

KLASIFIKASI STATUS MAHASISWA MENGGUNAKAN ALGORITMA XGBOOST PADA DATASET AKADEMIK

Muhammad Rizki Andrian Fitra¹, Neysa Talitha Jehian^{2*}, Thania Dealva Arsyad³,

Ayman Human Sukma⁴, Arnita⁵, Fanny Ramadhani⁶

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email: andrian25544@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia²

Email*: jxjia05@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia³

Email: thaniadealva.4233250002@mhs.unimed.ac.id

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia⁴

Email: ahsukma15@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia⁵

Email: arnita@unimed.ac.id

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia⁶

Email: fannyr@unimed.ac.id

(*) Corresponding Author

ABSTRAK

Monitoring status akademik mahasiswa menjadi aspek penting dalam mendukung efektivitas sistem pendidikan tinggi. Penelitian ini bertujuan untuk mengklasifikasikan status mahasiswa—*Dropout*, *Enrolled*, dan *Graduate*—menggunakan algoritma *XGBoost* pada dataset akademik dari situs *Kaggle* yang terdiri dari 4424 entri dan 37 fitur. Penelitian dilakukan dengan pendekatan kuantitatif dan metode eksperimen komputasional. Tahapan meliputi pra-pemrosesan data, pemisahan data latih dan uji (rasio 80:20), pelatihan model, serta evaluasi menggunakan metrik klasifikasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil menunjukkan bahwa model *XGBoost* mampu mengklasifikasikan status mahasiswa dengan akurasi sebesar 79.9%. Evaluasi lebih lanjut mengungkapkan bahwa model bekerja baik pada kelas mayoritas (*Graduate*), namun memiliki performa rendah pada kelas minoritas (*Enrolled*), terutama *recall* sebesar 0,48. Hal ini menunjukkan tantangan dalam menghadapi distribusi kelas yang tidak seimbang. Oleh karena itu, disarankan penggunaan teknik *balancing* data, setelah penyimbangan data akurasi model mencapai sebesar 82,65%, supaya meningkat lagi performa maka perlu *tuning hiperparameter* untuk peningkatan performa model. Hasil penelitian ini diharapkan dapat membantu institusi pendidikan dalam mendeteksi potensi *dropout* lebih awal dan mendukung pengambilan keputusan berbasis data.

Kata kunci: Klasifikasi, Mahasiswa, Status akademik, *XGBoost*

ABSTRACT

Monitoring student academic status is an important aspect of supporting the effectiveness of higher education systems. This study aims to classify student status—Dropout, Enrolled, and Graduate—using the XGBoost algorithm on an academic dataset from Kaggle consisting of 4424 entries and 37 features. The research was conducted using a quantitative approach and computational experiment methods. The stages included data

pre-processing, training and testing data split (80:20 ratio), model training, and evaluation using classification metrics such as accuracy, precision, recall, and F1-score. The results showed that the XGBoost model was able to classify student status with an accuracy of 79.9%. Further evaluation revealed that the model performed well on the majority class (Graduate), but had low performance on the minority class (Enrolled), especially with recall of 0,48. This indicates challenges in dealing with imbalanced class distributions. Therefore, the use of data balancing techniques must be implemented, thus the accuracy by the model reached 82,65%, for further refining and improving its performace, hyperparameter tuning is recommended for it. The results of this study are expected to assist educational institutions in detecting potential dropouts earlier and supporting data-driven decision making.

Keywords: Academic status, Classification, Students, XGBoost

1. PENDAHULUAN

Dalam dunia pendidikan tinggi, memahami dan memantau status akademik mahasiswa merupakan hal yang sangat penting bagi institusi pendidikan. Proses *monitoring* ini mendukung pelaksanaan pembelajaran dan penilaian yang efektif karena memberikan data penting untuk pemeriksaan, pengawasan, dan pengendalian. Dengan demikian, kesalahan dalam proses akademik dapat diminimalkan, dan kualitas pembelajaran dapat terus ditingkatkan. *Monitoring* dan evaluasi idealnya berjalan secara bersamaan untuk memastikan keberlanjutan proses akademik yang optimal [1].

Status mahasiswa—seperti *dropout*, terdaftar, atau lulus—tidak hanya mencerminkan performa akademik individu, tetapi juga menjadi indikator penting dari efektivitas sistem pendidikan secara keseluruhan. Tingginya angka mahasiswa yang tidak menyelesaikan studi berdampak pada reputasi institusi, efisiensi penggunaan sumber daya, serta tingkat *retensi* mahasiswa. Selain itu, tingkat kelulusan juga merupakan indikator penting yang memengaruhi *akreditasi* suatu perguruan tinggi [2]. *Dropout* juga menimbulkan dampak sosial yang signifikan bagi mahasiswa dan keluarga mereka, karena waktu dan biaya yang telah dikeluarkan tidak menghasilkan hasil yang diharapkan [3].

