Perbandingan Algoritma *Machine Learning* (LR, SVM, dan RF) untuk Pemodelan Hasil Seleksi Historis Penerima Beasiswa KIP Kuliah sebagai Pendukung *Pre-Screening*

Joko Kuswanto[*1], Handes Triwi Bowo[2], Anggraeni Agustin Muris[3], Pujianto[4], Jum Dapiokta[5], Riya Majalista[6], Yuli Ermawati[7]

[1,2,3,4,5,6,7]Program Studi Informatika Universitas Baturaja

E-mail: ko.8515@gmail.com[1], handestriwibowo27@gmail.com[2], miss.muris@gmail.com[3], Pujianto.mail@gmail.com[4],jumdapiokta@gmail.com[5], atsilajamayir@gmail.com [6], yulielektro.ubr@gmail.com [7]

Abstrak. Program Kartu Indonesia Pintar (KIP) Kuliah membantu calon mahasiswa dari keluarga kurang mampu mengakses pendidikan tinggi, tetapi proses seleksinya memerlukan verifikasi berkas dan survei lapangan yang memakan waktu. Penelitian ini membandingkan tiga algoritma *machine learning*, yaitu *Logistic Regression* (LR), *Support Vector Machine* (SVM), dan *Random Forest* (RF), dalam memodelkan keputusan seleksi historis KIP Kuliah di Universitas Baturaja. Data berupa 1.522 pendaftar periode 2022–2025 dengan label status tercantum atau tidak tercantum dalam daftar resmi penerima. Fitur hanya berasal dari komponen penilaian berkas (bobot 30% dalam seleksi kampus), sedangkan hasil survei lapangan (bobot 70%) tidak tersedia sebagai data terstruktur sehingga model tidak merepresentasikan kelayakan penerima secara utuh. Evaluasi memakai *Stratified 5-Fold Cross Validation* pada setiap tahun dengan SMOTE di dalam *pipeline* setiap lipatan, dengan metrik akurasi, presisi, *recall*, dan F1-score kelas diterima. RF memperoleh akurasi tertinggi pada seluruh tahun (0,636–0,699), sedangkan LR memperoleh *recall* dan F1-score kelas diterima tertinggi pada 2023–2025. Selisih F1-score antarmodel kecil dibandingkan simpangan baku antarlipatan sehingga tidak ditafsirkan sebagai keunggulan yang pasti. Model layak diposisikan sebagai alat bantu *pre-screening* yang tetap memerlukan verifikasi manusia, mekanisme banding, dan audit keadilan, bukan sebagai penentu otomatis penerima bantuan.
Kata Kunci :KIP Kuliah, *machine learning*, *Logistic Regression*, *Support Vector Machine*, *Random Forest*, seleksi historis, *pre-screening*.

Abstract. *The Indonesia Smart Card for Higher Education Program (Kartu Indonesia Pintar Kuliah or KIP Kuliah) helps prospective students from underprivileged families access higher education, but its selection process requires time-consuming document verification and field surveys. This study compares three machine learning algorithms, namely Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF), in modeling historical KIP Kuliah selection decisions at Baturaja University. The data consist of 1,522 applicants from 2022–2025, labeled by whether they appear on the official recipient list. Features come only from the document assessment component (30% weight in the campus selection), while field survey results (70% weight) are not available as structured data, so the models do not represent recipient eligibility in full. Evaluation uses Stratified 5-Fold Cross Validation for each year, with SMOTE applied inside each fold's pipeline, and accuracy, precision, recall, and F1-score of the accepted class. RF achieved the highest accuracy in every year (0.636–0.699), whereas LR achieved the highest recall and accepted-class F1-score in 2023–2025. F1-score differences between models were small relative to the between-fold standard deviation, so they are not interpreted as definite superiority. The models are suitable only as pre-screening aids that still require human verification, an appeals mechanism, and fairness audits, not as automatic decision-makers for aid recipients.*
Keywords:*KIP Kuliah, Machine learning, Logistic Regression, Support Vector Machine, Random Forest, Historical Selection, Pre-screening.*

1. PENDAHULUAN

1.1 Latar Belakang

Pendidikan tinggi merupakan investasi jangka panjang yang menentukan kualitas sumber daya manusia (SDM) dan daya saing bangsa [1], [2]. Namun, kesenjangan ekonomi membuat sebagian individu berpotensi tidak dapat melanjutkan studi. Pemerintah menjawab tantangan ini melalui program KIP Kuliah, yaitu bantuan biaya pendidikan bagi calon mahasiswa dari keluarga miskin atau rentan miskin [3]. Karena menyangkut keadilan sosial, penentuan kelayakan penerima merupakan tugas yang krusial dan sensitif [4].

Tingginya jumlah pendaftar membuat seleksi melibatkan data bervolume besar dengan banyak kriteria. Cara konvensional yang bertumpu pada verifikasi dokumen fisik, survei lapangan, dan penilaian sederhana rentan terhadap tiga masalah: inefisiensi waktu dan sumber daya, subjektivitas serta kesalahan manusia, dan kesalahan alokasi bantuan (*mismatched allocation*). Pendekatan klasifikasi *machine learning* (ML) memungkinkan sistem belajar dari data historis pendaftar berlabel "Diterima" atau "Ditolak" untuk memodelkan keputusan seleksi dan memperkirakan status pendaftar baru [5]. Perlu ditegaskan bahwa label tersebut adalah hasil seleksi historis, sehingga model yang dihasilkan menggambarkan keputusan seleksi berdasarkan data yang tersedia, bukan menilai kelayakan penerima secara utuh. Penerapan *data mining* pada seleksi beasiswa telah dilakukan sebelumnya [3], [6], [7], tetapi belum ada kajian yang secara khusus membandingkan LR, SVM, dan RF pada data pendaftar KIP Kuliah Universitas Baturaja. Inilah celah penelitian (*research gap*) yang diisi oleh penelitian ini.

Ketiga algoritma dipilih karena karakternya saling melengkapi: LR sebagai *baseline* yang mudah diinterpretasi, SVM yang mencari *hyperplane* optimal untuk data kompleks, dan RF sebagai model *ensemble* yang tahan terhadap *overfitting* serta mampu mengukur pentingnya fitur. Karena menurut prinsip *No Free Lunch* tidak ada algoritma yang unggul secara universal, perbandingan pada karakteristik data lokal menjadi perlu, dengan perhatian khusus pada F1-score dan *recall* kelas DITERIMA agar calon yang layak tidak terlewat (meminimalkan *false negative*).

