

Impact of SMOTE and ADASYN on Class Imbalance in Metabolic Syndrome Classification Using Random Forest Algorithm

Lutfiana Deka Nurhayati ^{1*}, Majid Rahardi ^{2*}

* Informatika, Universitas Amikom Yogyakarta, Indonesia

lutfianadeka@students.amikom.ac.id¹, majid@amikom.ac.id²

Article Info

Article history:

Received 2025-08-05

Revised 2025-08-29

Accepted 2025-09-10

Keyword:

ADASYN,
Metabolic Syndrome,
SMOTE,
Random Forest.

ABSTRACT

Metabolic Syndrome is a collection of medical conditions that can increase the risk of stroke, cardiovascular disease, and type 2 diabetes. Early detection of this condition requires a machine learning model capable of accurate classification to support timely treatment. However, class imbalance in data often hampers the performance of classification algorithms, particularly in recognizing minority classes, namely individuals diagnosed with Metabolic Syndrome. This study aims to analyze the effect of applying the SMOTE and ADASYN data balancing techniques in classifying Metabolic Syndrome using the Random Forest algorithm. These algorithms were chosen for their ability to produce accurate predictions, although their performance can decline when faced with imbalanced class distributions. The results showed that the model without data balancing techniques achieved 86% accuracy with a minority class recall of 75%. The application of SMOTE increased accuracy to 91% and recall to 93%, while ADASYN achieved 92% accuracy and a minority class recall of 95%. These findings indicate that the ADASYN technique combined with the Random Forest algorithm provides significant performance improvements in the classification of Metabolic Syndrome on imbalanced data.



This is an open access article under the [CC-BY-SA](#) license.

I. PENDAHULUAN

Sindrom Metabolik merupakan suatu kelainan medis yang meliputi kondisi obesitas sentral, resistensi insulin, hipertensi dan dislipidemia. Seseorang yang memiliki kelainan tersebut memiliki risiko lebih besar terkena penyakit stroke kardiovaskular, dan diabetes tipe 2 [1].

Menurut National Heart, Lung and Blood Institute, seseorang didiagnosa terkena Sindrom Metabolik apabila memiliki 3 dari 5 kriteria berikut: lingkar pinggang (>102 cm bagi pria dan >89 cm bagi wanita), tekanan darah ($\geq 130/85$ mmHg), kadar gula darah puasa (≥ 100 mg/dL), kadar trigliserida (≥ 150 mg/dL) dan kadar HDL (<40 mg/dL bagi pria dan <50 mg/dL bagi wanita) [2].

Berdasarkan laporan global, kasus Sindrom Metabolik diperkirakan mencapai sekitar 25% dari populasi dunia, kondisi ini menunjukkan peningkatan pada anak-anak seiring perubahan gaya hidup. Di Amerika Serikat, prevalensinya meningkat dari 25,3% pada tahun 1988-1994 menjadi sebesar 34,2% pada tahun 2007-2012 [3]. Sementara itu, di Indonesia

prevalensi Sindrom Metabolik dilaporkan sebesar 17,5%, dengan potensi mampu meningkatkan risiko stroke dan penyakit jantung yang mampu menyebabkan kematian [4]. Oleh karena itu, diperlukan sebuah model machine learning seperti klasifikasi untuk mendeteksi Sindrom Metabolik lebih dini agar dapat dilakukan penanganan lebih lanjut.

Dalam membangun model klasifikasi, algoritma Random Forest sering digunakan karena kemampuannya dalam menangani data yang kompleks terhadap outlier dan mampu menghasilkan prediksi yang akurat. Namun, keberadaan kelas yang tidak seimbang (class imbalance) dalam data mampu mempengaruhi hasil prediksi, karena model yang cenderung mempelajari kelas mayoritas dan mengabaikan kelas minoritas [5]. Sehingga, diperlukan pendekatan yang dapat digunakan untuk menangani masalah ketidakseimbangan kelas, seperti menggunakan teknik SMOTE (Synthetic Minority Over-sampling) dan ADASYN (Adaptive Synthetic Sampling). Kedua metode ini menambahkan sampel sintetis pada kelas minoritas melalui pendekatan yang berbeda [6].

Penelitian terdahulu oleh Sholekhah et al. menyebutkan bahwa algoritma KNN mampu menghasilkan akurasi sebesar 82% dan dinilai lebih efektif dibandingkan Naïve Bayes yang hanya mampu menghasilkan akurasi sebesar 79% dalam mengklasifikasikan Sindrom Metabolik dengan skema pembagian data 50:50 dalam mengklasifikasikan Sindrom Metabolik. Sehingga masih diperlukan eksplorasi algoritma lain guna meningkatkan akurasi [7].