Berbagai penelitian sebelumnya telah mengembangkan pendekatan *prediktif* untuk mengidentifikasi status kelulusan mahasiswa. Penelitian oleh [4] menggunakan algoritma *Random Forest* yang diimplementasikan melalui aplikasi *web* berbasis *Streamlit* (*Python*). Sementara itu, penelitian [5] membandingkan tiga algoritma *machine learning*—*XGBoost Classifier*, *Random Forest*, dan *AdaBoost Classifier*—untuk memprediksi kelulusan siswa pada mata pelajaran Matematika. Hasilnya menunjukkan bahwa *XGBoost* memiliki nilai metrik tertinggi (*Accuracy*, *Precision*, *Recall*, dan *F1-Score*) dibandingkan dua algoritma lainnya.

Berdasarkan temuan tersebut, penelitian ini memilih algoritma *XGBoost* sebagai metode klasifikasi untuk memprediksi status mahasiswa. Pemilihan ini tidak hanya didasarkan pada performa metrik yang unggul, tetapi juga karena keunggulan teknis *XGBoost* dalam menangani data yang kompleks serta kemampuannya dalam mengurangi risiko *overfitting*. *XGBoost* merupakan pengembangan dari metode *Gradient Tree Boosting* yang berbasis algoritma *ensemble* dan telah terbukti efektif dalam menyelesaikan permasalahan *machine learning* berskala besar [6]. Selain itu, algoritma ini dapat diskalakan untuk meningkatkan performa pohon keputusan [7].

Dengan semakin lengkapnya data akademik yang dikumpulkan oleh institusi pendidikan—seperti nilai, latar belakang sosial ekonomi, kehadiran, dan status

administrasi mahasiswa—pendekatan berbasis *data science* menjadi semakin relevan. Pemanfaatan teknologi untuk menganalisis data pendidikan memungkinkan institusi membuat keputusan yang lebih terbuka dan berbasis pada data yang tersedia. Pendekatan seperti *learning analytics* dan *educational data mining* perlu diintegrasikan dalam sistem pendidikan agar dapat memberikan gambaran nyata dan mendukung upaya pencegahan kegagalan belajar mahasiswa [8].

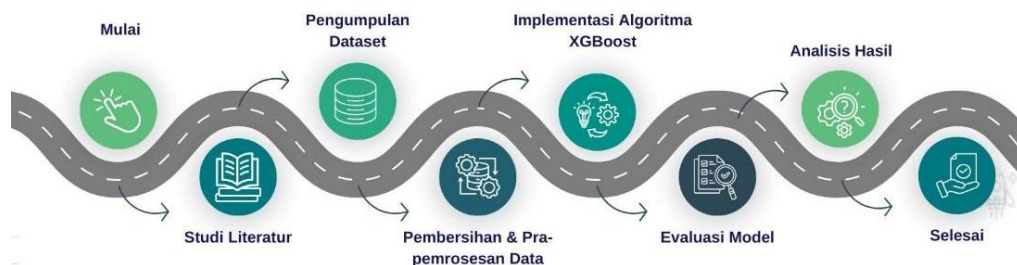
Dalam pengembangan model klasifikasi berbasis *machine learning*, proses pelatihan dan pengujian model merupakan tahapan krusial. Proses *training* adalah proses pembelajaran menggunakan data *training* untuk membentuk model, sedangkan proses *testing* dilakukan dengan data *testing* guna mengevaluasi performa model yang telah dibangun selama pelatihan [9]. Melalui dua proses ini, akurasi dan kemampuan *generalisasi* model dapat diukur secara objektif sebelum diterapkan pada data nyata.

Data akademik yang kompleks dan multidimensi memerlukan metode klasifikasi yang tangguh dan efisien dalam proses pengolahannya. *XGBoost*, sebagai salah satu algoritma *gradient boosting* yang dirancang untuk efisiensi dan performa tinggi, telah terbukti unggul dalam berbagai kompetisi *machine learning* [10]. Algoritma ini mampu menangani *dataset* berukuran besar dan fitur yang beragam tanpa kehilangan akurasi. Dengan tingkat prediksi yang tinggi dan minim *overfitting*, *XGBoost* terbukti efektif dalam mengidentifikasi faktor-faktor yang memengaruhi performa akademik mahasiswa [11]. Keunggulan lainnya adalah kemampuannya mengolah data tidak seimbang dan menyortir fitur-fitur paling berpengaruh dalam *dataset*, seperti nilai, kehadiran, latar belakang sosial ekonomi, dan status administrasi [12].

