Perbandingan Kinerja Metode Klasifikasi Untuk Memprediksi Putus Sekolah Dan Keberhasilan Akademik Siswa

William [1]

[1)] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Letjen S. Parman St No.1, RT.6/RW.16, Tomang, Grogol petamburan, West Jakarta City, Jakarta 11440
email : wiliamzwili135@gmail.com

ABSTRACT

Putus sekolah dan keberhasilan akademik siswa merupakan dua hal yang penting dalam pendidikan. Penelitian ini membandingkan kinerja metode klasifikasi untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Metode klasifikasi yang digunakan adalah Random Forest Classifier, AdaBoost, Decision Tree, Logistic Regression, dan XGBoost. Dataset yang digunakan berasal dari perguruan tinggi yang memiliki 4424 sampel dengan 36 fitur dan 3 kelas. Hasil penelitian menunjukkan bahwa metode Random Forest Classifier memiliki kinerja terbaik dengan akurasi 76%, diikuti oleh XGBoost 76%, AdaBoost 74%, Logistic Regression 74%, dan Decision Tree 71%. Oleh karena itu, metode Random Forest Classifier dapat digunakan untuk memprediksi putus sekolah dan keberhasilan akademik siswa dengan lebih akurat. Namun, perlu dicatat bahwa meskipun semua metode klasifikasi yang digunakan dalam penelitian ini telah mengalami perbaikan kinerja melalui penggunaan teknik ADASYN dan penyetelan parameter, mereka masih menghadapi tantangan dalam mengidentifikasi dengan akurat kasus-kasus dalam salah satu kelas minoritas. Oleh karena itu, langkah selanjutnya yang perlu diambil adalah melakukan penelitian lebih lanjut untuk mengoptimalkan parameter dengan lebih cermat dan juga mempertimbangkan pendekatan lain yang dapat lebih lanjut meningkatkan kinerja model, seperti mempertimbangkan penambahan informasi tambahan yang mungkin ada dalam dataset.

Key words

Algoritma Random Forest, Logistic Regression, Algoritma Boosting, Prediksi Keberhasilan Akademik, Putus Sekolah

1. Pendahuluan

Putus sekolah dan keberhasilan akademik siswa merupakan dua hal penting yang perlu diperhatikan dalam sistem pendidikan. Putus sekolah dapat berdampak negatif pada individu, keluarga, dan masyarakat. Sementara itu, keberhasilan akademik siswa merupakan indikator penting untuk keberhasilan pendidikan. Penyebab putus sekolah di Indonesia beragam, antara lain faktor ekonomi, faktor keluarga, faktor lingkungan, dan faktor individu. Faktor ekonomi merupakan faktor yang paling dominan. Faktor keluarga meliputi kurangnya perhatian orang tua, kekerasan dalam rumah tangga, dan konflik keluarga. Faktor lingkungan meliputi kondisi lingkungan yang tidak mendukung pendidikan, seperti tingginya angka kriminalitas dan narkoba. Faktor individu meliputi motivasi belajar yang rendah, kurangnya minat belajar, dan gangguan belajar. Keberhasilan akademik siswa juga dipengaruhi oleh berbagai faktor, antara lain faktor internal dan faktor eksternal. Faktor internal meliputi kemampuan kognitif, motivasi belajar, dan minat belajar. Faktor eksternal meliputi dukungan keluarga, dukungan guru, dan lingkungan belajar yang kondusif. Prediksi putus sekolah dan keberhasilan akademik siswa dapat dilakukan dengan menggunakan metode klasifikasi. Metode klasifikasi adalah metode yang digunakan untuk mengelompokkan data ke dalam dua atau lebih kelas. Metode klasifikasi yang digunakan untuk memprediksi putus sekolah dan keberhasilan akademik siswa antara lain metode decision tree, metode logistic regression, metode boosting, dan metode random forest.Penelitian ini bertujuan untuk membandingkan kinerja metode klasifikasi untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Penelitian ini menggunakan data dari Polytechnic Institute of Portalegre (IPP), Portugal. Data yang digunakan meliputi data siswa yang berisi variabel yang menggambarkan karakteristik demografi, sosial ekonomi, dan akademik siswa. Karakteristik demografi meliputi usia, jenis kelamin, status perkawinan, kewarganegaraan, kode alamat, dan kebutuhan khusus. Karakteristik sosial ekonomi meliputi pekerjaan pelajar, tempat tinggal orang tua, profesi orang tua, situasi pekerjaan orang tua, pelajar hibah, dan hutang siswa. Karakteristik akademik meliputi nilai masuk, tahun retensi di sekolah menengah atas, urutan pilihan mata pelajaran yang terdaftar, dan jenis mata pelajaran di sekolah menengah atas. Hasil penelitian ini diharapkan dapat memberikan informasi tentang metode klasifikasi yang paling efektif untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Informasi ini dapat digunakan untuk

