JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

Peningkatan Performa Model Machine Learning XGBoost Classifier melalui Teknik Oversampling dalam Prediksi Penyakit AIDS

Duta Firdaus Wicaksono, Ruri Suko Basuki*, Dicky Setiawan

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]112202006480@mhs.dinus.ac.id, [2,*]ruri.basuki@dsn.dinus.ac.id, [3]111202013089@mhs.dinus.ac.id
Email Penulis Korespondensi: ruri.basuki@dsn.dinus.ac.id

Abstrak−Data menunjukkan bahwa HIV (Human Immunodeficiency Virus) telah menyebabkan puluhan juta kematian global, dengan 630.000 orang meninggal akibat penyakit terkait HIV pada tahun 2022 dan 1,3 juta orang tertular HIV. Tanpa adanya pengobatan, HIV dapat berkembang menjadi AIDS (Acquired Immune Deficiency Syndrome), yang dapat melemahkan sistem kekebalan tubuh dan meningkatkan risiko terhadap infeksi dan penyakit lainnya. Meskipun telah ada kemajuan dalam pengobatan, deteksi dini AIDS tetap menjadi prioritas. Penelitian ini mengembangkan model prediksi AIDS dengan menggunakan machine learning, yang menjadi solusi efektif dalam memberikan prediksi kesehatan masa depan. Namun, masalah ketidakseimbangan data menyulitkan model dalam memprediksi kasus AIDS yang jarang terjadi. Untuk mengatasi ini, teknik oversampling digunakan untuk mengimbangi distribusi kelas minoritas. Penelitian ini mengeksplorasi teknik oversampling seperti SMOTE, ADASYN, dan Random Oversampling, serta menggabungkannya dengan algoritma XGBoost. Hasilnya menunjukkan bahwa kombinasi teknik Random Oversampling dengan XGBoost Classifier memberikan performa terbaik dengan akurasi 94,44%, presisi 90,72%, recall 98,74%, dan f1_score 94,65%. Penelitian ini diharapkan memberikan wawasan berharga bagi praktisi kesehatan dan masyarakat dalam upaya mengendalikan penyebaran penyakit AIDS secara global.

Kata Kunci: AIDS; Imbalance Data; Machine Learning; Oversampling; Prediksi; XGBoost Classifier

Abstract−The data shows that HIV (Human Immunodeficiency Virus) has caused tens of millions of global deaths, with 630,000 people dying from HIV-related illnesses in 2022 and 1.3 million people newly infected with HIV. Without treatment, HIV can progress to AIDS (Acquired Immune Deficiency Syndrome), weakening the immune system and increasing the risk of infections and other diseases. Despite advancements in treatment, early detection of AIDS remains a priority. This research develops an AIDS prediction model using machine learning, which proves to be an effective solution in providing future health predictions. However, data imbalance issues challenge the model in predicting rare AIDS cases. To solve this problem, oversampling techniques are employed to balance the distribution of minority classes. This study explores oversampling techniques such as SMOTE, ADASYN, and Random Oversampling, combined with the XGBoost algorithm. The results show that the combination of Random Oversampling technique with the XGBoost Classifier yields the best performance with an accuracy of 94.44%, precision of 90.72%, recall of 98.74%, and an f1_score of 94.65%. This research is expected to provide valuable insights for healthcare practitioners and the public in efforts to control the spread of AIDS globally.

Keywords: AIDS; Imbalance Data; Machine Learning; Oversampling; Prediction; XGBoost Classifier

1. PENDAHULUAN

Menurut data yang dihimpun dari WHO, HIV (Human Immunodeficiency Virus) sejauh ini telah merenggut 40,4 juta nyawa manusia di seluruh dunia dan pada tahun 2022, 630.000 orang meninggal karena penyakit terkait HIV dan 1,3 juta orang tertular HIV secara global. Masih dalam laman yang sama, dijelaskan bahwa jika HIV tidak diobati, maka dapat mengalami perkembangan menjadi AIDS (Acquired Immune Deficiency Syndrome) [1]. AIDS terjadi karena gejala yang timbul disebabkan oleh menurunnya fungsi sistem kekebalan tubuh yang diakibatkan oleh infeksi HIV, sehingga meningkatkan kerentanan individu terhadap berbagai jenis infeksi dan penyakit lainnya [2] [3]. AIDS telah menjadi fokus utama perhatian global sejak pertama kali diidentifikasi pada tahun 1980-an. Berdasarkan informasi pada [4], AIDS ditemukan pertama kalinya pada seseorang bernama Ken Horne yang merupakan seorang pekerja seks gay di San Francisco, California. Keprihatinan akan AIDS terkait dengan dampak kesehatan yang serius bagi individu yang terinfeksi dan dampaknya yang luas terhadap masyarakat secara keseluruhan. HIV-AIDS saat ini masih terus menjadi tantangan utama dalam masalah kesehatan masyarakat global [5], khususnya di negara-negara berkembang termasuk Indonesia. Meskipun telah ada kemajuan signifikan dalam pengobatan dan pencegahan AIDS, deteksi dini dan prediksi risiko infeksi tetap menjadi prioritas dalam upaya mengendalikan penyebaran penyakit ini.

Pengembangan model prediksi AIDS ini penting karena dapat memberikan kontribusi signifikan dalam meningkatkan efektivitas sistem pemantauan dan deteksi dini AIDS, yang pada gilirannya dapat membantu dalam upaya pencegahan penyebaran penyakit ini. Machine learning sebagai metode yang seringkali digunakan dalam pengembangan model prediksi, merupakan salah satu bidang dalam ranah teknis yang berfokus pada data, terletak di persimpangan antara ilmu komputer serta statistika, dan merupakan fondasi utama dari kecerdasan buatan serta analisis data [6]. Sama halnya dengan yang dikutip dari [7], machine learning bisa juga diartikan sebagai penerapan algoritma matematika dan komputer yang menggunakan pemahaman dari data untuk menghasilkan prediksi untuk masa depan. Proses pengambilan keputusan dapat dioptimalkan dengan menyajikan insight yang bersifat prediktif dan berjalan secara otomatis dengan menerapkan machine learning [8]. Berdasarkan uraian tersebut, tentunya

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

machine learning dapat menjadi solusi yang efektif dalam memberikan hasil prediksi dalam bidang kesehatan terkhusus pada penelitian ini yaitu prediksi penyakit AIDS.

Dalam penerapan model prediksi ini terdapat masalah ketidakseimbangan kelas dalam data. Data tidak seimbang (imbalance data) terjadi ketika terdapat satu atau beberapa kelas yang dominan terhadap keseluruhan data sebagai kelas mayoritas, sementara kelas lainnya merupakan kejadian yang jarang terjadi sebagai kelas minoritas [9]. Ketidakseimbangan ini sering kali mengarah pada kinerja model yang tidak memuaskan dalam memprediksi kasus-kasus yang langka atau minoritas, seperti kasus-kasus infeksi AIDS pada populasi yang rentan. Ketika metode klasifikasi diterapkan pada data yang tidak seimbang, pengklasifikasian cenderung mengabaikan peluang dari kelas minoritas karena kecenderungan nilai prediksi akan terpolarisasi pada kategori mayoritas [10]. Adanya data yang tidak seimbang membuat performa metode klasifikasi pada machine learning menjadi menurun [11]. Teknik oversampling menjadi salah satu pendekatan yang populer diterapkan sebagai upaya penyelesaian masalah kelas data yang tidak seimbang dalam machine learning. Pemilihan teknik oversampling disebabkan oleh kemampuannya untuk menghasilkan hasil yang lebih optimal daripada teknik undersampling [12]. Dengan oversampling, maka akan membuat salinan data dari kelas minoritas untuk menyamakan distribusi kelas dengan kelas mayoritas, sehingga memungkinkan model untuk belajar dengan lebih baik dari kedua kelas.