Dengan penerapan yang tepat, hasil klasifikasi ini dapat dimanfaatkan oleh institusi pendidikan untuk mendeteksi lebih dini mahasiswa yang berisiko *dropout*. Hal ini dapat menjadi dasar dalam mengoptimalkan strategi pendampingan dan bimbingan akademik. Oleh karena itu, penelitian ini tidak hanya berfokus pada penerapan algoritma *XGBoost* dalam klasifikasi status mahasiswa, tetapi juga sebagai upaya mendukung sistem pengambilan keputusan berbasis data dalam pendidikan tinggi.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional, di mana algoritma *machine learning* diterapkan pada *dataset* akademik mahasiswa untuk melakukan klasifikasi terhadap status akademik mereka (*Dropout*, *Terdaftar*, dan *Lulus*).



Gambar 1. Alur penelitian

Tahapan penelitian dimulai dari proses pengumpulan data, yaitu *dataset* yang terdiri atas beragam *variabel* seperti status pernikahan, program studi, kehadiran, nilai mata kuliah, latar belakang orang tua, serta indikator ekonomi makro seperti tingkat *inflasi* dan

Produk Domestik Bruto (PDB) yang diasumsikan berpotensi memengaruhi status akademik mahasiswa.

Langkah berikutnya adalah pra-pemrosesan data, yang mencakup pembersihan data, *encoding*, serta pemisahan *dataset* menjadi data latih dan data uji dengan rasio 80:20. Proses seleksi fitur dilakukan untuk menghindari *overfitting* dan meningkatkan akurasi model. Teknik *feature importance* dari model *XGBoost* dimanfaatkan untuk memilih variabel-variabel yang paling relevan terhadap klasifikasi.

Tahapan selanjutnya yaitu pengembangan model klasifikasi menggunakan algoritma *XGBoost*. Proses pelatihan (training) dilakukan pada data latih dengan pengaturan parameter awal (*default*), kemudian dilakukan penyetelan *hiperparameter* (*hyperparameter tuning*) menggunakan metode *Grid Search* guna memperoleh kombinasi parameter terbaik sekaligus meminimalkan risiko *overfitting*.

Setelah model selesai dilatih, dilakukan proses evaluasi dengan menggunakan beberapa metrik evaluasi klasifikasi, yakni *accuracy*, *precision*, *recall*, dan *F1-score*, serta *confusion matrix* untuk menganalisis distribusi prediksi pada setiap kelas. Hasil prediksi model dibandingkan dengan label aktual guna menilai efektivitas model, serta dilakukan analisis terhadap fitur-fitur yang paling berpengaruh dalam menentukan status mahasiswa.

3. HASIL DAN PEMBAHASAN

Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini bersumber dari situs *Kaggle* yang berisi informasi terkait status akademik mahasiswa. Dataset ini memiliki total 4424 entri dan 37 atribut, yang mencakup informasi seperti latar belakang keluarga, prestasi akademik, kondisi sosial ekonomi, dan status akhir mahasiswa (*Dropout*, *Graduate*, *Enrolled*).

Pembersihan dan Pra-Pemrosesan Data

Sebelum dilakukan proses pelatihan model, tahap pembersihan dan pra-pemrosesan data menjadi langkah yang sangat penting untuk memastikan bahwa data yang digunakan bersih, konsisten, dan siap untuk dianalisis. *Dataset* yang digunakan dalam penelitian ini terdiri dari 4424 entri dan 37 kolom. Berdasarkan hasil eksplorasi awal, diketahui bahwa dataset tidak memiliki data kosong. Hal ini dibuktikan dengan hasil pemeriksaan *df.isnull().sum()* yang menunjukkan nilai nol pada seluruh kolom. Dengan demikian, tidak diperlukan penanganan nilai hilang (*missing values*) seperti pengisian (*imputation*) atau penghapusan data.

Namun, meskipun tidak terdapat nilai kosong, ditemukan adanya karakter tidak lazim, yaitu karakter *tab* (*\t*) pada nama salah satu kolom, yaitu "*Daytime/evening attendance\t*". Karakter ini dapat menyebabkan *error* atau ketidaksesuaian saat pengolahan data, terutama saat pemanggilan kolom secara *eksplisit*. Untuk mengatasi hal ini, dilakukan proses pembersihan nama kolom menggunakan fungsi *.str.strip()* dari pustaka *Pandas*, yang akan menghapus spasi dan karakter yang tidak diinginkan di awal atau akhir nama kolom.

