akurasi sebesar 69,22%.

3.3.3 Permodelan SVM-GA Untuk Optimasi

Algoritma klasifikasi Support Vector Machine (SVM) ditambah dengan Genetic Algorithm untuk optimasi. Selain digunakan sebagai seleksi fitur, GA juga dapat digunakan sebagai optimasi untuk algoritma SVM. Proses ini digunakan untuk menemukan parameter terbaik bagi model SVM, yaitu C dan gamma, agar menghasilkan performa terbaik. Permodelan ini diterapkan terhadap 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 NM. Hasilnya akan ditampilkan sebagai berikut:

Tabel 9. Hasil Rata-Rata Akurasi SVM-GA Optimasi Tanpa NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	546	185	239	105	73,02%

$$\text{Accuracy} = \frac{546+239}{546+185+239+105} \text{ X } 100\% = 73{,}02\%$$

Pada Tabel 9, menampilkan hasil rata-rata akurasi SVM tanpa NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 546, fp 185, tn 239, fn 105. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 73,02%.

Tabel 10. Hasil Rata-Rata Akurasi SVM-GA Optimasi Dengan NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	308	131	293	116	70,88%

$$\text{Accuracy} = \frac{308+293}{308+131+293+116} \text{ X } 100\% = 70{,}88\%$$

Pada Tabel 10, menampilkan hasil rata-rata akurasi SVM dengan NM(Undersampling untuk menyeimbangkan data)(balance). Mendapatkan total tp 308, fp 131, tn 293, fn 116. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 70,88%.

3.3.4 Permodelan SVM-Genetic Algorithm Untuk Seleksi Fitur dan Optimasi

Algoritma klasifikasi Support Vector Machine ditambah dengan Genetic Algorithm untuk Seleksi Fitur dan Optimasi di terapkan tanpa menggunakan balancing undersampling dari NM dan diterapkan juga menggunakan balancing undersampling dari NM. 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 NM. Hasilnya akan ditampilkan sebagai berikut:

Tabel 11. Hasil Rata-Rata Akurasi SVM-GA Seleksi Fitur dan Optimasi Tanpa NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	534	163	261	117	73,95%

$$\text{Accuracy} = \frac{534+261}{534+163+261+117} \text{ X } 100\% = 73{,}95\%$$

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

Pada Tabel 11, menampilkan hasil rata-rata akurasi SVM tanpa NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 534, fp 163, tn 261, fn 117. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 73,95%.

Tabel 12. Hasil Rata-Rata Akurasi SVM-GA Seleksi Fitur dan Optimasi Dengan NM

Total Value Each Fold	TP	FP	TN	FN	Average Accuracy
	349	121	303	75	76,88%

$$\text{Accuracy} = \frac{349+303}{349+121+303+75} \text{ X } 100\% = 76{,}88\%$$

Pada Tabel 12, menampilkan hasil rata-rata akurasi SVM dengan NM(Undersampling untuk menyeimbangkan data)(imbalance). Mendapatkan total tp 349, fp 121, tn 303, fn 75. Setelah mendapat semua nilainya lalu dimasukan pada formula pencarian akurasi yang terdapat dalam formula 1, sehingga mendapatkan hasil akurasi sebesar 76,88%.

3.4 Hasil Perbandingan

Hasil akhir dari penelitian ini adalah membandingkan hasil evaluasi penerapan model algoritma klasifikasi SVM yang menggunakan GA untuk seleksi fitur dan optimasi algoritma, serta penerapan teknik undersampling Near Miss untuk menyeimbangkan kelas. Secara keseluruhan, selain teknik undersampling, penelitian ini mengkombinasikan model klasifikasi SVM dengan metode seleksi fitur dan optimasi untuk meningkatkan akurasi. Fokus utama dari penelitian adalah pada perbandingan hasil akurasi dan efektivitas seleksi fitur yang diperoleh dari kombinasi SVM, GA, dan Near Miss, baik sebelum maupun setelah penerapan teknik balancing kelas menggunakan Near Miss.