Berbagai penelitian terkait prediksi penyakit dengan memanfaatkan penerapan algoritma machine learning telah banyak dilakukan, terlebih dengan mengkombinasikan beberapa teknik oversampling yang ada. Selain pada penyakit AIDS, penelitian [11] menggunakan teknik random undersampling dan SMOTE untuk mendeteksi kanker serviks menggunakan algoritma naïve bayes, hasil menunjukkan bahwa SMOTE memberikan akurasi yang lebih tinggi dalam menghasilkan prediksi yakni sebesar 81,73%. Terdapat pula penelitian [13] juga menerapkan teknik SMOTE dengan beberapa algoritma machine learning dalam menangani ketidakseimbangan data dalam prediksi infeksi penyakit HIV, hasilnya menunjukkan bahwa penggunaan teknik SMOTE dan algoritma random forest menghasilkan akurasi tertinggi yakni 87,1%. Berbagai penyakit pun juga dapat diprediksi dengan menggunakan algoritma machine learning, seperti pada penelitian [14] pun melakukan prediksi literasi kesehatan penyakit menular dengan memanfaatkan teknik SMOTE dan algoritma logistic regression yang menghasilkan akurasi tertinggi sebesar 93,8% dimana akurasi ini meningkat dari yang sebelumnya tanpa dilakukan SMOTE hanya sebesar 82,8%. Berbeda dengan penelitian [15], setelah sebelumnya menggunakan SMOTE, penelitian ini menggunakan teknik ADASYN dalam melakukan oversampling, hasilnya menunjukkan peningkatan akurasi yang signifikan dari yang semula 80,13% menjadi 94,24% dengan menggunakan algoritma voting classifier (VC) yang merupakan gabungan dari random forest, KNN, dan logistic regression. Kemudian pada penelitian [16] juga mengembangkan model untuk melakukan prediksi penyakit HIV dengan mengaplikasikan teknik SMOTE bersamaan dengan beberapa algoritma, hasil menunjukkan bahwa juga terdapat peningkatan akurasi dari 76,5% menjadi 82,36% dengan menggunakan algoritma random forest. Bukti dari beberapa penelitian menunjukkan bahwa terjadi peningkatan kinerja atau performa yang signifikan setelah penerapan beberapa teknik oversampling dalam menanganai ketidakseimbangan data, namun penelitian yang secara khusus mengangkat pada permasalahan HIV-AIDS dengan mengaplikasikan beberapa teknik oversampling dan algoritma machine learning masih jarang ditemui dan tentunya menjadikan peluang untuk dapat terus digali.

Dalam upaya menanggapi tantangan tersebut serta mengacu pada temuan-temuan sebelumnya, penelitian ini bertujuan untuk melakukan eksperimen yang mengeksplorasi berbagai teknik oversampling yang tersedia, serta menggabungkannya dengan salah satu algoritma machine learning. Beberapa teknik oversampling yang dimanfaatkan dalam penelitian ini diantaranya SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), dan Random Oversampling. Kemudian secara khusus XGBoost (Extreme Gradient Boosting) dipilih sebagai algoritma machine learning yang akan digunakan dalam penelitian ini. Dengan demikian, tujuan utama dari penelitian ini adalah untuk mengevaluasi efektivitas teknik oversampling dalam meningkatkan performa model XGBoost dalam memprediksi penyakit AIDS. Secara khusus, penelitian ini akan mengukur peningkatan dalam metrik evaluasi seperti akurasi, presisi, recall, dan F1-score setelah menerapkan teknik oversampling. Penelitian ini juga bertujuan untuk mengevaluasi kestabilan model yang dihasilkan dan mengidentifikasi apakah terdapat peningkatan yang signifikan dalam kemampuan model untuk mengklasifikasikan kasus-kasus AIDS yang langka atau minoritas. Harapan dari penelitian ini adalah bahwa dengan mengimplementasikan teknik oversampling maka dapat meningkatkan performa model dalam memprediksi kasus AIDS dengan akurasi yang lebih tinggi daripada model yang tidak menggunakan oversampling. Dengan demikian, penelitian ini diharapkan mampu menyajikan pemahaman yang bernilai dan berharga bagi praktisi kesehatan dan masyarakat dalam upaya mereka untuk mengendalikan penyebaran penyakit AIDS secara global.

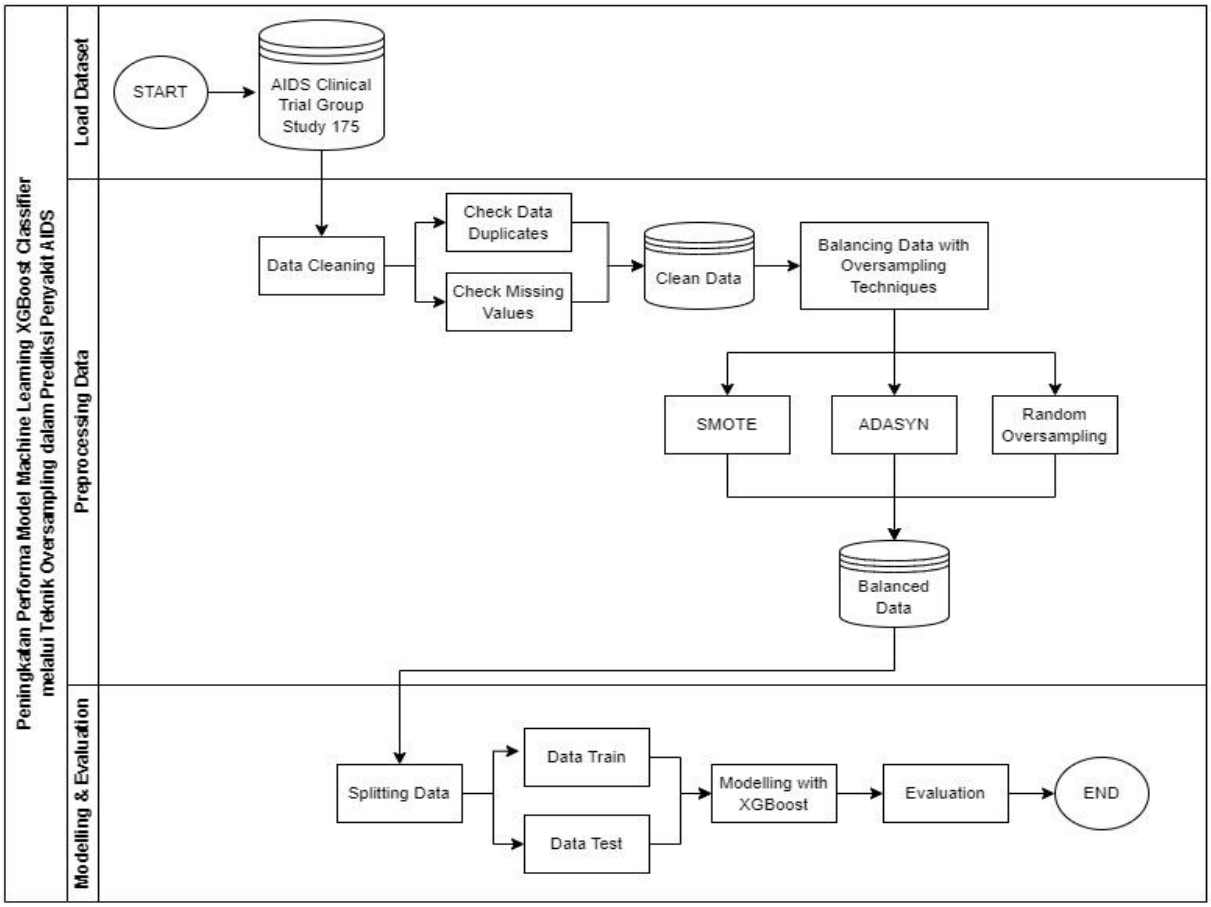
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Serangkaian proses dari penelitian mengenai evaluasi teknik oversampling dengan algoritma XGBoost Classifier pada dataset AIDS Clinical Trials Group Study 175 ditampilkan pada Gambar 1.