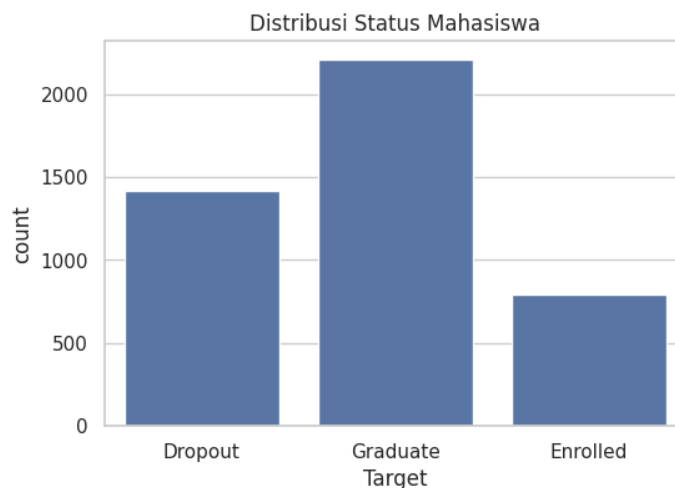
Selanjutnya, kolom target "Target" yang semula bertipe objek (*string*) dan berisi label kategorikal seperti "*Graduate*", "*Dropout*", dan "*Enrolled*", perlu diubah menjadi format numerik agar dapat digunakan dalam pelatihan model klasifikasi. Proses ini dilakukan menggunakan teknik *encoding* dengan *Label Encoder* dari pustaka *sklearn.preprocessing*. Hasil *encoding* ini menghasilkan label numerik untuk setiap kelas sebagai berikut:

Tabel 1. Tabel numerik target

Label Asli	Label Numerik
Dropout	0
Graduate	1
Enrolled	2

Visualisasi Distribusi Kelas

Sebelum dilakukan proses pelatihan model, tahap pembersihan dan pra-pemrosesan data menjadi langkah yang sangat penting untuk memastikan bahwa data yang digunakan bersih, konsisten, dan siap untuk dianalisis. *Dataset* yang digunakan dalam penelitian ini terdiri dari 4424 *entri* dan 37 kolom. Berdasarkan hasil eksplorasi awal, diketahui bahwa *dataset* tidak memiliki data kosong. Hal ini dibuktikan dengan hasil pemeriksaan *df.isnull().sum()* yang menunjukkan nilai nol pada seluruh kolom. Dengan demikian, tidak diperlukan penanganan nilai hilang (*missing values*) seperti pengisian (*imputation*) atau penghapusan data.



Gambar 2. Visualisasi distribusi status mahasiswa

Dari grafik tersebut terlihat bahwa kelas *Graduate* merupakan kelas yang paling dominan, disusul oleh *Dropout*, dan yang paling sedikit adalah kelas *Enrolled*. Ketidakseimbangan distribusi ini penting untuk diperhatikan, karena dapat memengaruhi performa model klasifikasi. Model cenderung bias terhadap kelas mayoritas jika tidak dilakukan penanganan ketidakseimbangan data, misalnya dengan teknik *re-sampling* seperti *SMOTE* atau *undersampling*.

Untuk memberikan gambaran yang lebih jelas mengenai proporsi masing-masing kelas, berikut adalah tabel distribusi jumlah data berdasarkan kelas target:

Tabel 2. Distribusi kelas target

Kelas	Label Numerik	Jumlah Mahasiswa
Dropout	0	1421

Kelas	Label Numerik	Jumlah Mahasiswa
Graduate	1	2319
Enrolled	2	684
Total		4424

Distribusi yang tidak seimbang ini menunjukkan bahwa model perlu diuji secara cermat, terutama pada kemampuan klasifikasi untuk kelas minoritas seperti *Enrolled*, agar hasil evaluasi tidak bias dan tetap akurat pada seluruh kategori.

Pelatihan Model

Model pembelajaran mesin yang digunakan dalam penelitian ini adalah *XGBoost Classifier*, yaitu implementasi algoritma *Extreme Gradient Boosting* yang sangat efisien untuk *data tabular*. *XGBoost* memiliki keunggulan dalam hal akurasi dan kecepatan pelatihan karena menerapkan teknik *boosting* berbasis pohon keputusan secara iteratif untuk meminimalkan kesalahan prediksi.

Pada proses pelatihan ini, model digunakan dengan parameter *default*, dengan pengecualian pada parameter *eval_metric* yang disetel ke '*mlogloss*', sesuai dengan kebutuhan klasifikasi *multi-class*. Parameter ini digunakan untuk mengevaluasi performa model selama proses pelatihan menggunakan *logarithmic loss* sebagai fungsi kerugian.