mengembangkan program intervensi untuk mencegah putus sekolah dan meningkatkan keberhasilan akademik.

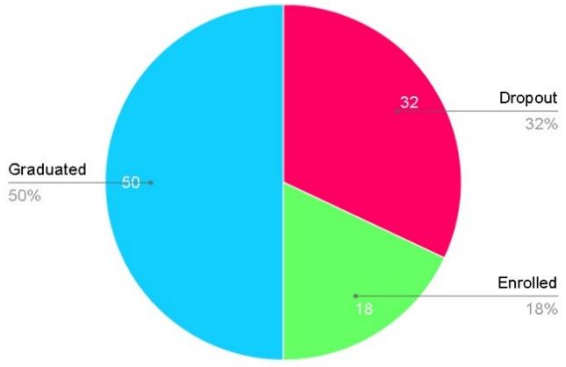
2. Metode Penelitian

Bagian dibawah ini menyajikan data yang digunakan, metode yang digunakan untuk mengatasi ketidakseimbangan data, dan menggunakannya untuk membangun dan mengevaluasi model klasifikasi.

2.1. Data

Dalam studi ini, penulis menggunakan data dari institusi yang terkait dengan mahasiswa yang mendaftar di program sarjana Polytechnic Institute of Portalegre, Portugal. Data ini mencakup informasi tentang mahasiswa yang terdaftar dari tahun akademik 2008/2009 hingga 2018/2019 dan berasal dari berbagai program sarjana seperti teknologi, layanan sosial, manajemen, jurnalisme, keperawatan, pendidikan, agronomi, desain. Data ini mencakup variabel-variabel yang berkaitan dengan faktor demografis, faktor sosial-ekonomi, serta jalur akademik mahasiswa [1].

Target data diklasifikasikan sebagai Dropout, Enrolled, dan Graduated, tergantung pada waktu yang diperlukan mahasiswa untuk lulus. Dropout berarti bahwa mahasiswa tidak berhasil lulus dalam rentang waktu yang ditentukan, Graduated berarti bahwa mahasiswa memerlukan waktu tambahan hingga tiga tahun untuk lulus, Enrolled berarti bahwa mahasiswa memerlukan lebih dari tiga tahun tambahan untuk lulus atau bahkan tidak lulus. Yang memberikan tiga tingkat risiko: mahasiswa 'risiko rendah' dengan probabilitas tinggi untuk lulus, mahasiswa 'risiko menengah/sedang', dengan probabilitas yang di mana jika diambil tindakan oleh institusi dapat berkontribusi pada keberhasilan untuk lulus, dan mahasiswa 'risiko tinggi', yang memiliki probabilitas tinggi untuk gagal.Pada Gambar 1, distribusi data di antara tiga kategori ini tidak seimbang, dengan dua kelas minoritas, yaitu Dropout dan Enrolled. Dropout mengisi 32% dari total data, dan Enrolled mengisi 18% dari total data, sedangkan kelas mayoritas, yaitu Graduated mengisi 50% dari total data.



Gambar 1. PieChart Distribusi Data

2.2. *Sampling Data Imbalanced*

Dalam penelitian ini, penulis mengatasi masalah data yang tidak seimbang dengan menggunakan metode ADASYN (Adaptive Synthetic Sampling) yang telah terintegrasi dalam pustaka imblearn [2]. Data tidak seimbang adalah kondisi di mana terdapat perbedaan yang signifikan dalam jumlah data antara kelas mayoritas dan kelas minoritas, yang dapat menyebabkan bias dalam model pembelajaran mesin terhadap kelas mayoritas. Metode ADASYN bekerja dengan menciptakan data sintetik untuk kelas minoritas berdasarkan data terdekat dari kelas minoritas tersebut [3]. Keunikan metode ini adalah bahwa ia mengambil keseimbangan antara jumlah data sintetik yang harus dibuat dan tingkat kesulitan kelas minoritas dalam diklasifikasikan oleh model. Dengan menerapkan ADASYN, penulis berhasil meningkatkan akurasi model pembelajaran mesin penulis pada data yang tidak seimbang, karena metode ini memungkinkan model untuk lebih efektif belajar dari data minoritas yang sulit diklasifikasikan.