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501



Gambar 1. Tahapan Penelitian

Dimulai dari pengumpulan data penelitian pada tahap pertama, melalui public dataset Kaggle yaitu dataset AIDS Clinical Trials Group Study 175 dengan 24 fitur dan total data sebanyak 2139. Kedua, dilakukan data cleaning untuk memastikan bahwa dalam dataset tidak terdapat missing values dan duplikasi data. Setelah data dinyatakan bersih, pada tahap ketiga ini metode oversampling diimplementasikan untuk mengatasi data yang tidak seimbang dengan teknik SMOTE, ADASYN, dan Random Oversampling. Kemudian, memasuki tahap keempat yaitu modelling yang diawali dengan pembagian data train dan test sebelum diterapkan pada algoritma XGBoost Classifier, dalam hal ini data train digunakan sebagai bahan pelatihan model sedangkan data test digunakan bahan pengujian pengujian model. Terakhir, evaluasi yang digambarkan dalam bentuk classification report dan confusion matrix sebagai hasil akhir dari penelitian.

2.2 Data Source

AIDS Clinical Trials Group Study 175 sebagai dataset penelitian ini yang diperoleh dari Kaggle. Sebagai bahan utama dalam penelitian, dataset ini berisi data – data pasien yang menderita penyakit AIDS. Terdapat data sebanyak 2139 dan 24 fitur dengan kelas label sebagai fitur targetnya. Informasi lebih rinci mengenai dataset ini dapat dilihat pada Tabel 1.

Tabel 1. Informasi Dataset

Nama Fitur	Deskripsi
time	Untuk merepresentasikan waktu failure atau cencoring pada pasien.
trt	Kategori tipe perawatan pasien : 0 = ZDV, 1 = ZDV + ddl, 2 = ZDV + Zal, 3 = ddl.
age	Umur pasien saat awal perawatan.
wtkg	Berat badan pasien saat awal perawatan.
hemo	Status hemophilia pasien : 0 = no, 1 = yes.

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

Nama Fitur	Deskripsi
homo	Keterangan homoseksual pasien : 0 = no, 1 = yes.
Nama Fitur	Deskripsi
drugs	Riwayat konsumsi narkoba pasien : 0 = no, 1 = yes.
karnof	Status fungsional pasien dalam rentang 0-100
oprior	Terapi antiretroviral non-ZDV : 0 = no, 1 = yes.
z30	Penggunaan ZDV dalam 30 hari : 0 = no, 1 = yes.
zpior	Penggunaan ZDV : 0 = no, 1 = yes.
preanti	Merepresentasikan jumlah hari terapi antiretroviral dalam int.
race	Keterangan ras pasien : 0 = white, 1 = non-white.
gender	Keterangan jenis kelamin pasien : 0 = female, 1 = male.
str2	Riwayat antiretroviral : 0 = naïve, 1 = experienced.
strat	Stratifikasi riwayat antiretroviral dalam int.
symptom	Status gejala : 0 = asymptomatic, 1 = symptomatic.
treat	Perawatan tambahan : 0 = hanya ZDV, 1 = lainnya.
offtrt	Tidak melakukan perawatan sebelum 96+/-5 minggu : 0 = no, 1 = yes.
cd40	Jumlah CD4 pada awal perawatan
cd420	CD4 dalam 20+/-5 minggu
cd80	CD8 pada awal perawatan
cd820	CD8 dalam 20+/-5 minggu
label	Indikator biner (1 = failure, 0 = cencoring)

2.3 Data Cleaning

Untuk lebih memahami data dan memastikan data siap digunakan dalam penelitian, maka terlebih dahulu dilakukan Preprocessing. Preprocessing adalah proses persiapan data sebelum memasuki proses klasifikasi, tujuannya untuk menjaga kualitas dan konsistensi data penelitian. Hal ini sangat penting untuk diterapkan karena berpengaruh terhadap hasil klasifikasi pada tahap berikutnya [17] [18]. Melalui tahap ini dilakukan pembersihan data dengan menghapus duplikasi data dan memastikan tidak terdapat nilai yang hilang (missing values).

2.4 Imbalance Data

Situasi di mana kelas dalam dataset memiliki distribusi yang tidak seimbang disebut sebagai data yang tidak seimbang (Imbalance Data). Hal ini menjadi permasalahan dalam klasifikasi karena algoritma klasifikasi akan cenderung memprediksi kelas data mayoritas daripada kelas data minoritas. Akibatnya, prediksi terhadap kelas minoritas akan mendapatkan hasil yang lebih buruk dibanding dengan kelas mayoritas [19]. Oleh karena itu, perlu adanya penanganan dalam mengatasi ketidakseimbangan data dengan menerapkan metode oversampling.

2.5 Oversampling

Untuk mengatasi masalah ketidakseimbangan data dalam dataset, di mana jumlah data salah satu kelas (kelas mayoritas) lebih besar dibanding kelas lainnya (kelas minoritas), oversampling digunakan. Tujuan dari metode ini

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

adalah untuk meningkatkan representasi kelas minoritas dengan meningkatkan jumlah data dalam kelas minoritas sehingga distribusi kelas menjadi lebih seimbang. Ini dapat dicapai dengan membuat sampel sintetis baru yang sebanding dengan data kelas minoritas saat ini [20]. Dalam penelitian ini, terdapat ketidakseimbangan data pada fitur label di mana kelas 0 (cencored) menjadi kelas mayoritas dengan data sebanyak 1618, sedankgan kelas 1 (failure) menjadi kelas minoritas dengan 521 data. Untuk menyeimbangkan datanya, eksperimen dilakukan dengan menerapkan beberapa teknik oversampling SMOTE, ADASYN, dan Random Oversampling.

2.5.1 SMOTE

Skenario pertama adalah dengan menerapkan SMOTE yang diperkenalkan oleh Nither V. Chawla, SMOTE (Synthetic Minority Oversampling Technique) sebagai salah satu teknik oversampling untuk menyelesaikan masalah ketidakseimbangan data. Tujuannya untuk meningkatkan kinerja metode klasifikasi dengan melakukan modifikasi dataset yang tidak seimbang dan menciptakan data sintetik baru yang bersasal dari kelas minoritas [21] [22]. Data buatan untuk kelas minoritas dibangkitkan guna menyeimbangkan kelas dara mayoritas dan minoritas. K-tetangga terdekat digunakan dalam membangkitkan data buatan tersebut [9].