Beberapa penelitian sebelumnya telah menunjukkan efektivitas penggunaan beberapa teknik balancing data. Salah satu penelitian oleh Fadmadika et al. menyebutkan bahwa penerapan teknik SMOTE mampu meningkatkan akurasi pada algoritma Random Forest, meskipun pada algoritma Gradient Boosting akurasi mengalami sedikit penurunan [5]. Selain itu, penelitian lain juga menyebutkan bahwa kombinasi algoritma Random Forest dengan teknik balancing SMOTE lebih unggul dibandingkan dengan kombinasi CatBoost dengan balancing SMOTE dalam memprediksi risiko diabetes[8]. Dalam konteks prediksi stroke, berdasarkan evaluasi F1-score penggunaan algoritma Random Forest dengan SMOTE lebih unggul dibandingkan dengan kombinasi penggunaan ADASYN dan ROS [9].

Penelitian oleh Benaya dan Prasetyo menyebutkan bahwa dalam klasifikasi kanker paru-paru, penggunaan algoritma Random Forest dengan ADASYN memberikan hasil yang cukup baik secara keseluruhan dengan nilai F1-score yang seimbang [10]. Studi lain menunjukkan bahwa kombinasi ADASYN lebih efektif dibandingkan dengan SVM kombinasi SMOTE dalam meningkatkan akurasi prediksi diabetes tipe 2 [11]. Selain itu, studi lain menyebutkan bahwa Random Forest dengan teknik ADASYN mampu meningkatkan akurasi klasifikasi hingga 5-10% lebih tinggi dibandingkan metode SMOTE dan SMOTE-ENN [12].

Meskipun telah banyak penelitian yang membahas implementasi teknik balancing data seperti SMOTE dan ADASYN, masih belum banyak studi yang secara spesifik membandingkan pengaruh kedua metode tersebut dalam konteks klasifikasi Sindrom Metabolik menggunakan algoritma Random Forest. Sebagian besar studi masih berfokus pada penyakit lain, seperti kanker, diabetes dan stroke serta menggunakan algoritma yang berbeda, seperti SVM, KNN dan CatBoost.

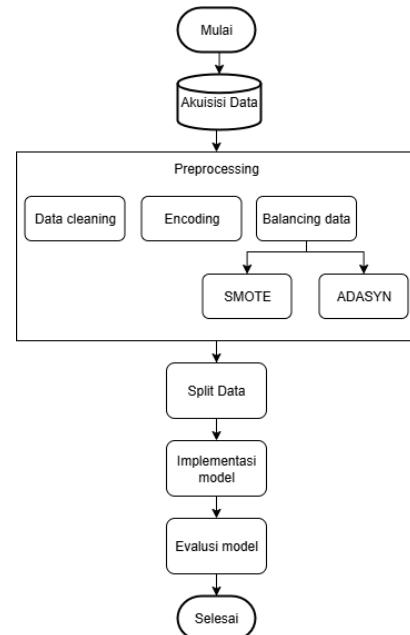
Berdasarkan pertimbangan tersebut, penelitian ini menggunakan SMOTE dan ADASYN sebagai metode penanganan class imbalance karena terbukti efektif dalam menangani class imbalance dalam dataset medis, sekaligus untuk membatasi lingkup penelitian. Penelitian ini juga difokuskan pada algoritma Random Forest karena telah terbukti menghasilkan performa yang baik dalam berbagai dataset medis, seperti stroke, diabetes dan kanker paru-paru.

Penelitian ini dilakukan untuk mengkaji pengaruh penerapan teknik oversampling SMOTE dan ADASYN terhadap performa algoritma Random Forest dalam klasifikasi Sindrom Metabolik. Fokus penelitian ini adalah untuk menganalisis perbandingan performa model menggunakan confusion matrix dan ROC-AUC. Hasil

penelitian ini diharapkan dapat memberikan kontribusi dalam pemilihan teknik penanganan class imbalance yang tepat, khususnya dalam pengembangan model klasifikasi penyakit, serta dapat menjadi acuan bagi peneliti berikutnya untuk mengembangkan metode klasifikasi yang lebih baik di bidang kesehatan, terutama dalam mengatasi masalah ketidakseimbangan kelas pada dataset medis.

II. METODE

Pada penelitian ini terdapat beberapa tahapan yang dilakukan, mulai dari akuisisi data hingga proses evaluasi seperti yang terdapat pada Gambar 1.