2.3. Algoritma Klasifikasi

Dalam studi ini menggunakan sklearn untuk algoritma machine learning. Sklearn adalah library machine learning Python yang menyediakan berbagai algoritma klasifikasi [4].

2.3.1. *Logistic Regression*

Logistic regression merupakan teknik analisis statistik yang powerful dan widely used untuk menggambarkan serta memperkirakan hubungan antara satu variabel dependen dan satu atau lebih variabel independen [5]. Metode ini sangat efektif dalam situasi di mana kita ingin memprediksi hasil biner atau kategorikal berdasarkan serangkaian prediktor. Berbeda dengan regresi linear yang memprediksi nilai kontinu, logistic regression berfokus pada estimasi probabilitas suatu peristiwa terjadi, yang didasarkan pada model statistik yang menghubungkan variabel dependen (biasanya biner) dengan satu atau lebih variabel independen [6].

Algoritma logistic regression bekerja dengan menyesuaikan fungsi logistik (juga dikenal sebagai fungsi sigmoid) dengan data, yang secara efektif memetakan variabel input ke variabel output [7]. Fungsi sigmoid ini memiliki bentuk karakteristik 'S' yang membatasi output antara 0 dan 1, menjadikannya ideal untuk memodelkan probabilitas.

2.3.2. *Decision Trees*

Decision Tree adalah salah satu metode pengklasifikasi yang paling populer dan intuitif dalam pembelajaran mesin dan penambangan data. Metode ini

memprediksi nilai variabel target berdasarkan serangkaian variabel input melalui struktur hierarkis yang menyerupai pohon [8]. Dalam struktur ini, setiap node internal mewakili sebuah "tes" pada atribut tertentu, setiap cabang merepresentasikan hasil dari tes tersebut, dan setiap node daun (atau node terminal) menyimpan label kelas [9].

Kekuatan utama Decision Tree terletak pada kemampuannya untuk memecah ruang fitur yang kompleks menjadi region-region yang lebih sederhana dan dapat diinterpretasikan. Proses ini memungkinkan model untuk menangkap interaksi non-linear antara fitur-fitur dan membuat keputusan berdasarkan serangkaian aturan logis yang mudah dipahami [10].

Algoritma pembentukan Decision Tree bekerja dengan cara membagi kumpulan data menjadi subset-subset yang lebih kecil secara rekursif. Pada setiap langkah, algoritma mencari fitur dan nilai ambang batas yang optimal untuk memisahkan data, dengan tujuan memaksimalkan homogenitas (atau meminimalkan impuritas) variabel target dalam setiap subset yang dihasilkan [11]. Proses ini berlanjut hingga mencapai kriteria penghentian tertentu, seperti kedalaman maksimum pohon atau jumlah minimum sampel di node daun.

2.3.3. *Random Forests*

Random Forest adalah algoritma pembelajaran mesin yang menggunakan kumpulan decision tree untuk meningkatkan akurasi dan ketahanan model. Ini adalah jenis algoritma bagging yang menggabungkan beberapa decision tree untuk mengurangi overfitting dan meningkatkan kinerja generalisasi [13]. Random forest adalah metode ensemble untuk klasifikasi, regresi, dan tugas lainnya yang beroperasi dengan membangun sejumlah besar pohon keputusan pada waktu pelatihan dan mengeluarkan kelas yang merupakan modus dari kelas (klasifikasi) atau rata-rata prediksi (regresi) dari pohon individu [14]. Dalam hutan acak, setiap pohon keputusan dilatih berdasarkan subset acak dari data pelatihan dan subset acak fitur. Keacakan ini membantu mengurangi korelasi antar pepohonan dan meningkatkan keanekaragaman ensembel [13].