Tabel 13. Perbandingan Hasil Akurasi Model Sebelum Menggunakan Near Miss (NM) dari masing-masing fold

Fold	SVM	SVM-GA Seleksi Fitur	SVM-GA Optimasi	SVM-GA Seleksi Fitur & Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur	Perubahan SVM ke SVM-GA Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur & Optimasi
1	60,01%	62,96%	70,37%	75,92%	+2,95%	+10,36%	+16,84%
2	60,01%	72,22%	75,92%	79,62%	+12,21%	+15,91%	+20,54%
3	60,01%	66,66%	70,37%	73,14%	+6,65%	+10,36%	+14,99%
4	60,01%	61.11%	72,22%	70,37%	+1,1%	+12,21%	+15,91%
5	61,11%	72,22%	77,77%	74,07%	+11,11%	+16,66%	+13,89%
6	60,74%	68,22%	73,83%	76,63%	+7,48%	+13,09%	+22,57%
7	60,74%	70,00%	71,96%	74,76%	+9,26%	+11,22%	+18,69%
8	60,74%	66,35%	74,76%	74,76%	+5,61%	+14,02%	+14,96%
9	60,74%	69,15%	71,96%	71,96%	+8,41%	+11,22%	+17,76%
10	60,74%	67,28%	71,02%	68,22%	+6,54%	+10,28%	+13,09%

Pada Tabel 13, menunjukkan hasil perbandingan akurasi dari masing-masing model *SVM*, *SVM-GA* Seleksi Fitur, *SVM-GA* Optimasi, dan *SVM-GA* Seleksi Fitur dan Optimasi. Hasil yang diperoleh menunjukkan peningkatan akurasi yang signifikan ketika kombinasi model tersebut diterapkan.

Tabel 14. Perbandingan Hasil Rata-Rata Akurasi Model Sebelum Menggunakan Near Miss (NM)

SVM	SVM-GA Seleksi Fitur	SVM-GA Optimasi	SVM-GA Seleksi Fitur & Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur	Perubahan SVM ke SVM-GA Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur & Optimasi
60,55%	67,62%	73,02%	73,95%	+7,07%	+12,47%	+13,40%

Pada Tabel 14, menunjukkan hasil perbandingan hasil rata-rata akurasi dari masing-masing model *SVM*, *SVM-GA* Seleksi Fitur, *SVM-GA* Optimasi, dan *SVM-GA* Seleksi Fitur dan Optimasi. Hasil peningkatan nya cukup tinggi pada bagian SVM ke SVM-GA yaitu sebesar 13,40%.

Tabel 15. Perbandingan Hasil Akurasi Model Setelah Menggunakan Near Miss (NM) dari masing-masing fold

Fold	SVM	SVM-GA Seleksi Fitur	SVM-GA Optimasi	SVM-GA Seleksi Fitur & Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur	Perubahan SVM ke SVM-GA Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur & Optimasi
1	49,41%	74,11%	71,76%	78,82%	+24,7%	+22,35%	+29,41%
2	52,94%	76,47%	78,82%	83,52%	+23,53%	+25,88%	+30,58%
3	50,58%	56,47%	65,88%	77,64%	+5,89%	+15,3%	+27,06%
4	50,58%	67,05%	68,23%	69,41%	+16,47%	+17,65%	+18,83%
5	64,70%	69,41%	69,41%	75,29%	+4,71%	+4,71%	+10,59%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

6	57,64%	72,94%	76,47%	82,35%	+15,3%	+18,83%	+24,71%
7	56,47%	65,88%	65,88%	77,64%	+9,41%	+9,41%	+21,17%
8	51,76%	68,23%	63,52%	71,76%	+16,47%	+11,76%	20%
9	54,76%	71,42%	72,61%	84,52%	+16,66%	+17,85%	+29,76%
10	53,57%	70,23%	76,19%	67,85%	+16,66%	+22,62%	+14,28%

Pada Tabel 15, menunjukkan hasil perbandingan akurasi setelah balancing dari masing-masing model *SVM*, *SVM-GA* Seleksi Fitur, *SVM-GA* Optimasi, dan *SVM-GA* Seleksi Fitur dan optimasi. Untuk hasil pada setiap fold mengalami peningkatan sangat tinggi di tahap perubahan SVM ke SVM-GA Seleksi Fitur dan Optimasi dominan di angka 27% keatas.