2.5.2 ADASYN

Teknik oversampling yang diterapkan selanjutnya yaitu ADASYN (Adaptive Syntethic), dengan mengambil bobot pembagian data dari kelas minoritas berdasarkan tingkat kesulitan pemahaman data oleh model machine learning. Dalam hal ini, data sintetis diperoleh dari kelas minoritas dengan tingkat kesulitan belajar yang lebih tinggi jika dibandingkan dengan data minoritas lain. [23]. Teknik ini memperbaiki pembagian data yang tidak seimbang dengan dua langkah. Diawali dengan mengurangi bias yang disebabkan oleh kelas yang tidak seimbang. Kemudian, ADASYN mampu mendorong batas keputusan klasifikasi ke arah sampel kelas yang menantang secara adaptif [9] [24].

2.5.3 Random Oversampling

Teknik oversampling terakhir yang diterapkan dalam eksperimen ini yaitu Random Oversampling. Dalam teknik ini, jumlah sampel dari kelas minoritas ditingkatkan secara acak dengan menambahkan salinan data dari kelas tersebut hingga jumlahnya sebanding dengan kelas mayoritas. Tujuannya untuk meningkatkan representasi kelas minoritas sehingga model machine learning dapat mempelajari pola-pola yang relevan dari kedua kelas secara seimbang, sehingga ketika diimplementasikan dapat meningkatkan performa model dalam mengklasifikasikan kelas minoritas [19] [25].

2.6 XGBoost Classifier

Setelah data diseimbangkan melalui teknik oversampling sebelumnya, eksperimen memasuki tahap modelling dengan algoritma XGBoost Classifier. Extreme Gradient Boosting atau yang biasa disebut XGBoost adalah salah satu algoritma klasifikasi machine learning yang dikembangkan oleh Chen dan Guestrin [26]. XGBoost bekerja berdasarkan prinsip ensemble learning, di mana ia menggabungkan prediksi dari beberapa model lemah (weak learners) yang disebut sebagai pohon keputusan (decision trees). XGBoost memperbaiki kelemahan algoritma gradient boosting tradisional dengan memperkenalkan beberapa inovasi, termasuk regularisasi, pohon keputusan berurutan, dan penanganan yang lebih baik terhadap data yang hilang (missing data). XGBoost juga mampu mengoptimalkan fungsi tujuan secara efisien dengan menggunakan pendekatan gradien stokastik. Hal ini memungkinkan XGBoost untuk menghasilkan model yang lebih cepat dan lebih baik dalam hal akurasi prediksi, bahkan untuk dataset yang sangat besar dan kompleks [27]. Metode ini memerlukan fungsi objektif yang bermanfaat dalam menilai kualitas model yang dihasilkan selaras dengan data training [28]. Dalam penelitian ini, algoritma XGBoost diimplementasikan pada data yang telah diseimbangkan oleh ketiga teknik oversampling yakni SMOTE, ADASYN, dan Random Oversampling. Berikut ini algoritma dari XGBoost Classifier:

1. Perhitungan nilai pelatihan yang hilang dan nilai regularisasi.

$$obj = L(\theta) + \Omega(\theta) \quad (1)$$

Keterangan:
θ diartikan sebagai parameter model terkait, Ω sebagai fungsi regularisasi, L fungsi pelatihan yang hilang.

2. Persamaan fungsi pelatihan yang hilang.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \quad (2)$$

Keterangan:
y_i diartikan sebagai nilai yang dianggap benar pada data sebenarnya dan $\hat{y}_i$ yaitu hasil prediksi dari model, sedangkan n sebagai jumlah iterasi nilai dari model.

2.7 Evaluation

Tahap terakhir yaitu evaluasi, tujuannya untuk mengevaluasi kinerja model klasifikasi yang telah dikembangkan. Evaluasi dilakukan untuk menilai dan memilih metode terbaik berdasarkan eksperimen yang telah dilakukan.

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

Dengan memanfaatkan confusion matrix sebagai metode yang paling umum diterapkan guna mengukur kinerja klasifikasi. Metode tersebut memungkinkan peneliti untuk melihat kualitas model dalam melakukan klasifikasi objek pengamatan ke dalam kelas-kelas yang berbeda. Evaluasi juga melibatkan perhitungan berbagai metrik kinerja, termasuk akurasi, presisi, recall, dan F1-score, untuk memberikan pemahaman yang lebih komprehensif tentang kualitas prediksi model [29]. Ilustrasi confusion matrix dapat dilihat pada Tabel 2.

Tabel 2. Confusion Matrix

Confusion Matrix		**Predicted Class**	
		P	**N**
Actual Class	**P**	TP (True Positive)	FN (False Negative)
	N	FP (False Negative)	TN (True Negative)

Metrik akurasi (accuracy) merepresentasikan banyaknya kelas yang terprediksi benar oleh model. Persamaan dari metrik akurasi [29], sebagai berikut :

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Metrik presisi (precision) menunjukkan rasio kelas yang terprediksi benar terhadap jumlah total kelas positif (aktual). Persamaan dari metrik presisi, sebagai berikut :

$$precision = \frac{TP}{TP+FP} \quad (4)$$

Kemampuan model dalam menemukan kembali semua instance positif yang benar (true positive atau false negative) dari semua instance positif yang sebenarnya dikenal sebagai recall. Dengan kata lain, recall dapat menggambarkan seberapa baik kualitas dari model dalam memprediksi kelas positif (aktual). Persamaan dari recall [29], sebagai berikut :

$$recall = \frac{TP}{TP+FN} \quad (5)$$

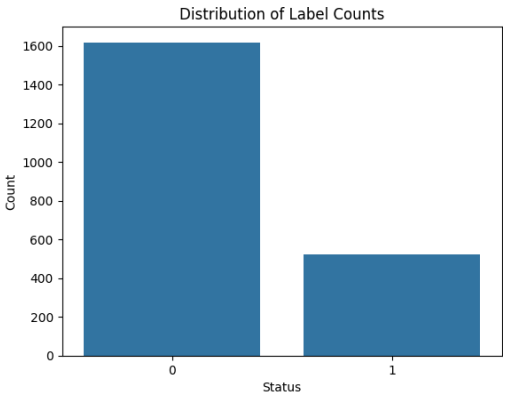
Dalam menilai kinerja kelas positif, skor F1 melakukan hitungan pada rata-rata harmonic keduanya untuk menutupi kekurangan dalam presisi dan recall. [29]. Persamaan dari F1 Score, sebagai berikut :

$$f1-score = 2 * \frac{precision*recall}{precision+recall} = 2 * \frac{2TP}{2TP+FP+FN} \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Pembahasan pertama yaitu mengenai dataset AIDS Clinical Trials Group Study 175 sebagai bahan utama dalam penelitian ini. Kumpulan data tersebut terdiri dari hasil studi klinis yang dilaksanakan pada pasien-pasien yang terinfeksi HIV. Dengan tujuan utama untuk menyelidiki dan mengevaluasi efektivitas berbagai terapi antiretroviral dalam mengendalikan perkembangan HIV pada pasien. Dataset ini mencakup beragam informasi klinis dan laboratorium yang dikumpulkan selama studi, termasuk informasi mengenai respons terhadap berbagai jenis terapi yang diberikan kepada pasien. Terdapat 24 fitur dalam dataset termasuk fitur target yaitu label. Namun, terdapat permasalahan dalam distribusi kelas yang tidak seimbang pada fitur target tersebut, seperti yang ditampilkan pada gambar 2 dan tabel 3 berikut.