2.3.4. *Adaptive Boosting*

Adaptive Boosting, juga dikenal sebagai AdaBoost, adalah algoritma pembelajaran mesin yang menggabungkan beberapa pengklasifikasi lemah untuk membuat pengklasifikasi yang kuat [15]. Algoritma ini bekerja dengan melatih serangkaian pengklasifikasi lemah secara berulang pada kumpulan data yang sama, dengan setiap pengklasifikasi berikutnya lebih menekankan pada sampel yang salah diklasifikasikan oleh pengklasifikasi sebelumnya [16]. Pengklasifikasi terakhir adalah jumlah tertimbang dari pengklasifikasi lemah, dengan bobot ditentukan oleh keakuratannya [15].

2.3.5. *Extreme Gradient Boosting*

Extreme Gradient Boosting atau yang dikenal juga dengan nama XGBoost adalah metode pembelajaran mesin yang digunakan untuk masalah klasifikasi dan regresi. Ini adalah metode pembelajaran ensembel yang menggabungkan beberapa pohon keputusan untuk membuat prediksi [17]. XGBoost merupakan versi perbaikan dari metode Gradient Boosting, yaitu algoritma peningkatan yang menggabungkan pembelajar yang lemah untuk menciptakan pembelajar yang kuat [18]. XGBoost menggunakan algoritma penurunan gradien untuk meminimalkan fungsi kerugian dan meningkatkan akurasi model [19]. XGBoost juga memiliki beberapa parameter yang dapat disesuaikan untuk meningkatkan performa model, seperti learning rate, jumlah pohon keputusan, dan kedalaman pohon keputusan [20].

2.4. Skema Eksperimen

Data dibagi menjadi dua set, yaitu set pelatihan (80%) dan set pengujian (20%). Kemudian, setiap algoritma klasifikasi menggunakan cross validation 10-fold untuk menghindari overfitting. Ini berarti bahwa set data pelatihan dibagi menjadi 10 blok, dan pelatihan setiap algoritma klasifikasi dilakukan dengan 9 blok, sementara satu blok sisanya digunakan untuk tujuan validasi. Proses ini diulang 10 kali, sekali untuk setiap blok, sehingga memungkinkan maksimalnya jumlah pengamatan yang digunakan untuk validasi sambil menghindari overfitting. Skor estimator validasi silang rata-rata terbaik dipilih. Metodologi ini juga mencakup prosedur untuk memastikan bahwa setiap kelas diwakili dengan baik dalam setiap lipatan. Kemudian, kinerja keseluruhan dari setiap model yang terpilih dievaluasi dengan set pengujian.

Karena target data yang tidak seimbang, akurasi bukanlah metrik yang paling sesuai untuk kinerja model, karena itu adalah metrik keseluruhan yang mungkin menghasilkan nilai tinggi berdasarkan kinerja yang baik hanya untuk kelas mayoritas. Dalam penelitian ini, penulis menggunakan metrik F1, yang mempertimbangkan trade-off antara presisi dan recall. Skor F1 dihitung untuk setiap kelas, dan skor F1 rata-rata untuk tiga kelas juga dihitung. Ini adalah metrik yang digunakan untuk penyetelan hiperparameter, seperti yang akan dijelaskan selanjutnya. Untuk model yang dioptimalkan, akurasi juga dihitung sebagai metrik keseluruhan [21]. Semua model telah melalui proses penyetelan hiperparameter, yang merupakan langkah penting dalam meningkatkan kinerja mereka. Salah satu teknik yang digunakan untuk menyesuaikan hiperparameter adalah dengan melakukan pencarian grid. Ini adalah pendekatan yang sangat komprehensif di mana berbagai konfigurasi parameter diuji secara sistematis, dan yang terbaik dipilih berdasarkan hasil validasi silang. Dalam hal ini, penulis memanfaatkan

metode Grid Search CV yang disediakan oleh pustaka Scikit-learn, dengan metrik F1-score dan akurasi sebagai fokus utama untuk memaksimalkan kinerja model.

2. Hasil dan Pembahasan

Pada bagian ini, kami menyajikan hasil pengujian kinerja metode-metode klasifikasi yang telah dilatih setelah penerapan teknik ADASYN (Adaptive Synthetic Sampling) untuk mengatasi ketidakseimbangan kelas dan optimasi hyperparameter. Analisis ini bertujuan untuk memberikan pemahaman mendalam tentang efektivitas masing-masing metode dalam konteks prediksi status mahasiswa (Dropout, Enrolled, dan Graduated).