Tabel 16. Perbandingan Hasil Rata-Rata Akurasi Model Setelah Menggunakan Near Miss (NM)

SVM	SVM-GA Seleksi Fitur	SVM-GA Optimasi	SVM-GA Seleksi Fitur & Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur	Perubahan SVM ke SVM-GA Optimasi	Perubahan SVM ke SVM-GA Seleksi Fitur & Optimasi
54,24%	69,22%	70,88%	+76,88%	+14,98%	+18,64%	+22.64%

Pada Tabel 16, menunjukkan hasil perbandingan akurasi setelah balancing dari masing-masing model *SVM*, *SVM-GA* Seleksi Fitur, *SVM-GA* Optimasi, dan *SVM-GA* Seleksi Fitur dan optimasi. Tahap permodelan ini mengalami peningkatan akurasi lebih tinggi dari permodelan sebelumnya.

Tabel 17. Perbandingan Hasil Seleksi Fitur GA

No	Seleksi Fitur (Imbalance)	Seleksi Fitur & Optimasi (Imbalance)	Seleksi Fitur (Balance)	Seleksi Fitur & Optimasi (Imbalance)
1	Status P3KE	Status P3KE	Status P3KE	Status DTKS
2	No. KIP	No. KIP	Jenis Kelamin	Status P3KE
3	Pekerjaan Ayah	No. KKS	Penghasilan Ibu	No. KIP
4	Penghasilan Ayah	Penghasilan Ayah	Status Ibu	No. KKS
5	Penghasilan Ibu	Penghasilan Ibu	Kepemilikan Rumah	Pekerjaan Ayah
6	Kepemilikan Rumah	Tahun Perolehan	Tahun Perolehan	Penghasilan Ibu
7	Tahun Perolehan	Sumber Listrik	Sumber Listrik	Status Ibu
8	Sumber Listrik	Luas Bangunan	Sumber Air	Tahun Perolehan
9	Luas Bangunan	Jarak Pusat Kota	Jarak Pusat Kota	Luas Tanah
10	Jarak Pusat Kota	Akreditasi	-	Jarak Pusat Kota
11	-	-	-	Akreditasi

Pada Tabel 17, menampilkan perbandingan hasil seleksi fitur pada berbagai kondisi, yaitu seleksi fitur biasa dan seleksi fitur yang dioptimasi menggunakan Genetic Algorithm (GA), baik pada dataset yang tidak seimbang (imbalance) maupun yang telah diseimbangkan (balance). Pada seleksi fitur biasa, fitur yang terpilih didasarkan pada relevansi dan signifikansi masing-masing fitur dalam klasifikasi data, tanpa mempertimbangkan interaksi antar fitur. pada dataset yang tidak seimbang (imbalance), fitur yang terpilih meliputi Status P3KE, No. KIP, dan Pekerjaan Ayah, yang menunjukkan beberapa atribut penting terkait kondisi sosial-ekonomi penerima beasiswa. Namun, setelah dilakukan optimasi menggunakan GA, terdapat perubahan signifikan dalam fitur yang terpilih. GA mampu mengoptimalkan pemilihan fitur dengan mempertimbangkan interaksi antar fitur, sehingga menghasilkan subset fitur yang lebih relevan untuk meningkatkan performa model klasifikasi. Misalnya, pada dataset yang tidak seimbang, fitur seperti Status DTKS dan Pekerjaan Ayah muncul sebagai pilihan utama setelah optimasi dengan GA, yang sebelumnya tidak terpilih dalam seleksi fitur biasa. Lalu didapatkanlah seleksi fitur dengan subset fitur yang lebih relevan pada tabel 16 sebagai berikut:

Tabel 17. Evaluasi Hasil Seleksi Fitur GA

No	Hasil Seleksi Fitur
1	Status P3KE
2	Penghasilan Ibu
3	Tahun Perolehan
4	Jarak Pusat Kota

Pada Tabel 17, menunjukan hasil seleksi fitur yang selalu tampil dalam 4 kali percobaan dengan model yang berbeda. Menurut model yang digunakan 4 fitur ini dianggap sangat penting atau relevan untuk klasifikasi class imbalance data beasiswa KIP-Kuliah.