Gambar 2. Distribusi Kelas label

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

Tabel 3. Jumlah Pembagian Data

Class	Jumlah
0	1618
1	521

Melalui analisis grafik pada Gambar 2 menunjukkan perbedaan jumlah data yang cukup signifikan antara kelas 0 dan kelas 1 dalam dataset. Terdapat 1618 data pada kelas 0, sementara hanya 521 data yang terdapat pada kelas 1. Kondisi ketidakseimbangan ini dapat berdampak negatif terhadap hasil klasifikasi, karena algoritma cenderung memihak pada kelas mayoritas dalam proses pembelajaran. [19]. Oleh karena itu, untuk mengatasi permasalahan tersebut diperlukan penanganan khusus. Metode oversampling yang akan dikombinasikan dengan XGBoost Classifier, menjadi fokus utama penelitian ini. Langkah-langkah ini diharapkan dapat memperbaiki distribusi kelas dalam dataset dan meningkatkan kualitas prediksi model dalam mengenali kedua kelas dengan lebih seimbang.

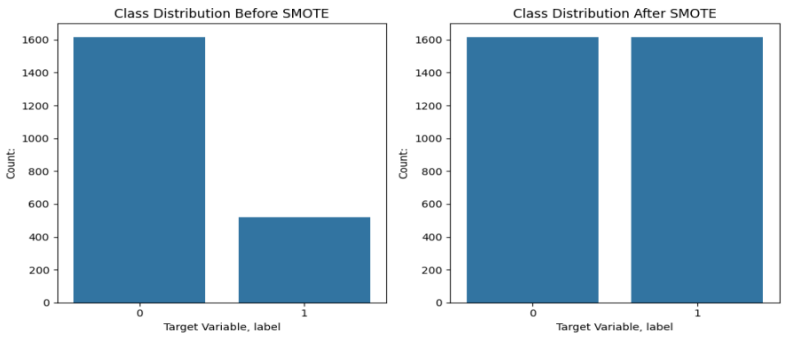
3.2 Data Cleaning

Sebelum beralih ke tahap berikutnya dalam penelitian, memastikan kualitas dan konsistensi data menjadi langkah penting yang tidak bisa diabaikan. Oleh karena itu, proses data cleaning dilakukan dengan cermat untuk memeriksa keberadaan nilai yang hilang (missing values) dan duplikasi data. Langkah pertama adalah mengidentifikasi nilai yang hilang dalam dataset, dilanjutkan dengan pengecekan kemungkinan adanya duplikasi data. Hasil data cleaning menegaskan bahwa dataset tidak mengandung nilai yang hilang maupun data yang terduplikat, dipastikan bahwa kualitas dataset terjaga dengan baik dan siap untuk analisis lebih lanjut. Dengan demikian, diharapkan hasil analisis yang dihasilkan akan lebih dapat diandalkan dan bebas dari bias yang dapat memengaruhi interpretasi hasil penelitian.

3.3 Oversampling

Dalam upaya mengatasi ketidakseimbangan data, teknik oversampling menjadi strategi yang efektif untuk menciptakan keseimbangan antara kelas mayoritas dan minoritas dalam dataset. Ini penting karena data yang tidak seimbang dapat menyebabkan bias dalam model pembelajaran mesin [20]. Terdapat beberapa jenis oversampling yang umum digunakan, seperti Random Oversampling, SMOTE, dan ADASYN, yang masing-masing memiliki pendekatan yang berbeda dalam menciptakan sampel baru untuk kelas minoritas. Dengan menggunakan oversampling, distribusi kelas dalam dataset diperbaiki, memungkinkan model untuk mempelajari pola dari kedua kelas dengan lebih baik dan menghasilkan prediksi yang lebih akurat.

Eksperimen pertama adalah penerapan teknik SMOTE. SMOTE mengambil peran penting dengan memodifikasi dataset dan menciptakan data sintetis baru dari kelas minoritas. Langkah pertama adalah menemukan titik-titik data minoritas yang strategis. Kemudian, dilakukan penciptaan data sintetis baru dengan cara menghubungkan titik-titik tersebut secara linier atau dengan menggunakan algoritme k-nearest neighbors atau tetangga terdekat [9]. Data sintetis yang dihasilkan akan digunakan untuk menyeimbangkan distribusi kelas antara mayoritas dan minoritas. Pendekatan ini membantu meningkatkan kualitas model pembelajaran mesin dengan memberikan akses yang lebih seimbang terhadap data minoritas, sehingga meningkatkan kemampuan model dalam memahami pola-pola yang mungkin terabaikan sebelumnya. Hasil dari teknik SMOTE dapat dilihat pada Gambar 3 berikut ini.



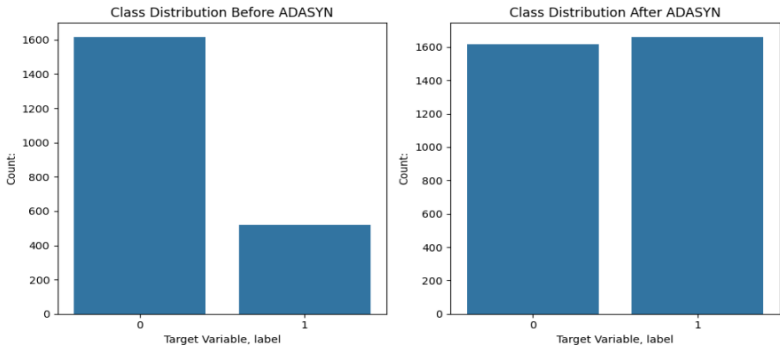
Gambar 3. Hasil Implementasi SMOTE

Berdasarkan analisis pada Gambar 3, terlihat bahwa SMOTE berhasil menyeimbangkan data dengan baik. Perbedaan jumlah data yang signifikan antara kelas mayoritas dan kelas minoritas telah terseimbangkan dengan adil. Keberhasilan ini didasari oleh pendekatan pembuatan sampel sintetis baru pada kelas minoritas berdasarkan tetangga terdekat. SMOTE memperkaya dataset dengan menciptakan observasi sintetis yang menyerupai data minoritas, namun dengan variasi minor. Pendekatan ini memberikan model pembelajaran mesin akses yang lebih luas terhadap pola dalam data minoritas, secara signifikan meningkatkan kinerja algoritma dalam melakukan prediksi dengan lebih akurat.

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

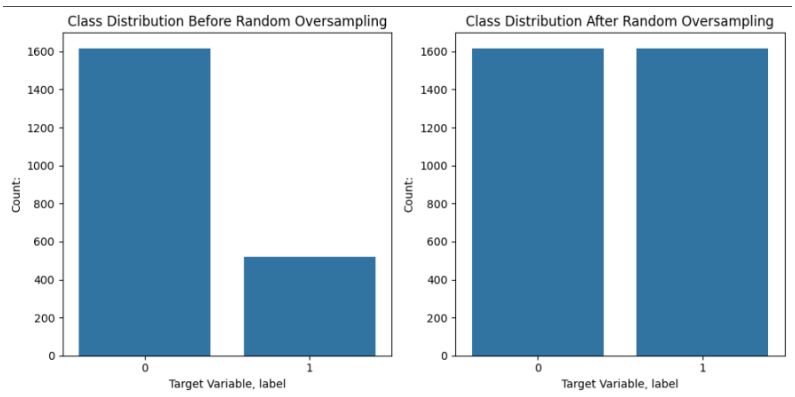
ADASYN, sebagai teknik oversampling kedua yang digunakan dalam penelitian ini, memperkaya pemahaman kita terhadap penyeimbangan data. Berbeda dengan SMOTE, ADASYN mengambil bobot pembagian data kelas minoritas berdasarkan tingkat kesulitan pemahaman oleh model machine learning. Dalam penerapannya, ADASYN memecahkan ketidakseimbangan kelas dengan dua langkah strategis. Pada permulaan, teknik ini mengurangi bias yang terjadi akibat dari ketidakseimbangan kelas, memungkinkan model untuk memahami kelas minoritas dengan lebih baik. Kemuan, ADASYN mendorong batas keputusan klasifikasi ke arah sampel kelas yang lebih menantang secara adaptif, memperbaiki kemampuan model dalam memahami pola-pola yang kompleks [9]. Dengan demikian, ADASYN memberikan kontribusi yang signifikan dalam meningkatkan performa model machine learning dalam menghadapi data tidak seimbang, memperkuat kemampuan prediksi dan analisis model secara keseluruhan. Hasil dari ADASYN dapat dilihat pada Gambar 4 berikut ini.