1.2 Rumusan Masalah, Tujuan, dan Batasan

Penelitian ini menjawab empat pertanyaan berikut:

1. Tahapan *preprocessing* dan *feature selection* apa yang paling efektif untuk menyiapkan dataset KIP Kuliah agar siap dimodelkan?
2. Bagaimana kinerja LR, SVM, dan RF dalam memodelkan keputusan seleksi historis KIP Kuliah berdasarkan komponen penilaian berkas?
3. Algoritma mana yang memperoleh nilai tertinggi berdasarkan akurasi, presisi, *recall*, dan F1-score, dan apakah selisihnya cukup besar dibandingkan variasi antarlipatan?
4. Bagaimana implikasi teknis dan etis penggunaan model sebagai alat bantu *pre-screening* di Universitas Baturaja?

Tujuannya adalah menerapkan tahapan *preprocessing* dan pemilihan fitur yang sesuai, membangun dan mengevaluasi ketiga model untuk memodelkan hasil seleksi historis, menganalisis pola tahun 2022–2025, serta merumuskan rekomendasi penggunaan model sebagai alat bantu *pre-screening* dengan pengawasan manusia. Ruang lingkupnya dibatasi pada data pendaftar tahun ajaran 2022–2025 dari Universitas Baturaja yang dianalisis per tahun (tidak digabung), fitur kategori ekonomi dan prestasi yang tersedia di dataset, tiga algoritma dengan konfigurasi baseline dan empat metrik standar, serta *blueprint* teknis model tanpa implementasi sistem utuh di kampus. Penelitian ini tidak menilai kelayakan sosial-ekonomi pendaftar secara utuh karena hasil survei lapangan (bobot 70%) tidak tersedia sebagai data terstruktur.

1.3 Manfaat Penelitian

Secara akademis, penelitian ini menambah kajian penerapan algoritma klasifikasi pada kasus sosial yang sensitif seperti seleksi beasiswa, menyediakan studi perbandingan terstruktur atas LR, SVM, dan RF, serta memberi bukti empiris tentang pentingnya memilih algoritma dan memperhatikan metrik *recall* pada keputusan yang menuntut keadilan. Bagi Universitas Baturaja, model dapat menjadi alat bantu *pre-screening* untuk memprioritaskan urutan verifikasi sehingga berpotensi mengurangi waktu verifikasi manual, dengan keputusan akhir tetap berada pada manusia, serta menyediakan dasar berbasis data (misalnya *feature importance*) untuk mengevaluasi kriteria seleksi. Secara sosial, penelitian ini diharapkan mendukung keadilan akses pendidikan tinggi melalui penggunaan model yang transparan dan diawasi, dan dalam jangka panjang peningkatan kualitas SDM regional.

2. LANDASAN TEORI

2.1 KIP Kuliah dan Seleksi Penerima

KIP Kuliah diatur dalam Permendikbud No. 10 Tahun 2020 sebagai bantuan biaya pendidikan bagi peserta didik dan mahasiswa baru dari keluarga miskin atau rentan miskin [3], [8]. Program ini bertujuan membuka akses pendidikan tinggi bagi siswa berpotensi akademik dari keluarga kurang mampu, meningkatkan semangat dan hasil

belajar, memastikan mahasiswa lulus tepat waktu, serta melahirkan lulusan mandiri dan produktif yang mampu memutus rantai kemiskinan. Secara umum, beasiswa dimaksudkan memperluas akses pendidikan bagi kelompok yang kurang mampu membiayai studinya [7].

2.2 Machine Learning dan Supervised Learning

ML adalah cabang kecerdasan buatan yang membuat komputer belajar dari data dan menghasilkan prediksi atau keputusan tanpa pemrograman eksplisit [5]. Pengembangan model ML lazimnya melalui enam tahap: penentuan kebutuhan model, pemrosesan data, perancangan fitur, pelatihan, evaluasi, dan penyediaan (*deployment*) model [5]. Penelitian ini berada dalam kerangka *supervised learning*, yakni model belajar dari pasangan input dan label yang telah diketahui [5], dan tergolong klasifikasi biner karena target yang diprediksi bersifat kategorikal: Diterima (kelas 1) atau Ditolak (kelas 0). Tugas regresi tidak digunakan karena target bukan nilai numerik kontinu.

2.3 Algoritma yang Dibandingkan

Logistic Regression menggabungkan fitur secara linear, lalu memetakan skor tersebut menjadi probabilitas 0–1 melalui fungsi sigmoid (Persamaan 1); koefisiennya diestimasi dengan *maximum likelihood estimation* dan kelas ditentukan dengan ambang keputusan, umumnya 0,5 [9]. Model ini efisien dan skalabel [10]. *Support Vector Machine* mencari *hyperplane* dengan *margin* terbesar di antara dua kelas; titik data terdekat yang menentukan margin disebut *support vector* [11]–[13]. Untuk data non-linear, SVM memakai fungsi kernel (misalnya RBF) yang memproyeksikan data ke ruang berdimensi lebih tinggi. *Random Forest* membangun banyak pohon keputusan dari sampel *bootstrap* (*bagging*) dengan pemilihan fitur acak pada setiap simpul, lalu menggabungkan prediksinya melalui suara mayoritas [14]–[16]. Perbandingan karakteristiknya diringkas pada Tabel 1.

$$P(Y = 1) = 1 / (1 + e^{-Z}), \quad Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{1}$$

Tabel 1. Perbandingan karakteristik LR, SVM, dan RF

Algoritma	Prinsip kerja	Kelebihan	Keterbatasan
Logistic Regression	Kombinasi linear fitur, fungsi sigmoid, dan probabilitas; koefisien diestimasi dengan MLE	Mudah diinterpretasi (*white box*); cepat dan skalabel; menghasilkan probabilitas	Mengasumsikan hubungan linear; sensitif terhadap *outlier*; lebih optimal bila fitur distandardisasi
Support Vector Machine	*Hyperplane* dengan *margin* maksimal; *kernel trick* untuk data non-linear	Kuat pada batas keputusan non-linear; solusi optimal global; efektif pada dimensi tinggi	Berat pada data besar; sensitif terhadap kernel dan parameter (C, γ); sulit diinterpretasi
Random Forest	*Ensemble* pohon keputusan (*bagging* dan *feature randomness*) dengan suara mayoritas	Akurasi tinggi dan stabil; tahan *overfitting*; menghasilkan *feature importance*; tanpa *scaling*	Sulit diinterpretasi (*black box*); komputasi dan memori lebih besar; cenderung condong ke fitur dominan