3.5 Pembahasan

Penelitian ini berfokus pada klasifikasi data beasiswa KIP-Kuliah Universitas Muhammadiyah Kalimantan Timur (UMKT) dengan menggunakan data yang dihimpun dari Bagian Kemahasiswaan UMKT selama periode 2021-2023.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

Penelitian ini mencakup beberapa tahap utama, mulai dari identifikasi masalah, pengumpulan data, pemrosesan awal data (data preprocessing), hingga penerapan model machine learning. Dua versi model diterapkan, yakni model pertama dengan teknik undersampling Near Miss dan model kedua tanpa teknik balancing tersebut. Keduanya kemudian dievaluasi untuk menentukan pendekatan yang paling efektif dalam mengklasifikasi data beasiswa KIP-Kuliah UMKT. Pada tahap akhir, penelitian ini mengevaluasi hasil yang diperoleh dan mengaitkannya kembali dengan permasalahan yang dirumuskan. Fokusnya adalah menjawab pertanyaan mengenai fitur-fitur yang terpilih melalui seleksi fitur menggunakan Genetic Algorithm (GA) serta dampaknya terhadap akurasi klasifikasi data beasiswa KIP-Kuliah UMKT. Selain itu, penelitian ini juga membahas sejauh mana peningkatan akurasi model Support Vector Machine (SVM) yang diterapkan dengan metode Near Miss untuk menangani ketidakseimbangan kelas dan seleksi fitur serta optimasi parameter menggunakan GA. Berdasarkan rumusan masalah yang telah ditetapkan, pembahasan hasil penelitian dirangkum sebagai berikut.

3.6 Hasil Peningkatan Akurasi

Hasil penelitian ini menunjukkan bahwa penerapan metode Genetic Algorithm (GA) untuk seleksi fitur dan optimasi algoritma, serta teknik Near Miss untuk menyeimbangkan data, memberikan peningkatan signifikan dalam akurasi model Support Vector Machine (SVM) dalam mengatasi class imbalance pada klasifikasi data Beasiswa KIP-Kuliah Universitas Muhammadiyah Kalimantan Timur. Dalam analisis awal, model dasar SVM tanpa penerapan teknik apapun menunjukkan akurasi sebesar 60,55%. Setelah menerapkan seleksi fitur menggunakan GA, akurasi model meningkat menjadi 67,62%, mencerminkan peningkatan sekitar 7,07%. Penelitian yang dilakukan oleh Dilla dan Wawan pada tahun 2021, melakukan klasifikasi banjir kota Samarinda menggunakan metode SVM dan seleksi fitur GA. Dari hasil pengujian GA meningkatkan akurasi SVM sebesar 13,45%. Sebelum penererapan seleksi fitur GA, SVM hanya mencapai 52,71%, namun setelah penerapan seleksi fitur menggunakan GA, akurasi meningkat menjadi 66,16%. Hal ini menunjukan bahwa penggunaan seleksi fitur menggunakan GA meningkatkan metode SVM .Selanjutnya ketika model dioptimasi menggunakan GA, akurasi meningkat lebih jauh menjadi 73.02% [24]. Penelitian yang dilakukan Pratiwi dan Putra pada tahun 2021, menunjukan peningkatan akurasi SVM sebanyak 0,60% dengan menggunakan optimasi GA. Sebelum melakukan optimasi parameter akurasi yang didapatkan sebesar 93.20% dan setelah menggunakan optimasi GA terjadi peningkatan akurasi menjadi 93.80%. Hal ini menunjukan bahwa optimasi parameter memiliki kontribusi terhadap performa model. Penerapan kombinasi seleksi fitur dan optimasi menghasilkan akurasi tertinggi pada penggunaan data yang belum seimbang (imbalance), yaitu 73.95%, dengan peningkatan sebesar 13,40% dibandingkan model dasar [20].