Model dilatih menggunakan data yang telah dibagi menjadi data latih dan data uji dengan rasio 80:20. Pemilihan rasio 80:20 dalam pembagian data latih dan data uji didasarkan pada pertimbangan keseimbangan antara kebutuhan pelatihan model dan validitas evaluasi performa. Rasio ini memberikan cukup banyak data bagi model untuk mempelajari pola secara optimal, serta menyediakan data uji yang cukup untuk mengukur kemampuan *generalisasi model*. Penelitian yang dilakukan [13] mendukung penggunaan rasio ini, di mana dalam studi mengenai prediksi tingkat obesitas menggunakan algoritma *LightGBM*, rasio 80:20 menghasilkan akurasi hingga 97%. Hasil ini menunjukkan bahwa rasio 80:20 efektif untuk digunakan pada *dataset* berukuran menengah dengan kompleksitas tinggi, seperti halnya *dataset* akademik yang digunakan dalam penelitian ini. Dengan demikian, pemilihan rasio 80:20 tidak hanya sejalan dengan praktik umum dalam *machine learning*, tetapi juga didukung oleh bukti empiris terkini yang relevan.

Setelah proses pelatihan selesai, model diuji menggunakan data uji dan menghasilkan akurasi sebesar 79,9%. Setelah pelatihan, dilakukan prediksi terhadap data uji, dan hasil akurasi model mencapai 79,9%. Nilai ini termasuk kategori cukup baik, karena secara umum akurasi $\geq 70\%$ sudah dianggap memadai dalam klasifikasi. Hal ini juga terlihat pada penelitian klasifikasi penyakit jantung menggunakan K-NN, dan klasifikasi botol plastik dengan SVM, di mana akurasi 70% dinilai sebagai hasil yang baik [14] [15]. Sedangkan dengan menyeimbangkan data dengan *SMOTE*, hasil akurasi model naik menjadi 82,65%, berhasil melewati target awal harapan akurasi model.

Tabel 3. Parameter model *xgboost*

Parameter	Nilai
Booster	gbtree (default)
Objective	multi:softprob
Eval Metric	mlogloss
Learning Rate	0.1

Parameter	Nilai
Max Depth	4
n_estimators	200

Tabel 4. Hasil akurasi model

Metode	Akurasi (%)
XGBoost Classifier	79.9
XGBoost dengan penimbangan data	82.65

Nilai akurasi ini menunjukkan bahwa *XGBoost* mampu melakukan klasifikasi yang cukup baik terhadap *dataset* ini, meskipun nilai ini masih dapat ditingkatkan dengan penyesuaian *hyperparameter* atau penanganan distribusi kelas yang tidak seimbang.

Evaluasi Model

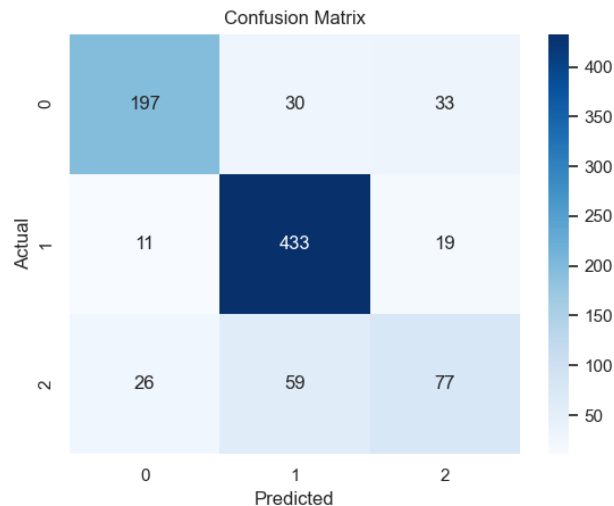
Evaluasi performa model *XGBoost* dilakukan menggunakan dua pendekatan utama, yaitu *confusion matrix* dan *classification report*. *Confusion matrix* memberikan gambaran distribusi prediksi model terhadap kelas-kelas aktual, sedangkan *classification report* menyajikan metrik evaluasi seperti *precision*, *recall*, dan *F1-score* untuk masing-masing kelas.

Berdasarkan hasil pelatihan, *confusion matrix* yang diperoleh ditunjukkan pada Gambar 3. Dari visualisasi tersebut dapat diamati bahwa sebagian besar prediksi model berada pada kelas "*Graduate*", yang memang merupakan kelas mayoritas pada *dataset* ini. Namun, terdapat pula sejumlah *misclassifications* pada kelas "*Dropout*" dan "*Enrolled*", yang menunjukkan bahwa model masih menghadapi tantangan dalam membedakan dua kelas minoritas tersebut.