2.4 Preprocessing, Feature Engineering, dan Data Tidak Seimbang

Kualitas data menentukan keandalan model. *Preprocessing* mencakup integrasi, pelabelan, transformasi, dan pembersihan data untuk menangani duplikat, *outlier*, serta nilai hilang [17]. Data kategorikal perlu diubah menjadi numerik melalui *encoding*, yaitu *label encoding* yang berisiko menimbulkan urutan ordinal semu atau *one-hot encoding* yang menghindarinya; pemilihan teknik *feature engineering* yang tepat terbukti memengaruhi akurasi [18]. Selain itu, distribusi kelas yang timpang (*imbalanced data*) membuat model bias ke kelas mayoritas, menghasilkan akurasi yang menyesatkan, dan mengabaikan kelas minoritas yang justru penting [19]. Penanganannya dapat berupa *oversampling*, *undersampling*, atau SMOTE yang membangkitkan sampel sintetik kelas minoritas dan dapat menekan *overfitting* akibat duplikasi data [19], [20].

2.5 Validasi Silang dan Metrik Evaluasi

Pengujian pada data latih cenderung terlalu optimis dan berujung *overfitting*. Karena itu digunakan *k-fold cross-validation*: data dibagi menjadi *k* lipatan, setiap lipatan bergantian menjadi data uji sementara lipatan lain menjadi data latih, dan kinerja akhir adalah rata-rata seluruh iterasi [5], [21]. Varian *stratified* menjaga proporsi kelas di tiap lipatan. Metrik dihitung dari *confusion matrix* (TP, TN, FP, FN) [5]:

$$Presisi = TP / (TP + FP) \tag{2}$$

$$Recall = TP / (TP + FN) \tag{3}$$

$$F1\text{-}score = 2 \times Presisi \times Recall / (Presisi + Recall) \tag{4}$$

$$Akurasi = (TP + TN) / (TP + TN + FP + FN) \tag{5}$$

Presisi mengukur ketepatan prediksi positif, *recall* mengukur kemampuan menemukan seluruh kasus positif, F1-score adalah rata-rata harmonik keduanya, dan akurasi menunjukkan proporsi prediksi yang benar secara keseluruhan.

2.6 Penelitian Terdahulu

Tabel 2 merangkum tiga studi terdahulu yang paling relevan dengan tahapan metodologis penelitian ini.

Tabel 2. Penelitian terdahulu yang relevan

Peneliti	Fokus kajian	Temuan utama	Relevansi
Qadrini *dkk.* (2022) [19]	Penerima Bidikmisi se-Jawa Timur 2017: *oversampling*, *undersampling*, dan SMOTE pada SVM dan RF	SVM dengan *oversampling* terbaik (F1-score 93,66%)	Urgensi penanganan data tidak seimbang
Gori *dkk.* (2024) [17]	*Preprocessing* dan klasifikasi untuk prediksi kinerja akademik siswa	Pemilihan fitur yang tepat berpengaruh signifikan pada performa model	Dasar tahap *feature selection*
Herdian *dkk.* (2024) [18]	*Label encoding* vs *one-hot encoding* pada regresi linear	Teknik *encoding* memengaruhi akurasi; menjadi panduan teknis *feature engineering*	Pengolahan fitur kategorikal

3. METODE PENELITIAN

3.1 Waktu, Tempat, serta Alat dan Bahan

Penelitian berlangsung sekitar Oktober 2025 hingga Mei 2026. Data diperoleh dari Bagian Akademik Universitas Baturaja setelah mendapat izin resmi, sedangkan seluruh pengolahan dilakukan secara daring di *Google Colaboratory* dengan bahasa Python dan pustaka Pandas, NumPy, Scikit-learn [22], Imbalanced-learn [23], Matplotlib, serta Seaborn, menggunakan laptop HP EliteBook 820 G2 (Intel Core i5-5200U, RAM 8 GB).

3.2 Data, Variabel, dan Pembobotan Fitur

Data bersifat sekunder, kuantitatif, dan terstruktur, dikumpulkan dengan teknik dokumentasi terhadap arsip digital (CSV/Excel) pendaftar KIP Kuliah tahun 2022–2025 sebanyak 1.522 baris. Data identitas pribadi tidak digunakan sebagai fitur demi kerahasiaan pendaftar. Variabel target adalah Status_Lulus_Label: 1 untuk pendaftar yang tercantum dalam daftar resmi penerima (DITERIMA, kelas minoritas) dan 0 untuk yang tidak tercantum (TIDAK DITERIMA). Label ini merupakan hasil keputusan seleksi akhir kampus yang dipengaruhi oleh penilaian berkas, survei lapangan, dan kemungkinan kuota tahunan, sehingga bukan ukuran kelayakan yang murni. Variabel input berupa enam fitur numerik berbobot dari atribut ekonomi dan prestasi.