Gambar 1. Tahapan penelitian

A. Akuisisi Data

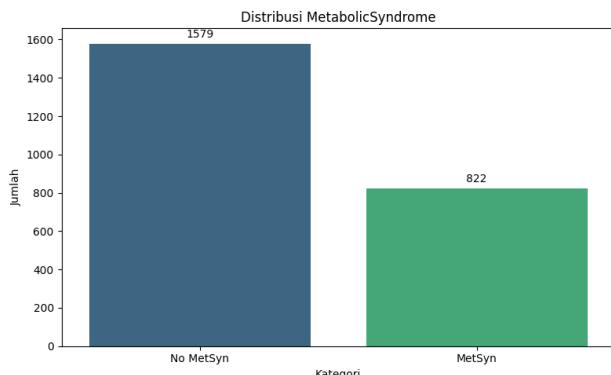
Pada tahapan pertama, dilakukan akuisisi data yang diperoleh dari sebuah repositori GitHub milik YBilodeau yang terdiri dari 15 fitur dengan 2401 entri data terkait Sindrom Metabolik.

TABEL I
DESKRIPSI FITUR DATASET

No	Fitur	Deskripsi
1	Seqn	Nomor urut responden
2	Age	Usia responden
3	Sex	Jenis kelamin
4	Marital	Status
5	Income	Pendapatan (dolar)
6	Race	Kelompok ras
7	WaisCirc	Lingkar pinggang (cm)
8	BMI	Indeks masa tubuh (kg/m^2)
9	Albuminuria	Kadar albumin dalam urin (mg/L)
10	UrAlbCr	Rasio albumin (mcg/L) terhadap kreatinin (mg/L) dalam urin

11	UricAcid	Kadar asam urat dalam darah (mg/dL)
12	BloodGlucose	Kadar glukosa dalam darah (mg/dL)
13	HDL	Kadar High Density Lipoprotein (HDL) dalam darah (mg/dL)
14	Triglycerides	Kadar trigliserida dalam darah (mg/dL)
15	MetabolicSyndrome	Diagnosa Sindrom Metabolik

Pada dataset ditemukan adanya ketidakseimbangan kelas pada variabel target yakni *MetabolicSyndrome*. Berdasarkan visualisasi distribusi kelas, jumlah kelas terdiagnosa Sindrom Metabolik (MetSyn) lebih sedikit dibandingkan yang tidak terdiagnosa Sindrom Metabolik (NoMetSyn). Secara kuantitatif jumlah kelas mayoritas (MetSyn) adalah 1579 dan jumlah kelas minoritas 822, sehingga rasio antara kelas mayoritas dan kelas minoritas adalah sekitar 2:1. Distribusi kelas target dapat dilihat pada Gambar 2.



Gambar 2. Distribusi kelas target MetabolicSyndrome

B. Preprocessing

Tahapan preprocessing ini dilakukan untuk mempersiapkan data sebelum digunakan untuk modeling. Pada proses ini dilakukan beberapa tahapan seperti data cleaning, label encoding dan balancing data. Tahapan data cleaning mencakup penanganan null, duplikasi data dan penghapusan fitur yang tidak relevan (feature selection). Feature selection dilakukan berdasarkan kriteria menurut National Heart, Lung and Blood Institute sehingga dipilih beberapa fitur seperti: *Age*, *Sex*, *WaistCirc*, *BMI*, *BloodGlucose*, *HDL* dan *Triglycerides* [2]. Selanjutnya, dilakukan label encoding untuk mengubah kolom kategorikal menjadi numerik, hal ini dilakukan karena model machine learning cenderung bekerja lebih baik pada data bertipe numerik [13]. Pada penelitian ini, encoding dilakukan pada fitur *Sex* dan *MetabolicSyndrome* yang bertipe kategorikal. Selain itu, pada tahapan preprocessing ini juga dilakukan proses balancing data guna menangani distribusi kelas yang tidak seimbang.

C. Balancing Data

Pada penelitian ini eksperimen balancing data menggunakan pendekatan oversampling SMOTE dan

ADASYN. Teknik oversampling bekerja dengan meningkatkan jumlah sampel pada kelas minoritas sehingga jumlah antara kelas mayoritas dan kelas minoritas menjadi lebih seimbang [14]. SMOTE (Synthetic Minority Oversampling Technique) merupakan metode oversampling yang bekerja dengan menciptakan data sintetis baru pada kelas minoritas [15]. SMOTE menggunakan pendekatan dengan memilih sampel acak dari kelas minoritas, kemudian mencari tetangga terdekat dari kelas minoritas tersebut, dan menghasilkan data baru yang terletak diantara kedua titik tersebut melalui proses interpolasi, seperti yang ditunjukkan pada formula berikut [16].

$$x_{syn} = x_i + \delta \cdot (x_{knn} - x_i)$$

x_{syn} = Data baru hasil sintesis
 x_i = Data yang akan direplikasi dari kelas minoritas yang dipilih acak
 δ = Bilangan acak antara 0 dan 1
 x_{knn} = Data terdekat dengan x_i dihitung dengan euclidean distance

Sementara itu, ADASYN (Adaptive Synthetic Sampling Approach) merupakan teknik penanganan ketidakseimbangan data yang berupa peningkatan dari SMOTE, dengan tujuan utama menghindari overfitting akibat dari penambahan data sintetis pada sampel minoritas ke dataset utama [17]. ADASYN bekerja dengan menambahkan sampel baru pada kelas minoritas yang sulit diklasifikasi dengan memperhatikan distribusi lokal. Berikut merupakan langkah-langkah dan formula dari ADASYN [18].