3.1. Metode Klasifikasi Standar

Tabel 1 menyajikan hasil evaluasi kinerja untuk tiga metode klasifikasi standar yang digunakan dalam penelitian ini: Logistic Regression, Decision Tree, dan Random Forest. Evaluasi dilakukan menggunakan metrik F1-score, yang merupakan rata-rata harmonik dari presisi dan recall, untuk setiap kelas ("Dropout", "Enrolled", dan "Graduated"), serta rata-rata F1-score dan akurasi keseluruhan model.

Hasil yang tercatat dalam tabel ini memberikan wawasan mendalam tentang kemampuan masing-masing metode klasifikasi dalam mengatasi kompleksitas dan ketidakseimbangan antar kelas. Metrik F1-score dipilih karena kemampuannya dalam memberikan gambaran yang lebih komprehensif tentang kinerja model, terutama ketika berhadapan dengan dataset yang tidak seimbang.Rata-rata F1-score dihitung untuk memberikan gambaran keseluruhan tentang performa metode klasifikasi. Metode Random Forest mencapai rata-rata F1-score tertinggi 0.71, diikuti oleh Logistic Regression 0.70, dan Decision Tree 0.66.

Dalam analisis prediksi status mahasiswa, tiga model machine learning yaitu, Random Forest, Logistic Regression, dan Decision Tree menunjukkan kinerja yang berbeda-beda untuk setiap kategori. Untuk kelas "Dropout", Random Forest unggul dengan F1-score 0.78, menunjukkan kemampuannya dalam menangkap interaksi kompleks antar variabel prediktor, diikuti oleh Logistic Regression 0.76 dan Decision Tree 0.72. Pada kelas "Enrolled", semua model menunjukkan kinerja yang lebih rendah, dengan Logistic Regression memimpin pada F1-score 0.51, menunjukkan tantangan dalam memprediksi status ini. Sementara itu, untuk kelas "Graduated", Random Forest kembali unggul dengan F1-score 0.84, sedikit lebih tinggi dari Logistic Regression 0.83 dan Decision Tree 0.80, menunjukkan akurasi yang baik dalam mengidentifikasi mahasiswa yang cenderung lulus. Secara keseluruhan, Random Forest memiliki rata-rata F1-score tertinggi 0.71 dan akurasi tertinggi 0.76, menandakan kinerja konsisten dan prediksi yang lebih tepat dibandingkan Logistic Regression dan Decision Tree, masing-masing dengan rata-rata F1-score 0.70 dan 0.66, serta akurasi 0.74 dan 0.71. Perbedaan ini, meskipun kecil, penting dalam konteks prediksi akademik.

Random Forest menunjukkan keunggulan dalam hampir semua metrik berkat kemampuannya menangani kompleksitas data pendidikan. Metode ini dapat menangkap interaksi non-linear antar variabel dan cenderung lebih tahan terhadap overfitting, menjelaskan kinerjanya yang superior. Sementara itu, meskipun lebih sederhana, Logistic Regression menunjukkan kinerja yang kompetitif, terutama untuk kelas "Enrolled", yang mungkin mengindikasikan adanya hubungan linear yang kuat antara beberapa variabel prediktor dan probabilitas seorang mahasiswa tetap terdaftar. Sebaliknya, Decision Tree secara konsisten menunjukkan kinerja yang sedikit lebih rendah, mungkin karena kecenderungannya untuk overfitting atau ketidakmampuannya menangkap pola yang lebih halus dalam data. Tantangan dalam memprediksi status "Enrolled" tercermin dalam F1-score yang relatif rendah di semua metode, mengindikasikan bahwa status ini paling sulit diprediksi, mungkin karena heterogenitas yang lebih tinggi atau faktor-faktor yang tidak tertangkap oleh variabel prediktor. Kinerja baik dalam memprediksi "Dropout" dan "Graduated" memberikan dasar kuat untuk program intervensi yang ditargetkan, memungkinkan institusi untuk mengalokasikan sumber daya lebih efektif dengan fokus pada mahasiswa yang berisiko dropout dan mendukung mereka yang berpotensi tinggi untuk lulus. Namun, meskipun Random Forest menunjukkan kinerja terbaik, kompleksitasnya yang lebih tinggi dibandingkan Logistic Regression mungkin menjadi pertimbangan dalam implementasi praktis. Logistic Regression, dengan kinerjanya yang kompetitif, mungkin lebih mudah diinterpretasikan dan diimplementasikan dalam sistem yang ada.