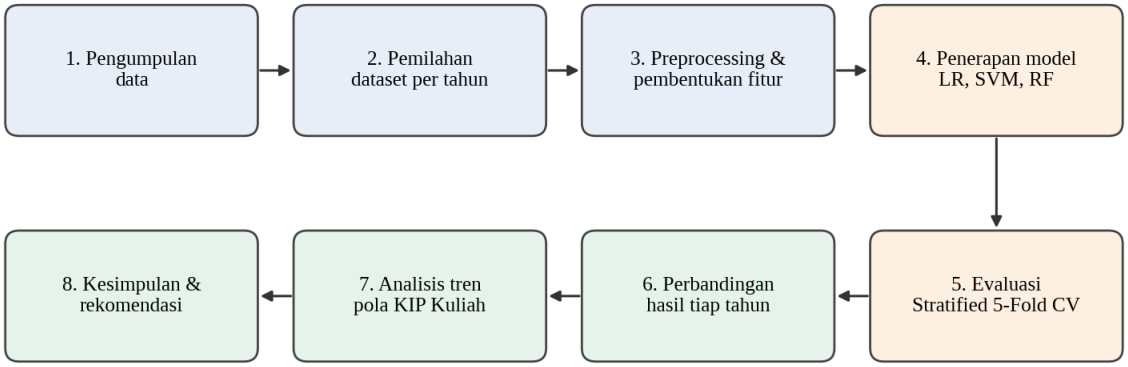
Pembentukan fitur mengacu pada bobot resmi yang dipakai pihak akademik Universitas Baturaja saat menilai kelengkapan berkas. Seleksi di kampus terdiri atas penilaian berkas (bobot 30%) dan survei lapangan (bobot 70%); penelitian ini hanya memakai komponen berkas karena hasil survei tidak tersedia sebagai data terstruktur. Dengan demikian, sekitar 70% informasi yang menentukan keputusan akhir tidak masuk ke fitur, dan model memodelkan hasil seleksi historis berdasarkan data berkas, bukan kelayakan penerima secara utuh. Istilah "prediksi" dalam naskah ini dimaknai dengan pengertian tersebut. Tabel 3 merangkum pembobotan tersebut.

Tabel 3. Fitur numerik berbobot dan kriteria pembobotannya

Fitur	Atribut asal	Kriteria (bobot)
Bobot_Jumlah_Bantuan	Status DTKS, P3KE, KIP, KKS	≥ 2 bantuan (4); 1 bantuan (2); tidak ada (0)
Bobot_Penghasilan_Ortu	Penghasilan ayah dan ibu	≤ Rp4.000.000 (4); > Rp4.000.000 (0)
Bobot_Status_Ortu	Status ayah dan ibu	Yatim piatu (4); yatim (3); piatu (2); masih lengkap (1)
Bobot_Rumah	Kepemilikan rumah	Menyewa (4); menumpang (3); warisan keluarga (2); rumah pribadi (1)
Bobot_Prestasi_Akademik	Prestasi siswa (peringkat)	Peringkat 1–3 (6); 4–6 (5); 7–10 (4); tidak ada (0)
Bobot_Prestasi_Non_Akademik	Prestasi siswa (tingkat kejuaraan)	Nasional (4); provinsi (3); kabupaten (2); sekolah/kelas (1); tidak ada (0); dipakai tingkat tertinggi

3.3 Alur Penelitian dan Skema Evaluasi

Penelitian dilakukan dalam delapan tahap (Gambar 1). Dataset dipisahkan per tahun karena jumlah pendaftar, proporsi kelas, dan kondisi seleksi berbeda setiap tahun; prosedur yang sama diterapkan pada tiap dataset agar perbandingan adil.



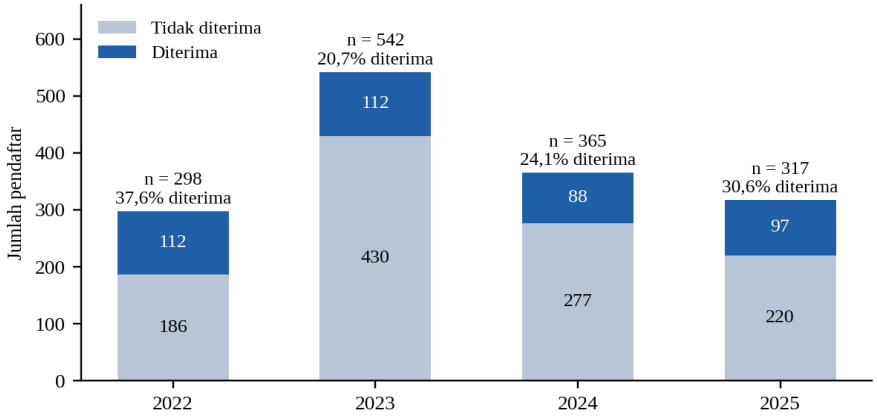
Gambar 1. Kerangka alur penelitian

Evaluasi memakai *Stratified 5-Fold Cross Validation* pada setiap dataset tahunan. Seluruh langkah berikut dibungkus dalam satu *pipeline* dan hanya dijalankan pada data latih di setiap lipatan: imputasi nilai kosong, standardisasi fitur (hanya untuk LR dan SVM yang sensitif terhadap skala, tidak untuk RF), penyeimbangan kelas dengan SMOTE, dan pelatihan model. Data validasi tidak dikenai SMOTE sehingga tetap mencerminkan distribusi asli, dan penempatan ini mencegah *data leakage*. Ketiga model dijalankan dengan konfigurasi baseline (parameter bawaan Scikit-learn) dan prosedur validasi yang sama tanpa *tuning hyperparameter*, sehingga perbandingan ini berlaku untuk konfigurasi baseline tersebut dan tidak menunjukkan batas kemampuan masing-masing algoritma. Validasi dilakukan silang di dalam setiap tahun; pengujian lintas tahun belum dilakukan. Kinerja dilaporkan sebagai rata-rata dan simpangan baku (SD) seluruh lipatan. Nilai tertinggi ditentukan dengan dua kriteria, yaitu akurasi tertinggi dan F1-score kelas DITERIMA tertinggi; bila model dipakai untuk *pre-screening*, F1-score kelas DITERIMA diprioritaskan. Karena selisih antarmodel dapat berada dalam rentang variasi antarlipatan, satu model tidak disebut unggul bila rentang rerata ± SD antarmodel saling tumpang tindih.

4. HASIL DAN PEMBAHASAN

4.1 Karakteristik Dataset

Gambar 2 menunjukkan bahwa jumlah data dan proporsi pendaftar yang diterima berbeda tiap tahun. Tahun 2023 memiliki data terbanyak (542), sedangkan proporsi DITERIMA tertinggi terjadi pada 2022 (37,6%) dan terendah pada 2023 (20,7%). Kelas DITERIMA merupakan kelas minoritas di semua tahun (409 dari 1.522 data, sekitar 26,9%), sehingga penanganan dengan SMOTE dan stratifikasi lipatan beralasan. Perubahan proporsi ini juga mendukung keputusan menganalisis data per tahun.