- 1) Menghitung rasio ketidakseimbangan local untuk setiap sampel minoritas, seperti formula berikut.

$$r_i = \frac{d_i}{N_i}$$

r_i = Rasio ketidakseimbangan lokal sampel minoritas i
 d_i = Jumlah tetangga mayoritas di sekitar sampel i
 N_i = Jumlah total tetangga yang digunakan

- 2) Menghitung bobot kontribusi sintetis setiap sampel minoritas, ditunjukkan pada formula berikut.

$$g_i = \frac{r_i}{\sum_{j=1}^{N_m} r_j}$$

g_i = Bobot dari data sintetis yang perlu dihasilkan sampel i
 r_i = Rasio ketidakseimbangan lokal sampel minoritas i
 N_m = Jumlah total sampel minoritas
 $\sum r_j$ = Total rasio ketidakseimbangan dari semua sampel minoritas

- 3) Menghitung jumlah total sampel sintetis yang harus dibuat, ditunjukkan pada formula berikut.

$$G = (N_m - N_{maj}) \cdot \beta$$

G = Jumlah sampel sintetis yang akan dibuat
 N_m = Jumlah total sampel minoritas dataset

- N_{maj} = Jumlah data kelas mayoritas
 β = Parameter tingkat oversampling (antara 0 dan 1)
- 4) Membangkitkan data sintetis, seperti ditunjukkan pada formula berikut.

$$X_{syn} = N_i + \delta \cdot (X_j - X_i)$$

X_{syn} = Data baru hasil sintesis
 N_i = Jumlah total tetangga yang digunakan
 δ = Bilangan acak antara 0 dan 1
 X_j = Tetangga dari sampel i
 X_i = Data yang akan direplikasi dari kelas minoritas yang dipilih acak

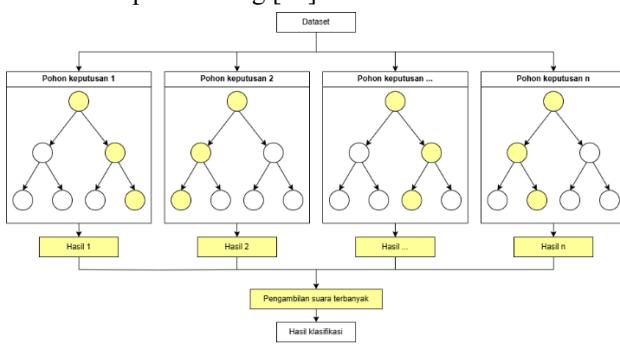
Dalam penelitian ini parameter yang digunakan untuk SMOTE dan ADASYN adalah menggunakan parameter default $k\text{-neighbors}=5$ dan $\text{sampling_strategy}=\text{'auto'}$ serta penyesuaian parameter $\text{random_state}=42$.

D. Split Data

Pada tahap ini, dataset yang telah melalui preprocessing data kemudian dibagi menjadi dua bagian, yaitu data latih (data) dan data uji (testing). Pembagian data dilakukan dengan menggunakan metode *train-test split* data dengan rasio 80:20, dimana 80% data digunakan untuk melatih model dan 20% digunakan untuk menguji kinerja model.

E. Implementasi Model Random Forest

Pada penelitian ini, algoritma Random Forest digunakan sebagai algoritma sebagai algoritma klasifikasi karena kemampuannya yang dikenal mampu mengatasi overfitting [19]. Algoritma ini bekerja dengan membangun beberapa pohon Keputusan dan menggabungkan hasil prediksi melalui pemungutan suara. Kelas yang mendapatkan hasil suara paling banyak akan digunakan sebagai hasil akhir klasifikasi. Dalam metode ini, setiap pohon dilatih menggunakan bootstrap aggregating (bagging) agar model lebih stabil dan tahan terhadap overfitting [20].



Gambar 2. Ilustrasi Algoritma Random Forest

Pada tahapan implementasi model ini, penelitian dilakukan dengan 3 skema pelatihan yang berbeda, yakni: Random Forest tanpa balancing data, Random Forest dengan kombinasi balancing SMOTE dan terakhir Random Forest dengan kombinasi balancing ADASYN. Serta dengan penyesuaian parameter $\text{random_state}=42$.