Gambar 4. Hasil Implementasi ADASYN

Seperti halnya SMOTE pada pembahasan sebelumnya, ADASYN juga berhasil dalam mengatasi data yang tidak seimbang. Hanya terdapat sedikit perbedaan dalam distribusi kelasnya setelah penerapan ADASYN, namun perbedaan tersebut bukanlah hasil yang buruk mengingat sangat sedikit selisihnya dan masih dapat dikatakan seimbang untuk dilakukan analisis. Dalam hal ini, kelas 1 sedikit lebih banyak dibanding kelas 0. Hasil penerapan ADASYN ini didasari oleh proses penyesuaian dinamis yang mempertimbangkan tingkat kesulitan dalam membedakan antara kelas mayoritas dan minoritas. Kemudian, proses penyesuaian dinamis yang mempertimbangkan tingkat kesulitan dalam membedakan antara kelas mayoritas dan minoritas.

Dalam penelitain untuk meningkatkan performa model XGBoost melalui teknik oversampling ini, teknik Random Oversamping juga digunakan sebagai metode oversampling terakhir. Random Oversampling dalam penerapannya memiliki langkah yang dimulai dengan memasukkan jumlah data training yang tersedia. Selanjutnya, dilakukan perhitungan untuk mengidentifikasi jumlah kelas mayoritas dan minoritas, serta menentukan selisih antara keduanya. kemudian, melakukan insiasi i=1 sebagai indeks perulangan. Dalam setiap iterasi, sistem memeriksa apakah indeks perulangan kurang dari sama dengan 1. Jika iya, maka dilakukan duplikasi acak dari data kelas minoritas untuk menyamakan distribusi kelas. Namun, jika indeks tidak memenuhi kondisi tersebut, data akan disimpan untuk proses pembangkitan lebih lanjut. Tahapan ini bertujuan untuk menciptakan keseimbangan dalam dataset dengan memperbanyak sampel dari kelas minoritas secara acak, sehingga dapat meningkatkan kualitas prediksi model dalam mengenali kelas minoritas [19]. Hasil penerapan Random Oversampling dapat dilihat pada Gambar 5 berikut ini.



Gambar 5. Hasil Implementasi Random Oversampling

Random Oversampling, sebagai teknik terakhir dalam penanganan ketidakseimbangan data, berhasil membawa perbaikan yang signifikan. Dengan melihat grafik pada Gambar 5, terlihat bahwa kedua kelas berada pada tingkat yang seimbang, menunjukkan kesuksesan dari pendekatan ini. Proses duplikasi acak dari sampel data kelas minoritas memberikan kontribusi yang penting dalam meningkatkan jumlah data pada kelas minoritas

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

sehingga sejajar dengan kelas mayoritas. Dengan demikian, model machine learning dapat mengakses informasi dari kedua kelas dengan proporsional, meminimalkan risiko bias dalam pembelajaran. Kesuksesan Random Oversampling dalam memperbaiki distribusi kelas ini memberikan keyakinan untuk melanjutkan penelitian ke tahap selanjutnya, yaitu modelling, dengan harapan menghasilkan model yang lebih kuat dan akurat dalam memprediksi kasus AIDS.

Ketiga teknik oversampling tersebut secara keseluruhan memberikan hasil yang memuaskan dalam menangani ketidakseimbangan data. Oleh karena itu, akan dilanjutkan penelitian lebih mendalam ke tahap modelling dengan algoritma XGBoost Classifier. Tahap ini penting guna memperdalam analisis dan evaluasi performa model. Melalui proses modelling yang teliti, diharapkan hasil terbaik dapat terjawab. Evaluasi pada tahap selanjutnya akan menghasilkan informasi yang lebih rinci mengenai efektivitas teknik oversampling dalam meningkatkan performa model XGBoost Classifier, serta memastikan bahwa model yang dihasilkan dapat memberikan prediksi yang akurat dan dapat diandalkan.

3.4 XGBoost Classifier Modelling

Serangkaian tahapan telah dilakukan guna mempersiapkan data dan permasalahan data yang tidak seimbang pun telah teratasi melalui teknik oversampling. Kini pembahasan mulai memasuki tahap modelling yang menjadi proses utama pada penelitian ini. XGBoost Classifier akan digunakan sebagai algoritma machine learning dalam proses modelling. Algoritma ini bekerja berdasarkan prinsip ensemble learning, di mana ia menggabungkan prediksi dari beberapa model lemah (weak learners) yang disebut sebagai pohon keputusan (decision trees). XGBoost memperbaiki kelemahan algoritma gradient boosting tradisional dengan memperkenalkan beberapa inovasi, termasuk regularisasi, pohon keputusan berurutan, dan penanganan yang lebih baik terhadap data yang hilang (missing data). XGBoost mampu untuk menghasilkan model klasifikasi dengan kualitas lebih baik dalam hal akurasi prediksi, bahkan untuk dataset yang sangat besar dan kompleks. Hal ini dikarenakan adanya optimalisasi fungsi tujuan secara efisien dengan menggunakan pendekatan gradien stokastik [27].

Proses modelling dimulai dengan pemisahan data menjadi dua subset: data pelatihan (80%) dan data pengujian (20%). Subset data pelatihan sebagai bahan untuk melatih model algoritma XGBoost dan memahami pola-pola data yang dianalisis dalam penelitian ini. Pembagian ini penting untuk menegaskan bahwa model yang dihasilkan mampu menggeneralisasi dengan baik terhadap data baru, serta meningkatkan akurasi prediksi yang akan dilakukan pada tahap pengujian.

3.5 Evaluation

Tahap modelling telah selesai, fokus saat ini adalah pada evaluasi model XGBoost menggunakan data pengujian. Berbagai metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score, diterapkan guna mengukur kinerja model. Selain itu, confusion matrix akan dibentuk untuk memberikan ilustrasi visual tentang hasil prediksi model. Confusion matrix menjadi alat yang sangat berguna dalam memberikan informasi yang lebih rinci dan mudah dipahami, dengan memperlihatkan jumlah prediksi yang benar dan salah untuk setiap kelas target. Tahap evaluasi ini menjadi penting untuk memastikan bahwa hasil modelling yang dikembangkan dapat melakukan prediksi yang akurat dan dapat diandalkan ketika diterapkan pada data baru.

Terkait dengan permasalahan data tidak seimbang yang terjadi dalam penelitian ini, sebelumnya telah dilakukan eksperimen teknik oversampling guna menyelesaikan permasalahan tersebut. Beberapa teknik oversampling yang diterapkan adalah SMOTE, ADASYN, dan Random Oversampling, di mana masing – masing teknik berhasil berperan dalam menyeimbangkan data. Kemudian hasil dari setiap teknik oversampling tersebut diimplementasikan ke dalam algoritma XGBoost Classifier untuk dilakukan modelling dan sebagai hasil akhir melalui metrik evaluasi dapat dilihat pada Tabel 4 berikut.