Ketika teknik Near Miss diterapkan, model dasar SVM menunjukkan akurasi awal sebesar 54,24%. Setelah penerapan seleksi fitur menggunakan GA, akurasi meningkat menjadi 69.22%, menunjukkan peningkatan sebesar 14,98% sama halnya dengan penelitian sebelumnya yang dilakukan oleh Syaputra dan Siswa pada tahun 2024, mengalami peningkatan akurasi sebanyak 25,80% setelah penggunaan GA sebagai seleksi fitur. Sebelum merapkan seleksi fitur akurasi yang didapat sebesar 56,58% dan setelah menggunakan seleksi fitur terjadi peningkatan akurasi menjadi 71,70%. Hal ini menjelaskan setelah melakukan balancing data dengan Near Miss, GA sebagai seleksi fitur tetap meningkat akurasi model. Selanjutnya, ketika model dioptimasi menggunakan GA dan menggunakan data yang sudah balancing akurasi meningkat menjadi 70,88%, yang mencerminkan manfaat positif dari optimasi parameter meskipun peningkatan tidak terlalu besar. Kombinasi seleksi fitur dan optimasi pada data yang telah diseimbangkan menggunakan Near Miss menghasilkan akurasi tertinggi, yaitu 76,88%, dengan peningkatan sebesar 22,64% dibandingkan model dasar. Secara garis besar akurasi model mengalami penurunan akurasi setelah melakukan balancing data menggunakan Near Miss dibandingkan sebelum dilakukan balancing. Namun penggunaan Near Miss dikombinasikan GA untuk seleksi fitur dan optimasi algoritma SVM berhasil mendapatkan akurasi tertinggi sebesar 76,88% [25]. Dari hasil penelitian ini, dapat disimpulkan bahwa penerapan metode GA untuk seleksi fitur dan optimasi algoritma memberikan kontribusi yang signifikan terhadap peningkatan akurasi model dasar. Kombinasi kedua metode tersebut, baik sebelum maupun setelah penerapan Near Miss, menunjukkan bahwa model yang lebih kompleks dan teroptimasi dapat menangani data yang tidak seimbang dengan lebih baik.

4. KESIMPULAN

Berdasarkan penelitian mengenai penerapan metode Genetic Algorithm (GA) yang dikombinasikan dengan teknik Near Miss pada algoritma Support Vector Machine (SVM) untuk mengatasi class imbalance pada data beasiswa KIP-Kuliah UMKT, diperoleh akurasi tertinggi sebesar 76,88%, yang menunjukkan peningkatan signifikan sebesar 16,33% dibandingkan dengan model baseline yang hanya mencapai 60,55%. Peningkatan ini menegaskan efektivitas kombinasi metode dalam menangani class imbalance dan meningkatkan akurasi model. Metode seleksi fitur dan optimasi menggunakan GA secara bersamaan memberikan hasil terbaik, berkontribusi signifikan terhadap akurasi model dibandingkan dengan penerapan metode secara terpisah. Meski demikian, penelitian ini memiliki keterbatasan, seperti penggunaan kernel RBF pada SVM yang mungkin belum optimal untuk semua jenis dataset. Oleh karena itu, penelitian selanjutnya dapat mengeksplorasi kernel SVM lainnya dan menerapkan metode pada dataset yang berbeda untuk melihat potensi peningkatan kinerja. Disarankan pula untuk mengeksplorasi metrik evaluasi tambahan seperti F1-score atau AUC-ROC guna memberikan gambaran yang lebih lengkap tentang performa model.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2358−2370
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6756

REFERENCES

[1] R. Susetyoko, W. Yuwono, and E. Purwantini, "Model Klasifikasi Pada Seleksi Mahasiswa Baru Penerima KIP Kuliah Menggunakan Regresi Logistik Biner," JIP (Jurnal Informatika Polinema), vol. 8, no. 4, 2023, doi: 10.33795/jip.v8i4.914.

[2] M. I. Bachtiar, H. Suyono, M. Fauzan, and E. Purnomo, "75 Method Comparison In The Decision Support System Of A Scholarship Selection," Jurnal Ilmiah KURSOR, vol. 11, no. 2, 2021, doi: 10.21107/kursor.v11i2.263.

[3] A. Nata and S. Royal, "Analisis Sistem Pendukung Keputusan Dengan Model Klasifikasi Berbasis Machine Learning Dalam Penentuan Penerima Program Indonesia Pintar," Journal of Science and Social Research, vol. 3, 2022, doi:10.54314/jssr.v5i3.1041.

[4] M. Wang, J. Yu, M. Zhou, W. Quan, and R. Cheng, "Joint Forecasting Model for the Hourly Cooling Load and Fluctuation Range of a Large Public Building Based on GA-SVM and IG-SVM," *Sustainability (Switzerland)*, vol. 15, no. 24, Dec. 2023, doi: 10.3390/su152416833.

[5] F. Hambidi Wiyanto, "Penerapan Senam Kaki Diabetes Terhadap Sensitivitas Kaki Pada Penderita Diabetes Melitus Di Wilayah Puskesmas Pucangsawit," *Public Health and Safety International Journal*, vol. 3, no. 2, pp. 2715–5854, 2023, doi: 10.55642.

[6] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.v4i1.2561.