F. Evaluasi Model

Setelah dilakukan pelatihan model, tahapan terakhir adalah evaluasi performa model menggunakan confusion matrix. Confusion matrix merupakan metode evaluasi yang digunakan dalam machine learning untuk mengukur kinerja dari sebuah model klasifikasi [20]. Metode ini, menganalisis performa model berdasarkan empat kategori berikut.

- 1) True Positive (TP): data positif diprediksi benar.
- 2) True Negatif (TN): data negatif diprediksi benar.
- 3) False Positive (FP): data negatif salah diprediksi sebagai positif.
- 4) False Negatif (FN): data positif salah diprediksi sebagai negative.

TABEL 2
CONFUSION MATRIX

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Empat kategori seperti terdapat pada Tabel 2 akan digunakan sebagai dasar untuk menghitung performa, seperti berikut:

- 1) Akurasi: Tingkat keakuratan dari sebuah model dalam mengklasifikasi data, dapat dihitung berdasarkan rumus berikut.

$$\text{Akurasi} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

- 2) Presisi: Tingkat ketepatan model dalam memprediksi kelas yang diminta, berikut rumus dari presisi.

$$\text{Presisi} = \frac{TP}{(TP + FP)}$$

- 3) Recall: Tingkat sensitivitas yang menunjukkan seberapa baik model dalam menemukan kembali informasi yang benar dari dataset, berikut rumus dari recall.

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

- 4) F1-Score: Tingkat kesimbangan antara presisi dan recall, ditunjukkan dengan rumus berikut.

$$\text{F1 Score} = \frac{2 \cdot (\text{Presisi} \cdot \text{Recall})}{\text{Presisi} + \text{Recall}}$$

Selain itu digunakan juga specificity guna mengukur tingkat ketepatan model dalam memprediksi kelas negatif, dengan rumus berikut [16]:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Selain evaluasi menggunakan confusion matrix, pada penelitian ini juga digunakan evaluasi menggunakan kurva ROC-AUC. Kurva ROC digunakan untuk mengevaluasi

performa klasifikasi secara visual dengan FPR (False Positive Rate) sebagai sumbu horizontal (x) dan TPR (True Positive Rate) sebagai sumbu vertikal (y). Sementara itu, AUC (Area Under the Curve) mengukur area dibawah kurva dengan metode trapesium. Nilai AUC berkisar antara 0 hingga 1, semakin nilai mendekati 1 maka performa klasifikasi semakin baik [21].

III. HASIL DAN PEMBAHASAN

A. Proses Balancing Data

Proses penanganan class imbalance dilakukan pada variabel target, yaitu *MetabolicSyndrome* yang merepresentasikan status diagnosis Sindrom Metabolik. Pada variabel target terdapat 2 kelas, yaitu MetSyn yang merepresentasikan individu terdiagnosis Sindrom Metabolik, dan NoMetSyn yang merepresentasikan individu yang tidak terdiagnosis Sindrom Metabolik.

Distribusi awal data sebelum penanganan class imbalance menunjukkan bahwa terdapat 1.579 individu yang tidak terdiagnosis Sindrom Metabolik dan 822 individu terdiagnosis Sindrom Metabolik. Hal ini menunjukkan bahwa variabel target memiliki distribusi kelas yang tidak seimbang. Ketidakseimbangan ini akan menyebabkan model cenderung lebih mudah mengenali kelas mayoritas (NoMetSyn) dan cenderung mengabaikan kelas minoritas (MetSyn). Hal ini tentunya bisa menyebabkan banyak kasus yang sebenarnya terkena Sindrom Metabolik tidak dapat terdeteksi oleh model, sehingga prediksinya kurang akurat.

Untuk mengatasi hal tersebut, dilakukan penanganan dengan pendekatan SMOTE dan ADASYN guna meningkatkan jumlah sampel pada kelas minoritas, sebagaimana terdapat pada Tabel 3.

TABEL 3
DISTRIBUSI KELAS METABOLICSYNDROME

Metode	MetSyn	NoMetSyn
Tanpa Balancing	822	1579
SMOTE	1579	1579
ADASYN	1613	1579

B. Uji Model dan Evaluasi

Setelah dilakukan penanganan class imbalance, kemudian dilakukan pelatihan model menggunakan algoritma Random Forest. Model berhasil memperoleh performa sebagaimana terdapat pada Tabel 4.