Tabel 4. Metrik Evaluasi

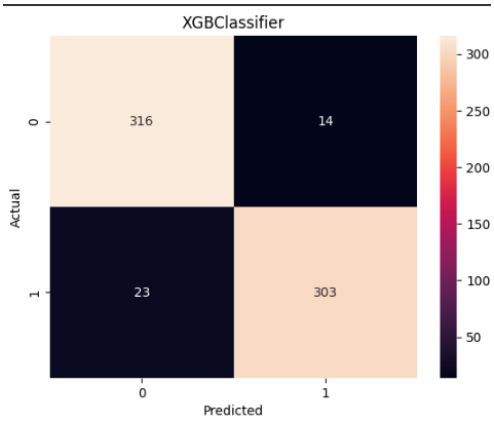
Teknik	Akurasi	Presisi	Recall	F1_Score
Normal	88,79%	81,18%	68,32%	74,19%
SMOTE	92,59%	91,13%	94,01%	92,55%
ADASYN	94,36%	92,94%	95,58%	94,25%
Random Oversampling	94,44%	90,72%	98,74%	94,56%

Metrik evaluasi menunjukkan bahwa penggunaan teknik oversampling secara signifikan meningkatkan performa model XGBoost Classifier dalam melakukan prediksi. Dibandingkan dengan penggunaan model tanpa oversampling (Normal), teknik oversampling, dalam hal ini SMOTE, ADASYN, dan Random Oversampling, menghasilkan peningkatan yang mencolok dalam semua metrik evaluasi. Hasil ini selaras dengan penelitian sebelumnya [14] yang juga menerapkan teknik oversampling sehingga mampu meningkatkan performa model. Terlihat bahwa Akurasi, Presisi, Recall, dan F1_Score meningkat secara konsisten seiring dengan penerapan teknik oversampling. Misalnya, menggunakan Random Oversampling, Akurasi meningkat menjadi 94,44%, sementara Recall mencapai 98,74%, menunjukkan peningkatan yang nyata dalam kemampuan model untuk mengidentifikasi dengan benar. Hasil ini menggambarkan peran penting teknik oversampling dalam mengatasi ketidakseimbangan data dan meningkatkan kualitas prediksi model dalam penelitian ini. Kemudian hasil confusion matrix dari masing

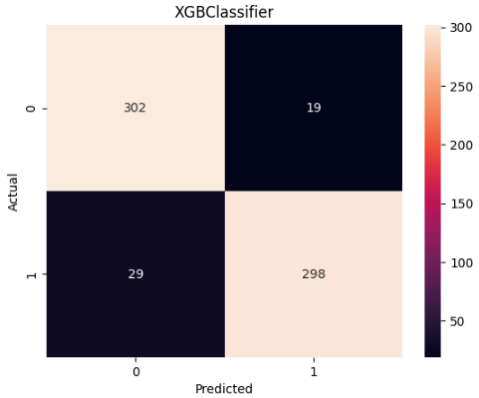
Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

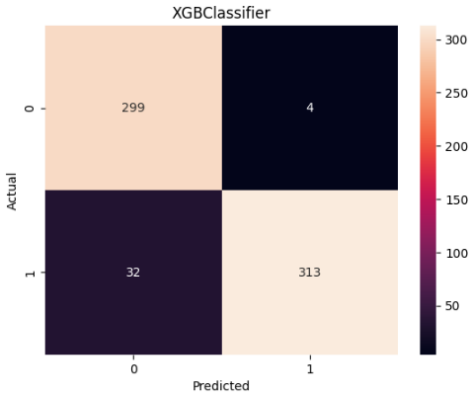
– masing teknik oversampling dengan XGBoost dapat dilihat pada Gambar 6, Gambar 7, dan Gambar 8 sebagai berikut.



Gambar 6. Confusion Matrix SMOTE



Gambar 7. Confusion Matrix ADASYN



Gambar 8. Confusion Matrix Random Oversampling

Berdasarkan evaluasi yang telah dilakukan, menunjukkan bahwa Random Oversampling merupakan teknik oversampling terbaik dalam meningkatkan akurasi model XGBoost, walaupun hanya unggul tipis 0,08% dari ADASYN. Dari confusion matrix, terlihat bahwa Random Oversampling berhasil memprediksi dengan tepat 299 kasus kelas 0 dan 313 kasus kelas 1. Dari hasil eksperimen ini, dapat disimpulkan bahwa Random Oversampling mampu secara signifikan meningkatkan performa model XGBoost Classifier. Kombinasi antara model XGBoost dan teknik Random Oversampling terbukti efektif dalam meningkatkan performa model prediktif pada studi kasus dataset AIDS Clinical Trials Group Study 175. Hal ini menunjukkan relevansi dan potensi penggunaan Random Oversampling dalam aplikasi praktis, terutama dalam bidang kesehatan.

4. KESIMPULAN

Melalui penelitian ini, telah dilakukan upaya untuk meningkatkan performa model XGBoost Classifier dengan studi kasus pada dataset AIDS Clinical Trials Group Study 175 yang memiliki ketidakseimbangan data pada fitur

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

target. Dengan menerapan tiga teknik oversampling, yakni SMOTE, ADASYN, Random Oversampling. Ketiganya berhasil menciptakan keseimbangan antara kelas minoritas dan mayoritas, yang menghasilkan peningkatan performa model. Berdasarkan hasil penelitian, implementasi model XGBoost tanpa menggunakan teknik oversampling memperoleh hasil akurasi 88,79%, presisi 81,18%, recall 68,32%, dan f1_score 74,19%. Kemudian dilakukan eksperimen dengan menerapkan teknik oversampling yang dikombinasikan model XGBoost. Hasilnya teknik Random Oversampling menjadi teknik terbaik yang mampu meningkatkan performa model XGBoost dengan perolehan akurasi 94,44%, presisi 90,72%, recall 98,74%, dan f1_score 94,65%. Untuk lebih detailnya pada confusion matrix terdapat sebanyak 299 kelas 0 terprediksi benar dan sebanyak 313 kelas 1 terprediksi benar oleh kombinasi Random Oversampling dan XGBoost Classifier. Dengan demikian, dapat disimpulkan bahwa kombinasi model XGBoost dengan Random Oversampling menjadi pendekatan yang efektif dalam meningkatkan performa model prediktif pada studi kasus dataset AIDS Clinical Trials Group Study 175. Penemuan ini memiliki kontribusi yang signifikan dalam pengembangan metodologi analisis data kesehatan. Hal ini mendukung pengambilan keputusan yang lebih akurat dalam bidang kesehatan masyarakat, khususnya dalam penelitian penyakit AIDS. Tentunya, hasil penelitian ini akan memberikan dampak positif yang besar bagi upaya pencegahan dan pengendalian penyakit tersebut.