Gambar 2. Distribusi data pendaftar KIP Kuliah per tahun (total 1.522: 409 diterima, 1.113 tidak diterima)

4.2 Hasil Evaluasi Model

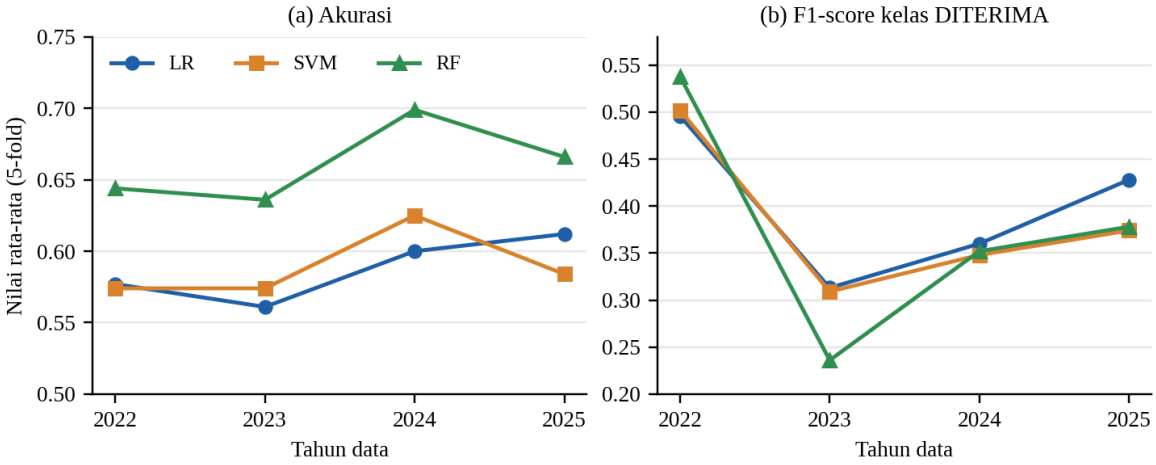
Tabel 4 merekap hasil evaluasi seluruh model. Berdasarkan akurasi, RF tertinggi pada seluruh tahun (0,644; 0,636; 0,699; dan 0,666 untuk 2022–2025) dengan puncak pada 2024; hal ini kemungkinan terkait dengan kemampuan RF menangkap pola non-linear dan interaksi antarfitur, namun dugaan ini belum diuji. Bila kriterianya F1-score kelas DITERIMA, RF hanya tertinggi pada 2022 (0,538), sedangkan LR tertinggi pada 2023 (0,313), 2024 (0,360), dan 2025 (0,428) (Gambar 3 dan Tabel 5).

Tabel 4. Rekap hasil *Stratified 5-Fold Cross Validation* (kelas 1 = DITERIMA; cetak tebal = tertinggi per tahun pada akurasi dan F1)

Tahun	Algoritma	Akurasi (rerata)	Akurasi (SD)	Presisi kelas 1	*Recall* kelas 1	F1 kelas 1 (rerata)	F1 kelas 1 (SD)
2022	*Random Forest*	**0,644**	0,026	0,523	0,579	**0,538**	0,094
2022	SVM	0,574	0,030	0,453	0,590	0,502	0,056
2022	*Logistic Regression*	0,577	0,053	0,455	0,553	0,496	0,044
2023	*Logistic Regression*	0,561	0,061	0,236	0,492	**0,313**	0,026
2023	SVM	0,574	0,071	0,237	0,466	0,309	0,077
2023	*Random Forest*	**0,636**	0,080	0,221	0,278	0,236	0,083
2024	*Logistic Regression*	0,600	0,089	0,304	0,453	**0,360**	0,094
2024	*Random Forest*	**0,699**	0,071	0,359	0,351	0,352	0,180
2024	SVM	0,625	0,051	0,301	0,422	0,348	0,090
2025	*Logistic Regression*	0,612	0,059	0,402	0,474	**0,428**	0,025
2025	*Random Forest*	**0,666**	0,052	0,466	0,331	0,378	0,045
2025	SVM	0,584	0,060	0,358	0,403	0,374	0,036

Tabel 5. Model terbaik per tahun berdasarkan dua kriteria penentuan

Tahun	Terbaik menurut akurasi	Akurasi	Terbaik menurut F1 kelas DITERIMA	F1-score
2022	*Random Forest*	0,644	*Random Forest*	0,538
2023	*Random Forest*	0,636	*Logistic Regression*	0,313
2024	*Random Forest*	0,699	*Logistic Regression*	0,360
2025	*Random Forest*	0,666	*Logistic Regression*	0,428



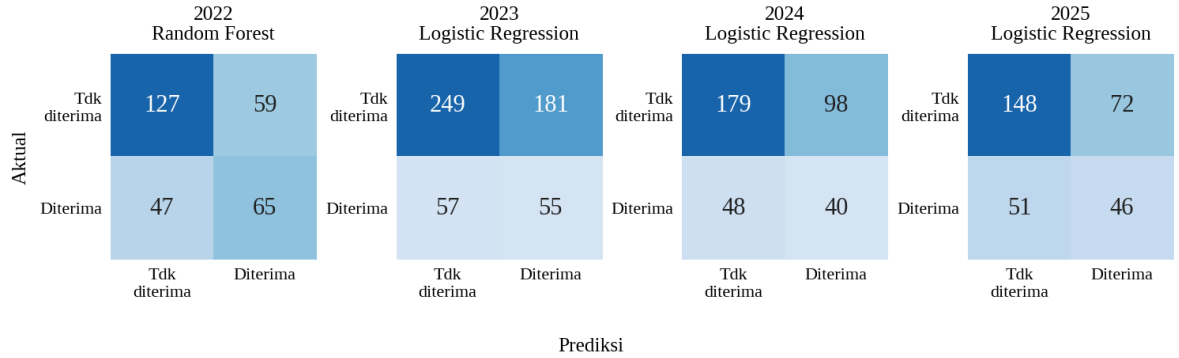
Gambar 3. Tren (a) akurasi dan (b) F1-score kelas DITERIMA pada periode 2022–2025