Hasil evaluasi ini menunjukkan bahwa Random Forest memberikan kinerja terbaik secara keseluruhan dalam memprediksi status mahasiswa, dengan keunggulan khusus dalam mengidentifikasi mahasiswa yang berpotensi dropout atau lulus. Namun, Logistic Regression juga menunjukkan kinerja yang kompetitif, terutama mengingat kesederhanaannya relatif terhadap Random Forest. Tantangan utama yang teridentifikasi adalah prediksi akurat untuk status "Enrolled", yang mungkin memerlukan penelitian lebih lanjut dan pengembangan model yang lebih canggih. Temuan-temuan ini memberikan dasar yang kuat untuk pengembangan sistem prediksi dan intervensi yang lebih efektif dalam konteks pendidikan tinggi.

Tabel 1. Evaluasi Metode Klasifikasi Standar

	Logistic Regression	Decision Tree	Random Forest
F1-score Dropout	0.76	0.72	0.78
F1-score Enrolled	0.51	0.46	0.50
F1-score Graduated	0.83	0.80	0.84
Rata-rata F1-score	0.70	0.66	0.71
Akurasi	0.74	0.71	0.76

2. Metode Klasifikasi Boosting

Tabel 2 menampilkan hasil evaluasi klasifikasi yang mendetail untuk dua metode yang berbeda, yaitu Adaptive Boost dan Extreme Gradient Boost. Analisis ini menggunakan metrik F1-score untuk menilai kinerja pada tiga kategori spesifik: "Dropout," "Enrolled," dan "Graduated." F1-score adalah ukuran penting dalam tugas klasifikasi karena mempertimbangkan baik presisi maupun recall, sehingga menjadi metrik yang seimbang untuk mengevaluasi efektivitas model yang digunakan.

Dari hasil yang ditabulasikan, terlihat jelas bahwa dalam kategori "Dropout," metode Extreme Gradient Boost mencapai F1-score yang sedikit lebih tinggi sebesar 0.78, dibandingkan dengan 0.76 yang diperoleh oleh Adaptive Boost. Hal ini menunjukkan bahwa Extreme Gradient Boost sedikit lebih efektif dalam mengidentifikasi siswa yang kemungkinan besar akan drop out. Sebaliknya, dalam kategori "Enrolled," Adaptive Boost mengungguli Extreme Gradient Boost, dengan meraih F1-score sebesar 0.48 dibandingkan dengan 0.46. Ini mengindikasikan bahwa Adaptive Boost mungkin memiliki keunggulan dalam mengenali siswa yang tetap terdaftar. Lebih lanjut, dalam kategori "Graduated," Extreme Gradient Boost kembali menunjukkan keunggulannya dengan F1-score sebesar 0.84, sementara Adaptive Boost sedikit tertinggal dengan F1-score sebesar 0.82. Temuan ini menyoroti kemampuan Extreme Gradient Boost yang lebih baik dalam mengklasifikasikan siswa yang berhasil lulus dengan akurat.

Ketika meninjau rata-rata F1-score di seluruh kategori ini, kedua metode klasifikasi menunjukkan kinerja keseluruhan yang serupa, dengan rata-rata F1-score sekitar 0.69. Rata-rata ini menekankan efektivitas yang sebanding dari kedua metode dalam menangani tugas klasifikasi. Namun, penting untuk dicatat bahwa keunggulan tipis dari Extreme Gradient Boost juga tercermin dalam akurasi keseluruhannya yang lebih tinggi sebesar 0.76, melampaui akurasi Adaptive Boost yang sebesar 0.74. Perbedaan kecil ini menunjukkan bahwa meskipun kedua metode sama-sama kuat, Extreme Gradient Boost mungkin menawarkan kemampuan prediktif yang sedikit lebih unggul dalam konteks ini.