[7] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl Inf Syst*, vol. 65, no. 1, pp. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.

[8] O. Liashenko, T. Kravets, and Y. Kostovetskyi, "Machine Learning and Data Balancing Methods for Bankruptcy Prediction," *Ekonomika* , vol. 102, no. 2, pp. 28–46, 2023, doi: 10.15388/Ekon.2023.102.2.2.

[9] E. Kocyigit, M. Korkmaz, O. K. Sahingoz, and B. Diri, "Enhanced Feature Selection Using Genetic Algorithm for Machine-Learning-Based Phishing URL Detection," *Applied Sciences (Switzerland)*, vol. 14, no. 14, Jul. 2024, doi: 10.3390/app14146081.

[10] M. S. Hosen and S. S. Gutlapalli, "A Study of Innovative Class Imbalance Dataset Software Defect Prediction Methods," *Asian Journal of Applied Science and Engineering*, vol. 10, no. 1, pp. 52–55, Dec. 2021, doi: 10.18034/ajase.v10i1.52.

[11] I. R. Pratama, M. Maimunah, and E. R. Arumi, "Sistem Klasifikasi Penjualan Produk Alat Listrik Terlaris Untuk Optimasi Pengadaan Stok Menggunakan Naïve Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2135, Oct. 2022, doi: 10.30865/mib.v6i4.4418.

[12] F. Sulianta, *Basic Data Mining from A to Z*. Feri Sulianta, 2023.

[13] I. M. Hamdani, Nurhidayat, A. Karman, N. Fuady, and A. Hermina., "Edukasi dan Pelatihan Data Science dan Data Preprocessing," INTISARI Jurnal Inovasi Pengabdian Masyarakat, vol. 2, no. 1, pp. 19–26, 2024, doi: 10.58227/intisari.v2i1.125.

[14] John Baer, "Domain Specificity and the Limits of Creativity Theory," The Journal of Creative Behavior vol. 1–46, 2023, doi: 10.1002/jocb.002.

[15] T.-T. Wong and P.-Y. Yeh, "Reliable Accuracy Estimates from k -Fold Cross Validation," *IEEE Trans Knowl Data Eng*, vol. PP, p. 1, Apr. 2020, doi: 10.1109/TKDE.2019.2912815.

[16] Budhi Gustiandi, *Langkah Awal Menguasai Bahasa Pemrograman Phyton*. Penerbit BRIN, 2023. doi: 10.55981/brin.633.

[17] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

[18] 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.

[19] S. Katoch, S. S. Chauhan, and V. Kumar, "A review On Genetic Algorithm Past, Present, and Future," Multimedia Tools and Applications, vol. 80, 2021, doi: 0.1007/s11042-020-10139-6.

[20] W. R. Pratiwi and R. E. Putra, "Perbandingan Performa Algoritma GA-SVM dan BOA-SVM dalam Mengklasifikasi Artikel Berita Berbahasa Indonesia," *Journal of Informatics and Computer Science*, vol. 02, 2021, doi: 10.26740/jinacs.v2n04.p252-258.

[21] Pratiwi B, Handayani A, and Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi WSN Menggunakan Confusion Matrix," JURNAL INFORMATIKA UPGRIS, Vol. 6, No. 2, 2020, doi: 10.26877/jiu.v6i2.6552.

[22] R. M. Mathew and R. Gunasundari, "A Cluster-based Undersampling Technique for Multiclass Skewed Datasets," *Engineering, Technology and Applied Science Research*, vol. 13, no. 3, pp. 10785–10790, Jun. 2023, doi: 10.48084/etasr.5844.

[23] W. H. Lam, W. S. Lam, and P. F. Lee, "A Bibliometric Analysis of a Genetic Algorithm for Supply Chain Agility," *Mathematics*, vol. 12, no. 8, Apr. 2024, doi: 10.3390/math12081199.

[24] Y. Dilla, E. Wawan, J. Pranoto, and N. Adzmi Verdikha, "Evaluasi Support Vector Machine Dengan Optimasi Metode Genetic Algorithm Pada Klasifikasi Banjir Kota Samarinda," Jurnal Sains Komputer dan Teknologi Informasi, vol. 6, no. 1, 2023, doi: 10.33084/jsakti.v6i1.5462.

[25] 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, Jul. 2024, doi: 10.34148/teknika.v13i2.876.