Beberapa hal patut dicatat dari Tabel 4. *Pertama*, *recall* kelas DITERIMA milik LR lebih tinggi daripada RF pada 2023–2025 (0,492 vs 0,278; 0,453 vs 0,351; 0,474 vs 0,331), artinya RF yang lebih akurat justru lebih sering melewatkan pendaftar yang tercantum sebagai penerima. *Kedua*, selisih F1-score antarmodel kecil dibandingkan variasi antarlipatan. Pada 2024, F1-score LR, RF, dan SVM masing-masing 0,360; 0,352; dan 0,348 dengan SD 0,094; 0,180; dan 0,090. Pada seluruh tahun, rentang rerata ± SD antarmodel saling tumpang tindih; pada 2025 tumpang tindihnya tipis (LR 0,403–0,453; RF 0,333–0,423). Karena itu, istilah "tertinggi" dalam naskah ini

bermakna numerik dan belum menunjukkan keunggulan yang konsisten secara statistik, yang memerlukan uji lanjutan (lihat subbab 4.6). *Ketiga*, seluruh F1-score kelas DITERIMA turun tajam pada 2023 lalu berfluktuasi; penulis menilai data 2023 lebih sulit dikenali dan prediksi kelas DITERIMA lebih sukar daripada klasifikasi keseluruhan. Karena itu, akurasi saja tidak cukup untuk memilih model pada seleksi KIP Kuliah. Pola yang lebih dapat dipertahankan adalah pertukaran (*trade-off*): RF lebih akurat pada kelas mayoritas, sedangkan LR menangkap lebih banyak pendaftar yang tercantum sebagai penerima dengan konsekuensi lebih banyak *false positive*.

4.3 Analisis Confusion Matrix

Gambar 4 memperlihatkan *confusion matrix* model terbaik menurut F1-score kelas DITERIMA pada tiap tahun. *False negative* (pendaftar yang tercantum sebagai penerima tetapi diprediksi tidak diterima) berjumlah 47, 57, 48, dan 51 untuk 2022–2025, sedangkan *false positive* mencapai 181 pada LR 2023, yakni model cenderung memprediksi DITERIMA bagi pendaftar yang sebenarnya tidak tercantum. Kesalahan pada kelas DITERIMA berdampak langsung pada calon penerima yang tidak teridentifikasi, sehingga model sebaiknya dipakai sebagai alat bantu *pre-screening*, bukan penentu akhir.



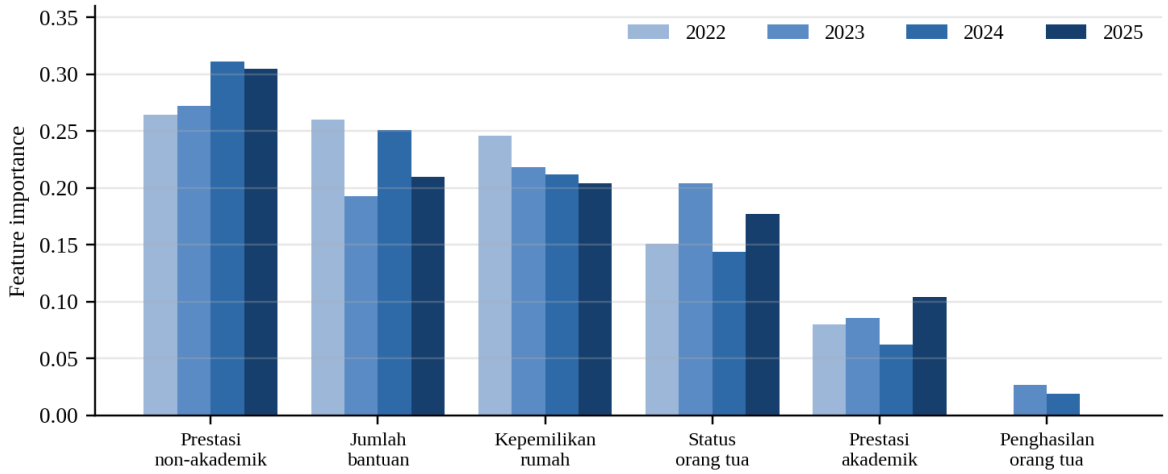
Gambar 4. *Confusion matrix* model terbaik berdasarkan F1-score kelas DITERIMA per tahun

4.4 Analisis Pola dan Pengaruh Fitur

Pada RF, *feature importance* (Tabel 6 dan Gambar 5) menempatkan Bobot_Prestasi_Non_Akademik sebagai fitur paling berpengaruh di keempat tahun (0,264; 0,272; 0,311; 0,305), diikuti jumlah bantuan sosial, kepemilikan rumah, dan status orang tua, sedangkan penghasilan orang tua berkontribusi paling kecil (0,000–0,027).

Tabel 6. Nilai *feature importance Random Forest* per tahun (cetak tebal = tertinggi)

Fitur	2022	2023	2024	2025
Bobot_Prestasi_Non_Akademik	**0,264**	**0,272**	**0,311**	**0,305**
Bobot_Jumlah_Bantuan	0,260	0,193	0,251	0,210
Bobot_Rumah	0,246	0,218	0,212	0,204
Bobot_Status_Ortu	0,151	0,204	0,144	0,177
Bobot_Prestasi_Akademik	0,080	0,086	0,062	0,104
Bobot_Penghasilan_Ortu	0,000	0,027	0,019	0,000



Gambar 5. *Feature importance Random Forest* tahun 2022–2025

Koefisien LR menunjukkan pola yang berubah antartahun. Menurut penulis, koefisien prestasi akademik negatif pada 2022–2023 lalu menjadi positif dan dominan pada 2024–2025, yang mungkin mencerminkan penyesuaian kebijakan seleksi; koefisien penghasilan orang tua positif pada 2023 tetapi negatif pada 2024; status orang tua relatif konsisten positif; jumlah bantuan bernilai negatif terutama pada 2024, yang ditafsirkan sebagai kecenderungan pemerataan bantuan; kondisi rumah cenderung negatif; dan prestasi non-akademik mendekati nol. Perlu dicatat bahwa tafsiran ini masih berupa dugaan dan kedua pendekatan tidak sepenuhnya sejalan (prestasi non-akademik terpenting bagi RF, tetapi berkoefisien kecil pada LR), sehingga konfirmasi dengan ketentuan seleksi kampus diperlukan.