Meskipun baik Adaptive Boost maupun Extreme Gradient Boost menunjukkan kekuatan di berbagai area, pemilihan antara kedua metode ini sebaiknya mempertimbangkan kategori klasifikasi spesifik yang diminati, serta pentingnya presisi dan recall dalam konteks studi. Perbedaan yang halus dalam kinerja mereka, sebagaimana yang disoroti oleh F1-score dan tingkat akurasi, memberikan wawasan berharga tentang kekuatan masing-masing dan potensi aplikasinya dalam penambangan data pendidikan dan analisis prediktif.

Tabel 2. Evaluasi Metode Klasifikasi Boosting

	Adaptive Boost	Extreme Gradient Boost
F1-score Dropout	0.76	0.78
F1-score Enrolled	0.48	0.46
F1-score Graduated	0.82	0.84
Rata-rata F1-score	0.69	0.69
Akurasi	0.74	0.76

4. Kesimpulan dan Saran

Dalam penelitian ini, penulis telah melakukan evaluasi komprehensif terhadap kinerja beberapa metode klasifikasi, meliputi Adaptive Boost, Extreme Gradient Boost, Logistic Regression, Decision Tree, dan Random Forest, dalam konteks penanganan data yang tidak seimbang. Teknik ADASYN dan GridSearchCV dengan 10-fold cross-validation telah diterapkan untuk meningkatkan ketepatan model dalam klasifikasi kelas minoritas.

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa metode klasifikasi Logistic Regression dan Random Forest secara konsisten menunjukkan kinerja unggul dibandingkan metode lainnya. Khususnya, jika dilihat dari nilai rata-rata F1-score, Logistic Regression dan Random Forest mencapai nilai tertinggi masing-masing sebesar 0.70 dan 0.71. Ini mengindikasikan bahwa kedua metode tersebut memiliki kemampuan yang lebih baik dalam mengatasi ketidakseimbangan kelas, terutama dalam memprediksi kelas minoritas, seperti "Dropout" dan "Enrolled."

Meskipun demikian, meskipun ada peningkatan kinerja yang signifikan, penting untuk dicatat bahwa metode klasifikasi ini masih memiliki ruang untuk perbaikan. Penggunaan GridSearchCV dalam mencari parameter optimal merupakan langkah positif dalam meningkatkan kinerja model. Namun, terdapat kemungkinan bahwa parameter ini dapat dioptimalkan lebih lanjut, atau metode klasifikasi yang lebih canggih dapat diterapkan untuk hasil yang lebih baik.

Secara keseluruhan, penelitian ini memberikan wawasan signifikan mengenai pendekatan untuk mengatasi masalah ketidakseimbangan data dalam klasifikasi. Random Forest muncul sebagai pilihan

terbaik dalam menangani masalah ini, dengan dukungan dari Extreme Gradient Boosting dan Logistic Regression dalam hal akurasi dan F1-score.

Namun, harus diakui bahwa meskipun semua metode klasifikasi yang digunakan dalam penelitian ini telah menunjukkan peningkatan kinerja melalui teknik ADASYN dan penyetelan parameter, tantangan masih ada dalam identifikasi kasus-kasus secara akurat di kelas minoritas. Oleh karena itu, penelitian lanjutan sangat diperlukan untuk lebih mengoptimalkan parameter yang ada dan mempertimbangkan pendekatan lain yang dapat meningkatkan kinerja model lebih lanjut. Ini mungkin termasuk eksplorasi tambahan informasi yang terdapat dalam dataset atau pengembangan model prediktif yang lebih inovatif.

REFERENSI

[1] D. and M. J. and B. L. M. T. and R. V. Martins Mónica V. and Tolledo, 2021, Early Prediction of student's Performance in Higher Education: A Case Study, Trends and Applications in Information Systems and Technologies, vol. 1, hal. 166–175.

[2] G. Lemaître, F. Nogueira, dan C. K. Aridas, 2017, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, vol. 18, no. 17, hal. 1–5, [Daring]. Tersedia pada: http://jmlr.org/papers/v18/16-365.html

[3] Haibo He, Yang Bai, E. A. Garcia, dan Shutao Li, 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, dalam 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, hal. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[4] F. Pedregosa dkk., 2011, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, vol. 12, hal. 2825–2830.