UCAPAN TERIMA KASIH

Dengan ketulusan hati, saya ucapakan terima kasih dan rasa syukur kepada Tuhan Yang Maha Esa atas limpahan nikmat, petunjuk, dan kesehatan yang telah diberikan sehingga saya mampu menyelesaikan penelitian ini. Kepada keluarga, khususnya orang tua saya yang telah senantiasa mendoakan saya dalam segala kondisi, tanpa mereka saya tidak akan menjadi pribadi yang kuat seperti sekarang. Tak lupa, kepada dosen pembimbing yang telah memberikan arahan, masukan, dan bimbingan berharga kepada saya demi kelancaran penelitian ini. Kemudian, kepada teman, sahabat, atau orang terdekat saya, terima kasih atas dukungan dan ilmu yang bermanfaat sehingga saya dapat berkembang. Terakhir, kepada pihak – pihak terkait yang telah memberikan akses dalam pengumpulan data. Semua dukungan ini sangat berarti bagi saya dalam proses penyelesaian penelitian saya, semoga apa yang telah saya lakukan dan tuangkan melalui penelitian ini dapat berkontribusi positif bagi perkembangan ilmu pengetahuan.

REFERENCES

[1] World Health Organization, “HIV and AIDS.” Accessed: Feb. 08, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hiv-aids
[2] R. S. Gumarianto, S. Lardo, and A. Chairani, “Hubungan antara Hitung Jumlah CD4 dengan Kejadian Wasting Syndrome pada Pasien HIV/AIDS Di RSPAD Gatot Soebroto Periode Januari-Desember 2020,” J. Kedokt. dan Kesehat. Publ. Ilm. Fak. Kedokt. Univ. Sriwij., vol. 9, no. 2, pp. 133–142, 2022.
[3] D. A. Putri, R. J. Sitorus, and N. Najmah, “Perilaku Berisiko Penularan HIV-AIDS pada Lelaki Seks Lelaki: Studi Literatur,” Heal. Inf. J. Penelit., pp. e1112–e1112, 2023.
[4] G. Ayala and A. Spieldenner, “HIV is a story first written on the bodies of gay and bisexual men,” American Journal of Public Health, vol. 111, no. 7. American Public Health Association, pp. 1240–1242, 2021.
[5] A. Arabia et al., “Evaluasi Sistem Surveillance HIV/AIDS Di Kota Bogor,” Media Kesehat. Politek. Makassar, vol. 18, no. 2, pp. 277–290, 2023.
[6] L. Zhang et al., “A review of machine learning in building load prediction,” Appl. Energy, vol. 285, p. 116452, 2021.
[7] A. Roihan, P. A. Sunarya, and A. S. Rafika, “Pemanfaatan Machine Learning dalam Berbagai Bidang,” J. Khatulistiwa Inform., vol. 5, no. 1, p. 490845, 2020.
[8] R. G. Wardhana, G. Wang, and F. Sibuea, “PENERAPAN MACHINE LEARNING DALAM PREDIKSI TINGKAT KASUS PENYAKIT DI INDONESIA,” J. Inf. Syst. Manag., vol. 5, no. 1, pp. 40–45, 2023.
[9] M. Aqsha and N. Sunusi, “PERFORMA KLASIFIKASI DATA TIDAK SEIMBANG DENGAN PENDEKATAN MACHINE LEARNING (STUDI KASUS: DIABETES INDIAN PIMA),” J. Mat. UNAND, vol. 12, no. 2, pp. 176–193, 2024.
[10] P. R. Sihombing and I. F. Yuliati, “Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia,” MATRIK J. Manajemen, Tek. Inform. Dan Rekayasa Komput., vol. 20, no. 2, pp. 417–426, 2021.
[11] N. S. Rahmi, N. W. S. Wardhani, M. B. Mitakda, R. S. Fauztina, and I. Salsabila, “SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data : (Case Study of Early Detection of Cervical Cancer in Indonesia),” in 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), 2022, pp. 1–6. doi: 10.1109/ICITDA55840.2022.9971421.
[12] C. Haryawan and Y. M. K. Ardhana, “ANALISA PERBANDINGAN TEKNIK OVERSAMPLING SMOTE PADA IMBALANCED DATA,” J. Inform. dan Rekayasa Elektron., vol. 6, no. 1, pp. 73–78, 2023.
[13] J. He et al., “Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation,” Front. Public Heal., vol. 10, p. 967681, 2022.
[14] R. Zhou et al., “Prediction Model for Infectious Disease Health Literacy Based on Synthetic Minority Oversampling Technique Algorithm,” Comput. Math. Methods Med., vol. 2022, p. 8498159, 2022, doi: 10.1155/2022/8498159.
[15] R. M. Munshi, “Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction,” PLoS One, vol. 19, no. 1, pp. e0296107-, Jan. 2024, [Online]. Available:

Volume 8, Nomor 2, April 2024, Page 736-747
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i2.7501

https://doi.org/10.1371/journal.pone.0296107
[16] S. U. Nisa, A. Mahmood, F. S. Ujager, and M. Malik, "HIV/AIDS predictive model using random forest based on socio-demographical, biological and behavioral data," Egypt. Informatics J., vol. 24, no. 1, pp. 107–115, 2023, doi: https://doi.org/10.1016/j.eij.2022.12.005.
[17] L. B. Adzy, A. Pambudi, U. M. Sukabumi, P. Bantuan, I. Jaminan, and S. K. Sukabumi, "Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima," vol. 6, no. 1, pp. 1–10, 2023.
[18] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," J. Media Inform. Budidarma, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.
[19] R. D. Fitriani, H. Yasin, and T. Tarno, "PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," J. Gaussian, vol. 10, no. 1, pp. 11–20, 2021, doi: 10.14710/j.gauss.v10i1.30243.
[20] P. Y. Saputra, M. Z. Abdullah, and A. P. Kirana, "Improvisasi Teknik Oversampling MWMOTE Untuk Penanganan Data Tidak Seimbang," J. Media Inform. Budidarma, vol. 5, no. 2, p. 398, 2021, doi: 10.30865/mib.v5i2.2811.
[21] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE," Aiti, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.
[22] N. N. Sholihah and A. Hermawan, "Implementation of Random Forest and Smote Methods for Economic Status Classification in Cirebon City," J. Tek. Inform., vol. 4, no. 6, pp. 1387–1397, 2023.
[23] F. S. Dhitama and F. A. Bachtiar, "Penentuan Kelayakan Debitur Menggunakan Metode Decision Tree C4.5 dan Oversampling Adaptive Synthetic (ADASYN)," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 10, pp. 3712–3721, 2020.
[24] G. Ahmed et al., "DAD-Net: Classification of Alzheimer's Disease Using ADASYN Oversampling Technique and Optimized Neural Network," Molecules, vol. 27, no. 20, pp. 1–21, 2022, doi: 10.3390/molecules27207085.
[25] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," Int. J. Artif. Intell. Res., vol. 4, no. 2, p. 86, 2021, doi: 10.29099/ijair.v4i2.152.
[26] Z. E. Aydin and Z. K. Ozturk, "Performance Analysis of XGBoost Classifier with Missing Data," 1st Int. Conf. Comput. Mach. Intell., no. March, 2021, [Online]. Available: https://www.researchgate.net/publication/350135431
[27] Nayan Kumar Sinha, "Developing A Web based System for Breast Cancer Prediction using XGboost Classifier," Int. J. Eng. Res., vol. V9, no. 06, pp. 852–856, 2020, doi: 10.17577/ijertv9is060612.
[28] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," J. Math. Theory Appl., vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.
[29] H. Nuraliza, O. N. Pratiwi, and F. Hamami, "Analisis Sentimen IMBd Film Review Dataset Menggunakan Support Vector Machine (SVM) dan Seleksi Feature Importance," J. Mirai Manaj., vol. 7, no. 1, pp. 1–17, 2022.

Submitted: **19/02/2024**; Accepted: **22/04/2024**; Published: **23/04/2024**