4.5 Rekomendasi Penerapan

Model ML direkomendasikan hanya sebagai sistem pendukung keputusan untuk seleksi awal (*pre-screening*), yaitu memprioritaskan urutan verifikasi, bukan menolak atau menetapkan penerima secara otomatis dan bukan pengganti keputusan akhir. Bila tujuan utamanya akurasi klasifikasi keseluruhan, RF memiliki nilai numerik yang lebih tinggi; bila prioritasnya memastikan pendaftar yang layak tidak terlewat, LR lebih dapat dipertimbangkan karena F1-score dan *recall* kelas DITERIMA-nya lebih tinggi pada 2023–2025, dengan catatan bahwa selisihnya belum diuji secara statistik. Karena pola data bersifat dinamis, model perlu dievaluasi secara berkala.

4.6 Keterbatasan dan Implikasi Etis

Keterbatasan data dan label. Label berasal dari keputusan seleksi historis yang memuat penilaian survei lapangan, sedangkan model hanya memakai komponen berkas. Akibatnya, kinerja model tidak dapat ditafsirkan sebagai ketepatan menilai kelayakan penerima. Label historis juga dapat memuat subjektivitas dan pembatasan kuota. Data per tahun relatif kecil (misalnya 2023 sebanyak 542 data dengan kelas DITERIMA 20,7%) dan proporsi kelas berubah antartahun, sehingga variasi antarlipatan cukup besar.

Keterbatasan evaluasi. Pertama, validasi dilakukan silang di dalam tahun yang sama, sehingga belum diketahui apakah model yang dilatih pada tahun-tahun sebelumnya (misalnya 2022–2023) masih bekerja pada pendaftar tahun berikutnya (2024); perubahan karakteristik pendaftar dan pola seleksi (*concept drift*) belum terukur. Kedua, perbedaan antarmodel belum diuji secara statistik dan belum disertai interval kepercayaan. Ketiga, ketiga model dijalankan dengan konfigurasi baseline tanpa *tuning*, sehingga hasil tidak menunjukkan batas kemampuan masing-masing algoritma. Keempat, metrik yang dilaporkan belum mencakup PR-AUC, F2-score, maupun *recall* pada kapasitas verifikasi tertentu yang lebih dekat dengan skenario operasional. Kelima, SMOTE belum dibandingkan dengan kondisi tanpa SMOTE, sehingga kontribusi *resampling* terhadap kinerja belum dapat dipisahkan.

Risiko bias dan keadilan. Fitur penghasilan, status orang tua, dan kepemilikan rumah merupakan kriteria yang sah, tetapi dapat mereproduksi pola bias pada seleksi historis dan membuat kelompok tertentu lebih sering terlewat. Sebelum dipakai secara operasional, model perlu diaudit antarkelompok yang tersedia dan diizinkan (misalnya asal sekolah, wilayah, atau status orang tua) dengan membandingkan *false negative rate* dan *recall* antarkelompok. Atribut kelompok dipakai untuk audit, bukan sebagai fitur model.

Privasi, tata kelola, dan pengawasan manusia. Data pendaftar bersifat sensitif sehingga penggunaannya perlu mengikuti prinsip pelindungan data pribadi, yaitu minimisasi data, pembatasan akses, dan dasar hukum yang jelas. Model hanya dipakai untuk memprioritaskan urutan verifikasi (*human in the loop*): setiap pendaftar tetap dapat diverifikasi petugas, keputusan akhir berada pada tim seleksi, pendaftar berhak mengajukan banding, dan skor model tidak boleh menjadi satu-satunya alasan penolakan.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Model LR, SVM, dan RF yang dilatih pada 1.522 data pendaftar KIP Kuliah Universitas Baturaja 2022–2025 dapat menangkap sebagian pola keputusan seleksi historis berdasarkan komponen penilaian berkas, yang diolah melalui *preprocessing*, pembentukan fitur berbobot, dan pengujian dengan *Stratified 5-Fold Cross Validation*. Karena survei lapangan (bobot 70%) tidak menjadi fitur dan label berasal dari daftar penerima resmi, model tidak mengukur kelayakan penerima secara utuh.
2. RF memperoleh akurasi tertinggi pada seluruh tahun (0,636–0,699). Karena kelas DITERIMA merupakan minoritas (sekitar 26,9%), akurasi saja tidak memadai, dan F1-score kelas DITERIMA seluruh model tergolong rendah (0,236–0,538).
3. Berdasarkan F1-score dan *recall* kelas DITERIMA, LR tertinggi pada 2023–2025 dan RF pada 2022. Selisih antarmodel kecil dibandingkan SD antarlipatan dan rentang rerata ± SD saling tumpang tindih, sehingga keunggulan tersebut tidak dapat dinyatakan konsisten secara statistik. Pemilihan model lebih tepat didasarkan pada kebutuhan: RF bila akurasi keseluruhan diutamakan, LR bila menangkap pendaftar yang tercantum sebagai penerima (*recall*) diutamakan.
4. *False negative* tetap muncul pada setiap tahun (47–57 pendaftar pada model terbaik menurut F1-score). Model hanya layak sebagai alat bantu *pre-screening* untuk memprioritaskan verifikasi, dengan keputusan akhir oleh manusia dan mekanisme banding, bukan sebagai penentu otomatis penerima atau penolak bantuan.
5. Proporsi pendaftar diterima dan kinerja model berubah antartahun (kelas DITERIMA 37,6% pada 2022 dan 20,7% pada 2023). Karena validasi baru dilakukan di dalam tahun yang sama, stabilitas model lintas tahun belum terbukti dan model perlu dievaluasi ulang pada setiap periode seleksi.
6. Hasil penelitian berlaku untuk konfigurasi baseline tanpa *tuning*, tanpa uji statistik formal, dan tanpa pembanding SMOTE, serta belum dilengkapi metrik operasional seperti PR-AUC, F2-score, dan *recall* pada kapasitas verifikasi tertentu.
7. Fitur ekonomi dan sosial berpotensi mereproduksi bias seleksi historis. Penggunaan model perlu disertai audit keadilan antarkelompok, tata kelola dan perlindungan data pendaftar, serta prinsip *human in the loop*.