TABEL 4
EVALUASI PERFORMA MODEL

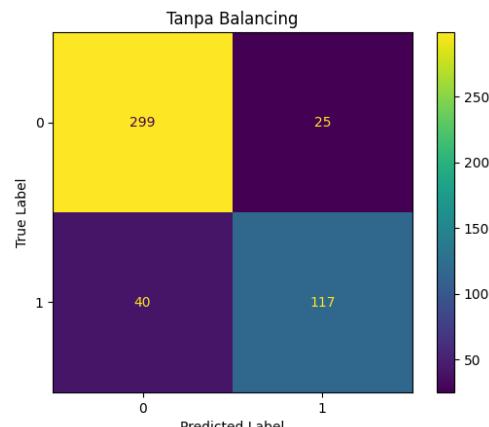
Matrix Evaluasi	Tanpa Balancing	SMOTE	ADASYN
Akurasi	86%	91%	92%
Presisi	86%	91%	92%
Recall	86%	91%	92%
F1-score	86%	91%	92%

Recall MetSyn	75%	93%	95%
Recall NoMetSyn	92%	88%	89%
Specificity	92%	88%	89%

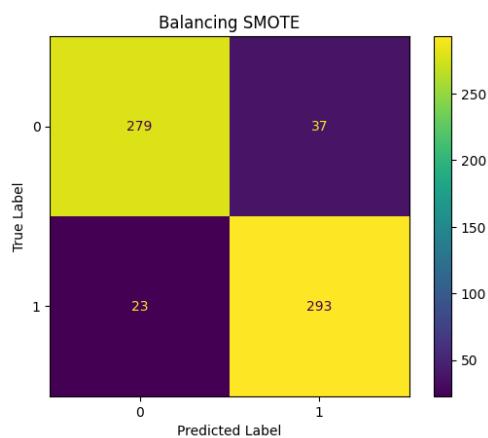
Pada pelatihan model tanpa penanganan class imbalance, diperoleh akurasi model sebesar 86%, presisi 86%, recall 86%, dan F1-score 86%. Namun, pada kelas mayoritas (NoMetSyn) nilai recall menunjukkan sebesar 92% sedangkan pada kelas minoritas (MetSyn) nilai recall sebesar 75% serta specificity sebesar 92%. Hal ini menandakan bahwa model masih cenderung bias terhadap kelas mayoritas dan masih kurang mampu mengidentifikasi individu yang seharusnya terdiagnosis Sindrom Metabolik.

Sementara itu, pada model dengan balancing data SMOTE akurasi model meningkat menjadi sebesar 91%, presisi, recall dan F1-score juga mengalami peningkatan menjadi sebesar 91%. Sementara recall pada kelas minoritas juga mengalami peningkatan menjadi sebesar 93%, yang menunjukkan bahwa model dengan balancing SMOTE lebih mampu mengidentifikasi kelas minoritas dibandingkan dengan tanpa penanganan balancing. Meskipun demikian, recall pada kelas mayoritas dan specificity terjadi sedikit penurunan menjadi sebesar 88%, hal ini dikarenakan model mengalami peningkatan dalam mengenali kelas minoritas (terdiagnosis Sindrom Metabolik).

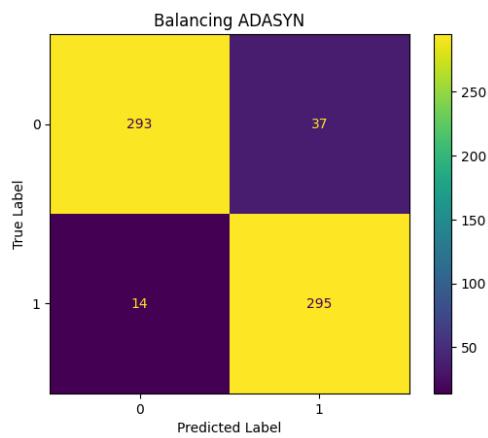
Pada skema pelatihan terakhir, model dengan balancing data ADASYN mampu memperoleh hasil akurasi terbaik. Akurasi model, presisi, recall dan F1-score meningkat menjadi sebesar 92%. Selain itu, recall pada kelas minoritas meningkat menjadi 95% yang merupakan nilai tertinggi dari pengujian sebelumnya. Recall pada kelas mayoritas dan specificity juga sedikit meningkat menjadi 89% dibandingkan dengan SMOTE, yang menunjukkan bahwa model masih baik dalam mengenali kelas mayoritas. Berdasarkan hasil tersebut, ADASYN merupakan pendekatan terbaik dalam kasus ini, karena selain mampu meningkatkan kemampuan model dalam mendeteksi individu yang terdiagnosis Sindrom Metabolik (minoritas), tetapi juga masih tetap menjaga kemampuan dalam mendeteksi individu tidak terdiagnosis Sindrom Metabolik (majoritas).