[5] K. Lu, Logistic Regression in Biomedical Study, 2022, 2022 International Conference on Biotechnology, Life Science and Medical Engineering (BLSME 2022), [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:248935866

[6] J. Friedman, T. Hastie, dan R. Tibshirani, 2000,Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), The Annals of Statistics, vol. 28, no. 2, doi: 10.1214/aos/1016218223.

[7] M. Zanchak, V. Vysotska, dan S. Albota, 2021, The Sarcasm Detection in News Headlines Based on Machine Learning Technology, dalam 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), IEEE, hal. 131–137. doi: 10.1109/CSIT52700.2021.9648710.

[8] Y. Yue, L. Jia, H. Zhai, M. Kong, dan M. Li, 2020, CFS-DT : a Combined Feature Selection and Decision Tree based Method for Octane Number Prediction, dalam 2020 4th Annual International Conference on Data Science and Business Analytics (ICDSBA), IEEE, hal. 100–103. doi: 10.1109/ICDSBA51020.2020.00033.

[9] J. R. Quinlan, 1986, Induction of decision trees, Mach Learn, vol. 1, no. 1, hal. 81–106, doi: 10.1007/BF00116251.

[10] E. Momeni, M. R. Sahebi, dan A. Mohammadzadeh, 2020, CLASSIFICATION OF HIGH-RESOLUTION SATELLITE IMAGES USING FUZZY LOGICS INTO DECISION TREE, Malaysian Journal of Geosciences, vol. 4, no. 1, hal. 07–12, doi: 10.26480/mjg.01.2020.07.12.

[11] L. Wang dan Y. Zhang, 2020, Clustering Reduction Method Analysis of Rough Set and Decision Tree based on Weight Matrix Analysis, IOP Conf Ser Mater Sci Eng, vol. 750, no. 1, hal. 012205, doi: 10.1088/1757-899X/750/1/012205.

[12] N. Nakaryakova, S. Rusakov, dan O. Rusakova, 2020, PREDICTION OF THE RISK GROUP (BY ACADEMIC PERFORMANCE) AMONG FIRST COURSE STUDENTS BY USING THE DECISION TREE METHOD, Applied Mathematics and Control Sciences, no. 4, hal. 121–136, doi: 10.15593/2499-9873/2020.4.08.

[13] S. Abdullah dan G. Prasetyo, 2020, EASY ENSEMMBLE WITH RANDOM FOREST TO HANDLE IMBALANCED DATA IN CLASSIFICATION, Journal of Fundamental Mathematics and Applications (JFMA), vol. 3, no. 1, hal. 39–46, doi: 10.14710/jfma.v3i1.7415.

[14] L. Breiman, Random Forests, Mach Learn, 2001, vol. 45, no. 1, hal. 5–32, doi: 10.1023/A:1010933404324.

[15] C. Han dan H. Jia, 2022, Multi-Modal Representation Learning with Self-Adaptive Thresholds for Commodity Verification, dalam China Conference on Knowledge Graph and Semantic Computing. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:251740987

[16] Z. Zheng dan Y. Yang, 2021, Adaptive Boosting for Domain Adaptation: Toward Robust Predictions in Scene Segmentation, IEEE Transactions on Image Processing, vol. 31, hal. 5371–5382, [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:232417741

[17] N. A. Akbar, A. Sunyoto, M. Rudyanto Arief, dan W. Caesarendra, 2020, Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm, dalam 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, hal. 110–114. doi: 10.1109/ICIMCIS51567.2020.9354286.

[18] M. Alqahtani, H. Mathkour, dan M. M. Ben Ismail, 2020, IoT Botnet Attack Detection Based on Optimized Extreme Gradient Boosting and Feature Selection, Sensors, vol. 20, no. 21, hal. 6336, doi: 10.3390/s20216336.

[19] Z. Yan dan H. Wen, 2020, Electricity Theft Detection Base on Extreme Gradient Boosting in AMI, dalam 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, hal. 1–6. doi: 10.1109/I2MTC43012.2020.9128712.

[20] T. Chen dan C. Guestrin, 2016, XGBoost, dalam Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, hal. 785–794. doi: 10.1145/2939672.2939785.

[21] D. Johannßen, C. Biemann, S. Remus, T. Baumann, dan D. Scheffer, 2020, GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text: Companion Paper. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:220320932