5.2 Saran

1. Melengkapi fitur dengan hasil survei lapangan dalam bentuk data terstruktur (sesuai bobot 70% dalam seleksi kampus), serta kondisi tempat tinggal yang lebih rinci, latar belakang sekolah, tanggungan keluarga, dan indikator sosial ekonomi lain agar model lebih mendekati proses seleksi yang sebenarnya.
2. Menguji algoritma lain (XGBoost, *Decision Tree*, KNN, *Neural Network*) dan melakukan *tuning hyperparameter* yang setara untuk semua model dengan melaporkan ruang pencarian, parameter final, dan prosedurnya.
3. Menambahkan validasi temporal (misalnya latih 2022–2023 uji 2024, latih 2022–2024 uji 2025), uji statistik atau interval kepercayaan untuk perbandingan antarmodel, metrik operasional (PR-AUC, F2-score, dan *recall* pada kapasitas verifikasi), serta *ablation study* SMOTE.
4. Melakukan audit keadilan antarkelompok, memantau *concept drift*, dan menyusun tata kelola data serta prosedur banding sebelum model dipakai secara operasional.
5. Bagi kampus, mengevaluasi dan merinci skema pembobotan seleksi yang saat ini masih sederhana agar lebih proporsional menggambarkan kondisi pendaftar.

DAFTAR PUSTAKA

[1] D. Amrizal, D. F. Nasution, and A. Amran, "Efektivitas pelayanan program Kartu Indonesia Pintar (KIP) dalam rangka peningkatan kualitas pendidikan di SMA Negeri 1 Tanjung Balai," *J. SOMASI (Sosial Hum. Komunikasi)*, vol. 1, no. 1, pp. 11–20, 2020, doi: 10.53695/js.v1i1.27.

[2] F. Y. Liriwati, Rulitawati, and Zulhimma, "Revolusi perguruan tinggi di era revolusi industri 4.0," in *Pros. Semin. Nas. Pendidik. Progr. Pascasarj. Univ. PGRI Palembang*, 2019, pp. 280–288.

[3] G. Suganda, M. Asfi, R. T. Subagio, and R. P. Kusuma, "Penentuan penerima bantuan beasiswa Kartu Indonesia Pintar (KIP) Kuliah menggunakan Naïve Bayes classifier," *JSiI (J. Sist. Informasi)*, vol. 9, no. 2, pp. 193–199, 2022, doi: 10.30656/jsii.v9i2.4376.

[4] F. Nuraeni, D. Kurniadi, and G. F. Dermawan, "Pemetaan karakteristik mahasiswa penerima Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan algoritma K-Means++," *J. Sisfokom (Sist. Inf. dan Komput.)*, vol. 11, no. 3, pp. 437–443, 2023, doi: 10.32736/sisfokom.v11i3.1439.

[5] P. Chyan *et al.*, *Pengantar Machine Learning*, vol. 1. PT Mifandi Mandiri Digital, 2024.

[6] A. I. Kushartanto and R. T. Aldisa, "Data mining perbandingan algoritma K-Nearest Neighbor dan Naïve Bayes dalam prediksi penerimaan beasiswa," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 196–207, 2023, doi: 10.47065/josyc.v5i1.4566.
[7] Baskoro, Sriyanto, and L. S. Rini, "Prediksi penerima beasiswa dengan menggunakan teknik data mining di Universitas Muhammadiyah Pringsewu," in *Semin. Nas. Has. Penelit. dan Pengabdi. Masy.*, 2021, pp. 87–94.
[8] H. Ali and N. Hendrastuty, "Comparison of Naïve Bayes classifier, support vector machine, random forest algorithms for public sentiment analysis of KIP-K program on Twitter," *J. Tek. Inform.*, vol. 5, no. 6, pp. 1701–1712, 2024, doi: 10.52436/1.jutif.2024.5.6.4030.
[9] F. Handayani, K. S. Kusuma, H. L. Asbudi, R. G. Purnasiwi, and R. Kusuma, "Komparasi Support Vector Machine, Logistic Regression dan Artificial Neural Network dalam prediksi penyakit jantung," *JEPIN (J. Edukasi dan Penelit. Inform.)*, vol. 7, no. 3, pp. 329–334, 2021, doi: 10.26418/jp.v7i3.48053.
[10] M. S. T. Putra and Y. Azhar, "Perbandingan model Logistic Regression dan Artificial Neural Network pada prediksi pembatalan hotel," *JISKA (J. Inform. Sunan Kalijaga)*, vol. 6, no. 1, pp. 29–37, 2021, doi: 10.14421/jiska.2021.61-04.
[11] S. Nurhayati, Kusrini, and E. T. Luthfi, "Prediksi mahasiswa drop out menggunakan metode Support Vector Machine," *J. Ilm. SISFOTENIKA*, vol. 5, no. 1, pp. 82–93, 2015.
[12] M. R. Qisthiano, "Klasifikasi terhadap prediksi kelulusan mahasiswa dengan menggunakan metode Support Vector Machine (SVM)," *Semin. Nas. Teknol. dan Multidisiplin Ilmu*, vol. 2, no. 2, pp. 203–207, 2022, doi: 10.51903/semnastekmu.v2i1.170.
[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[14] J. Kuswanto and L. Hakim, "Penerapan algoritma Random Forest untuk memprediksi performa akademik mahasiswa," *Decode J. Pendidik. Teknol. Inf.*, vol. 5, no. 1, pp. 262–270, 2025, doi: 10.51454/decode.v5i1.1103.
[15] B. Prasojo and E. Haryatmi, "Analisa prediksi kelayakan pemberian kredit pinjaman dengan metode Random Forest," *J. Nas. Teknol. dan Sist. Inf.*, vol. 7, no. 2, pp. 79–89, 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.
[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[17] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing data dan klasifikasi untuk prediksi kinerja akademik siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.
[18] C. Herdian, A. Kamila, and I. G. A. M. Budidarma, "Studi kasus feature engineering untuk data teks: Perbandingan label encoding dan one-hot encoding pada metode linear regresi," *Technol. J. Ilm.*, vol. 15, no. 1, p. 93, 2024, doi: 10.31602/tji.v15i1.13457.
[19] L. Qadrini, H. Hikmah, and M. Megasari, "Oversampling, undersampling, SMOTE SVM dan Random Forest pada klasifikasi penerima Bidikmisi se-Jawa Timur tahun 2017," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 386–391, 2022, doi: 10.47065/josyc.v3i4.2154.
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, vol. 2, pp. 1137–1143.
[22] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[23] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

Ucapan Terima Kasih

Tim Penulis mengucapkan terima kasih kepada semua rekan yang saling membantu terselesainya karya ini