Classification report yang dihasilkan memperlihatkan metrik evaluasi sebagai berikut:

Tabel 5. *Classification report*

Label	Precision	Recall	F1-Score	Support
Dropout	0.84	0.76	0.80	260
Enrolled	0.60	0.48	0.53	162
Graduate	0.83	0.94	0.88	463
Accuracy			0.80	885
Macro Avg	0.76	0.72	0.74	885
Weighted Avg	0.79	0.80	0.79	885



Gambar 3. *Confusion matrix*

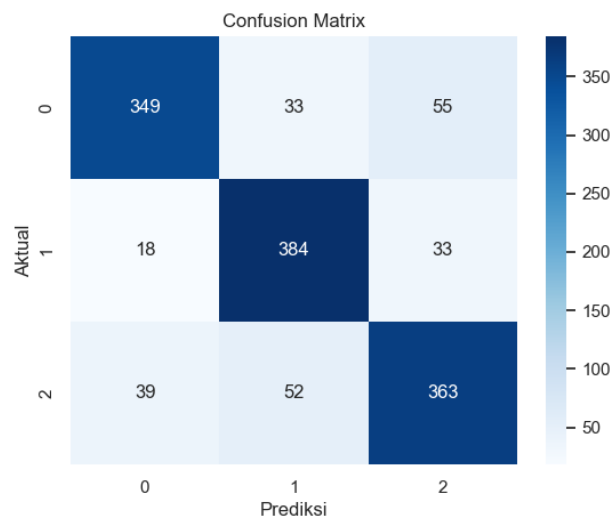
Berdasarkan tabel di atas, dapat dilihat bahwa:

1. Kelas "*Graduate*" memiliki nilai *precision* dan *recall* yang tinggi, sedangkan kelas "*Dropout*" dan "*Enrolled*" memiliki nilai *precision* dan *recall* yang berbeda akibat ketidakseimbangan data.
2. Kelas "*Enrolled*" memiliki performa yang rendah, dengan *recall* hanya sebesar 0.48, yang menunjukkan bahwa sebagian besar mahasiswa yang masih terdaftar salah diklasifikasikan sebagai *Dropout* atau *Graduate*.

Hal ini mengindikasikan adanya ketidakseimbangan kelas yang berdampak pada akurasi model dalam mengenali kelas minoritas. Untuk mengatasi hal ini, dapat dilakukan metode penyeimbangan kelas melalui oversampling *SMOTE*. Jumlah data pada tiap kelas akan disetarakan dengan jumlah data dari kelas mayoritas yaitu kelas "*Graduate*", sehingga ketiga kelas memiliki ukuran data yang sama yaitu sebanyak 2319 data.

Tabel 6. *Classification Report* setelah penyeimbangan data

Label	Precision	Recall	F1-Score	Support
Dropout	0.86	0.80	0.83	437
Enrolled	0.82	0.88	0.85	435
Graduate	0.80	0.80	0.80	454
Accuracy			0.83	1326
Macro Avg	0.83	0.83	0.83	1326
Weighted Avg	0.83	0.83	0.83	1326



Gambar 4. *Confusion Matrix* setelah penimbangan data

Berdasarkan tabel di atas, dapat dilihat bahwa:

1. Ketiga kelas "Graduate", "Dropout" dan "Enrolled" telah memiliki nilai *precision* dan *recall* yang tinggi, rata-rata diatas 0.80 akibat setelah menimbangan data.
2. Kelas "Graduate" dan "Dropout" memiliki performa yang paling rendah, dengan *recall* sebesar 0.80, tetapi nilainya sudah mendekati dan merata terhadap kelas "Enrolled" akibat setelah menimbangan data.

Tabel 7. Perbandingan hasil yang diharapkan dan hasil aktual

Aspek	Hasil yang Diharapkan	Hasil Aktual
Algoritma	XGBoost mampu mengklasifikasi dengan akurasi tinggi ($\geq 80\%$)	Akurasi model sebesar 79.9% tanpa penimbangan data, sedangkan setelah menimbangan data maka akurasi naik menjadi 82,65%
Distribusi Prediksi Kelas	Model mampu mengenali semua kelas dengan seimbang	Model sangat baik mengenali Graduate, cukup baik pada Dropout, lemah pada Enrolled, setelah penimbangan data, ketiga kelas mampu dikenal oleh model dengan baik
Ketepatan Prediksi	Precision dan recall merata antar kelas	Precision Dropout: 0.86, Enrolled: 0.82, Graduate: 0.80
Generalisasi Model	Diharapkan mampu bekerja baik di data uji	Performa cukup baik dengan pengklasifikasian saja, ternyata lebih baik lagi melalui penimbangan data
Keperluan Praproses Tambahan	Minim praproses karena data bersih	Praproses dilakukan untuk encoding dan perbaikan nama kolom

Tabel 6 menyajikan perbandingan antara ekspektasi awal dari penelitian dengan hasil aktual yang diperoleh dari implementasi di *Google Colab* menggunakan algoritma *XGBoost*. Secara umum, model menunjukkan performa yang cukup baik dengan akurasi sebesar 79.9%, meskipun sedikit di bawah target awal yaitu 80%. Hal ini menunjukkan

bahwa model mampu melakukan klasifikasi dengan cukup baik, namun masih terdapat ruang untuk perbaikan melalui penyimbangan data.

Dalam hal distribusi prediksi kelas, harapannya adalah model mampu mengenali semua kelas secara seimbang. Namun, hasil menunjukkan bahwa kelas *Graduate* memiliki akurasi dan *recall* tertinggi, sementara *Enrolled* menjadi kelas yang paling sulit diprediksi secara akurat. Hal ini kemungkinan besar disebabkan oleh ketidakseimbangan distribusi data atau pola kompleks dalam fitur kelas tersebut. Ketepatan model dalam memprediksi ditunjukkan melalui nilai *precision* dan *recall* yang bervariasi antar kelas. Kelas *Dropout* memperoleh *precision* cukup tinggi (0.84), menunjukkan bahwa model cukup yakin dalam mengidentifikasi mahasiswa yang keluar. Sementara itu, kelas *Enrolled* memiliki *precision* dan *recall* paling rendah, menandakan perlunya perhatian khusus dalam penanganan data kelas ini.

Setelah menerapkan penyimbangan data dengan *SMOTE*, performa model berhasil mendapatkan akurasi sebesar 82.65% dan ketiga kelas dapat dikenal oleh model secara seimbang, nilai *precision* dan *recall* tiap kelas merata dan rata-rata diatas 0.80.

Dari sisi generalisasi, model menunjukkan performa yang baik terhadap data uji. Namun, untuk memastikan konsistensi performa dan mengurangi risiko *overfitting*, disarankan penggunaan teknik validasi silang seperti *k-fold cross-validation*. Terakhir, meskipun data awalnya tidak mengandung nilai kosong (*null*), tetap dilakukan beberapa tahapan pra-pemrosesan, seperti pembersihan nama kolom dan *encoding* label target ke format numerik. Hal ini menunjukkan bahwa pra-pemrosesan tetap menjadi langkah penting bahkan ketika data tampak bersih.

4. KESIMPULAN DAN SARAN

Kesimpulan

Hasil pelatihan model *XGBoost* menunjukkan akurasi sebesar 79.9%, yang tergolong cukup baik untuk kasus klasifikasi *multi-class* dengan distribusi data yang tidak seimbang. Berdasarkan *confusion matrix* dan *classification report*, model menunjukkan performa terbaik pada kelas *Graduate*, diikuti oleh *Dropout*, sementara *Enrolled* memiliki performa terendah dengan nilai *recall* sebesar 0.48. Hal ini mengindikasikan bahwa sebagian besar mahasiswa yang masih aktif sering salah diklasifikasikan, kemungkinan akibat jumlah data yang sedikit atau fitur yang tumpang tindih antar kelas. Secara umum, model belum sepenuhnya mencapai ekspektasi awal untuk menghasilkan performa yang seimbang pada seluruh kelas. Hanya dua kelas yang menunjukkan *F1-score* > 0.7, sedangkan kelas *Enrolled* masih tertinggal. Oleh karena itu, dibutuhkan pendekatan tambahan seperti penyeimbangan data, dengan menerapkan *SMOTE*, hasil pelatihan model *XGBoost* menunjukkan akurasi sebesar 82,65%, melewati harapan awal dan menghasilkan performa yang seimbang pada ketiga kelas.

Saran

Penelitian selanjutnya disarankan untuk menambahkan fitur eksternal yang relevan, seperti data ekonomi regional, latar belakang akademik siswa, atau aspek psikologis dan motivasional. Selain itu, eksplorasi algoritma lain seperti *Random Forest*, *SVM*, atau *Neural Network* dapat digunakan untuk perbandingan performa model. Evaluasi juga sebaiknya diperluas dengan teknik validasi silang, seperti *k-fold cross-validation*, guna memastikan kestabilan dan generalisasi model terhadap data baru. Penelitian ini diharapkan dapat memberikan kontribusi dalam upaya deteksi dini mahasiswa berisiko

dropout, serta mendukung pengambilan keputusan berbasis data dalam lingkungan pendidikan tinggi.

5. DAFTAR PUSTAKA

- [1] M. Daheri and J. Warlizasusi, "Sistem monitoring perkembangan akademik peserta didik di sekolah," *MATAAZIR: Jurnal Administrasi dan Manajemen Pendidikan*, vol. 5, no. 2, pp. 208–217, 2024, doi: <https://doi.org/10.56874/jamp.v5i2.1812>.
- [2] A. Salim, N. Afni, R. Komarudin, and Y. I. Maulana, "Prediksi kelulusan mahasiswa dengan metode Naive Bayes," *Technologia: Jurnal Ilmiah*, vol. 13, no. 3, pp. 207–214, 2022, doi: [10.31602/tji.v13i3.7312](https://doi.org/10.31602/tji.v13i3.7312).
- [3] U. F. Laili, C. Umatin, and M. U. Ridwanulloh, "Analisis potensial drop out mahasiswa dengan K-Means++ clustering dalam upaya peningkatan kualitas IAIN Kediri," *Paedagoria: Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, vol. 14, no. 2, pp. 145–153, 2023, doi: <https://doi.org/10.31764/paedagoria.v14i2.14077>.
- [4] M. Putra and E. Harahap, "Machine learning pada prediksi kelulusan mahasiswa menggunakan algoritma Random Forest," *Jurnal Riset Matematika*, pp. 127–136, 2024, doi: <https://doi.org/10.29313/jrm.v4i2.5102>.
- [5] S. A. A. Kharis, A. Zili, E. Zubir, and F. I. Fajar, "Prediksi kelulusan siswa pada mata pelajaran matematika menggunakan educational data mining," *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 7, no. 1, pp. 21–29, 2023, doi: [10.21009/jrpms.071.03](https://doi.org/10.21009/jrpms.071.03).
- [6] S. E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan metode Extreme Gradient Boosting (XGBOOST) pada klasifikasi nasabah kartu kredit," *Journal of Mathematics: Theory and Applications*, pp. 21–26, 2022, doi: <https://doi.org/10.31605/jomta.v4i1.1792>.
- [7] R. G. Gunawan, E. S. Handika, and E. Ismanto, "Pendekatan machine learning dengan menggunakan algoritma XGBoost (Extreme Gradient Boosting) untuk peningkatan kinerja klasifikasi serangan SYN," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 3, pp. 453–463, 2022, doi: <https://doi.org/10.37859/coscitech.v3i3.4356>.
- [8] S. A. A. Kharis and A. H. A. Zili, "Learning analytics dan educational data mining pada data pendidikan," *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 6, no. 1, pp. 12–20, 2022, doi: [10.21009/jrpms.061.02](https://doi.org/10.21009/jrpms.061.02).
- [9] I. M. B. Adnyana, "Penerapan algoritma Support Vector Machine (SVM) untuk prediksi waktu tunggu alumni mendapatkan pekerjaan," *Jurnal Teknologi Informasi dan Komputer*, vol. 9, no. 1, 2023, doi: <https://doi.org/10.36002/jutik.v9i1.2420>.
- [10] R. Syafei and D. A. Efrilianda, "Machine learning model using Extreme Gradient Boosting (XGBoost) feature importance and Light Gradient Boosting Machine (LightGBM) to improve accurate prediction of bankruptcy," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 64–72, 2023, <https://doi.org/10.15294/rji.v1i2.71229>.
- [11] Y. N. Sukmaningtyas, R. M. Akbar, and G. R. U. Asyafiiyah, "Penerapan predictive

- analytics untuk analisis faktor-faktor yang mempengaruhi performa akademik siswa,” *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 2, pp. 127–145, 2024, doi: <https://doi.org/10.29240/arcitech.v4i2.12048>.
- [12] G. Velarde, A. Sudhir, S. Deshmane, A. Deshmunkh, K. Sharma, and V. Joshi, “Evaluating XGBOOST for balanced and imbalanced data: Application to fraud detection,” *arXiv*, pp. 1-17, 2023, doi: <https://doi.org/10.48550/arXiv.2303.15218>.
- [13] N. F. Fahrudin, K. R. Putra, S. Umaroh, and G. B. Lautan, “Influence of Data Scaling and Train/Test Split Ratios on LightGBM Efficacy for Obesity Rate Prediction,” *MIND (Multimedia Artificial Intelligent Networking Database) Journal*, vol. 9, no. 2, pp. 220–234, 2024, doi: <https://doi.org/10.26760/mindjournal.v9i2.220-234>.
- [14] M. E. I. Lestari, "Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung," *Faktor Exacta*, vol. 7, no. 4, pp. 366–371, 2015, doi: 10.25126/jtiik.202073622.
- [15] D. Nasien, R. Darwin, A. Cia, A. L. Winata, J. Go, R. MC, R. C. Wijaya, and K. C. Lo, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," *JEKIN - Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 10–17, 2024, doi: <https://doi.org/10.58794/jekin.v4i1.640>.