Gambar 3.a Confusion Matrix Tanpa Penanganan Balancing



Gambar 3.b Confusion Matrix dengan Penanganan Balancing SMOTE

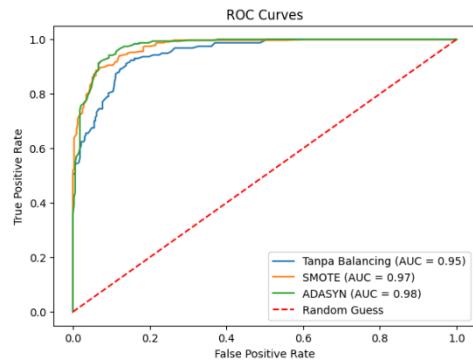


Gambar 3.c Confusion Matrix dengan Penanganan Balancing ADASYN

Berdasarkan Gambar 3.a model tanpa penanganan balancing data menunjukkan bahwa model cenderung bias terhadap kelas mayoritas (0) yaitu individu tidak terdiagnosa Sindrom Metabolik serta kurang akurat dalam memprediksi kelas minoritas (1) individu terdiagnosa Sindrom Metabolik. Model berhasil memprediksi 117 True Positif, 299 True Negatif, 25 False Positif dan 40 False Negatif.

Sementara pada Gambar 3.b model dengan penanganan balancing SMOTE, kinerja model dalam mengenali kelas minoritas (1) mengalami peningkatan. Model berhasil memprediksi 293 True Positif, 279 True Negatif, 37 False Positif dan 23 False Negatif.

Pada Gambar 3.b model dengan penanganan balancing ADASYN menunjukkan performa yang lebih baik dibandingkan metode sebelumnya. Model berhasil memprediksi 295 True Positif, 293 True Negatif, 37 False Positif dan 14 False Negatif.



Gambar 4 ROC-AUC

Setelah dilakukan evaluasi menggunakan confusion matrix, kemudian dilakukan evaluasi menggunakan kurva ROC-AUC. Pada Gambar 4 menunjukkan bahwa nilai AUC pada model tanpa balancing adalah sebesar 0,95, model dengan SMOTE sebesar 0,97 dan pada model dengan ADASYN adalah sebesar 0,98. Hal ini, menunjukkan bahwa penggunaan metode balancing sangat berpengaruh pada performa model. Dalam kasus ini, berdasarkan nilai AUC metode balancing ADASYN merupakan pendekatan terbaik dalam mendeteksi kasus Sindrom Metabolik.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa penerapan teknik penanganan class imbalance sangat berpengaruh dalam meningkatkan kemampuan model dalam mendeteksi kelas minoritas, yakni individu terdiagnosa Sindrom Metabolik. Meskipun penggunaan teknik SMOTE dan ADASYN dapat membuat model cenderung menyesuaikan diri dengan data buatan karena adanya penambahan sampel sintesis.

Pada kasus ini, performa model menggunakan kombinasi algoritma Random Forest dan ADASYN mampu memberikan hasil terbaik. Hal ini terlihat dari peningkatan akurasi, recall kelas minoritas, presisi, F1-score serta nilai AUC meskipun specificity sedikit lebih rendah. Kombinasi ini, menjadikannya sebagai metode yang paling direkomendasikan untuk mendeteksi kasus Sindrom Metabolik.

IV. KESIMPULAN

Penelitian ini bertujuan untuk mengkaji pengaruh penerapan teknik penanganan class imbalance, yaitu SMOTE dan ADASYN, terhadap performa klasifikasi Sindrom Metabolik menggunakan algoritma Random Forest. Hasil penelitian menunjukkan bahwa teknik balancing mampu meningkatkan kemampuan model dalam mendeteksi kelas minoritas (MetSyn), yang sebelumnya sulit dikenali oleh model tanpa balancing data. Teknik ADASYN dengan kombinasi Random Forest memberikan performa terbaik dengan akurasi, presisi, recall, dan F1-score sebesar 92%, recall kelas minoritas dan specificity sebesar 95%, serta nilai AUC sebesar 0,98. Hasil ini menunjukkan bahwa pemilihan metode balancing yang tepat sangat penting dalam pengembangan model klasifikasi dengan distribusi data yang tidak seimbang, khususnya dalam konteks prediksi medis.

Meskipun demikian, SMOTE dan ADASYN memiliki keterbatasan. Sampel sintetis yang ditambahkan bisa membuat model terlalu menyesuaikan diri dengan data buatan, sehingga memungkinkan adanya risiko overfitting, meskipun Random Forest relatif tahan terhadap hal ini. Oleh karena itu, hasil model perlu diperhatikan dengan hati-hati terutama pada dataset baru.

Penelitian selanjutnya diharapkan dapat menganalisa dengan perbandingan berbagai algoritma klasifikasi lainnya seperti XGBoost atau LightGBM, serta mengeksplorasi teknik penanganan class imbalance lainnya seperti Random Over Sampling (ROS), Random Under Sampling (RUS), Borderline-SMOTE, dan sebagainya untuk memperoleh performa model yang lebih optimal.

UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada pemilik repositori GitHub, yaitu YBilodeau yang telah membagikan dataset maupun menyediakan implementasi kode yang menjadi referensi penting bagi penelitian ini.

DAFTAR PUSTAKA

- [1] R. Zeltser, Supreya Swarup, Intisar Ahmed, Yulia Grigorova, "Metabolic Syndrome," in *StatPearls*, StatPearls Publishing, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK459248/>
- [2] and B. I. National Heart, Lung, "What Is Metabolic Syndrome?," May 18, 2022. [Online]. Available: <https://www.nhlbi.nih.gov/health/metabolic-syndrome>
- [3] S. M. Mohamed, M. A. Shalaby, R. A. El-Shiekh, H. A. El-Banna, S. R. Emam, and A. F. Bakr, "Metabolic syndrome: risk factors, diagnosis, pathogenesis, and management with natural approaches," *Food Chem. Adv.*, vol. 3, p. 100335, Dec. 2023, doi: 10.1016/J.FOCHA.2023.100335.
- [4] M. F. Dwi Yulianto, T. Y. Miko Wahyono, and H. Helda, "Sindrom Metabolik dan Kejadian Stroke pada Penduduk Berusia ≥ 15 Tahun di Indonesia: Analisis Data Riskesdas 2018," *J. Epidemiol. Kesehat. Indones.*, vol. 7, no. 1, p. 59, 2023, doi: 10.7454/epidkes.v7i1.6959.
- [5] F. Fadmadika *et al.*, "Pengaruh smote terhadap performa algoritma random forest dan algoritma gradient boosting dalam memprediksi penyakit stroke," vol. 7, pp. 837–846, 2024, doi: 10.37600/tekinkom.v7i2.1575.
- [6] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 24–32, 2022, doi: 10.28932/jutisi.v8i1.4004.
- [7] F. Sholekhah, A. D. Putri, R. Rahmaddeni, and L. Efrizoni, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 507–514, 2024, doi: 10.57152/malcom.v4i2.1249.
- [8] V. No, R. Irfannandhy, L. B. Handoko, and N. Ariyanto, "Edumatic : Jurnal Pendidikan Informatika Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes," vol. 8, no. 2, pp. 714–723, 2024, doi: 10.29408/edumatic.v8i2.27990.
- [9] F. A. E. Putri, "Pengaruh penanganan ketidakseimbangan kelas pada dataset penyakit stroke terhadap performa Algoritma Random Forest," Universitas Islam Negeri Maulana Malik Ibrahim, 2024. [Online]. Available: <http://etheses.uin-malang.ac.id/id/eprint/70971>
- [10] D. Benaya, "Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru," *JOINTER J. Informatics Eng.*, vol. 5, no. 01, pp. 27–31, 2024, doi: 10.53682/jointer.v5i01.331.
- [11] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.
- [12] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.
- [13] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, vol. 15, no. 1, p. 93, 2024, doi: 10.31602/tji.v15i1.13457.
- [14] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.
- [15] A. Syukron, S. Sardiarinto, E. Saputro, and P. Widodo, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *J. Teknol. Inf. dan Terap.*, vol. 10, no. 1, pp. 47–50, 2023, doi: 10.25047/jtit.v10i1.313.
- [16] A. A. G. W. S. Erlangga, I. G. A. Gunadi, and I. M. G. Sunarya, "Kombinasi Oversampling dan Undersampling dalam Menanganai Class Imbalanced dan Overlapping pada Klasifikasi Data Bank Marketing," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 7, no. 1, pp. 32–42, 2024, doi: 10.31598/jurnalresistor.v7i1.1515.
- [17] J. Al Amien, Yoze Rizki, and Mukhlis Ali Rahman Nasution, "Implementasi Adasyn Untuk Imbalance Data Pada Dataset UNSW-NB15 Adasyn Implementation For Data Imbalance on UNSW-NB15 Dataset," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 3, no. 3, pp. 242–248, 2022, doi: 10.37859/coscitech.v3i3.4339.
- [18] M. Tiara *et al.*, "PEMANFAATAN ALGORITMA ADASYN DAN SUPPORT VECTOR MACHINE DALAM MENINGKATKAN AKURASI PREDIKSI KANKER PARU-PARU," vol. 8, no. 5, pp. 8773–8778, 2024.
- [19] I. Setiawan, I. F. Yasin, Y. T. Desianti, and A. Surakarta, "Komparasi Kinerja Algoritma Random Forest , Decision Tree , Naïve Bayes , dan KNN dalam Prediksi Tingkat Depresi Mahasiswa Menggunakan Student Depression Dataset," vol. 6, no. 1, pp. 47–58, 2025.
- [20] Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, and Khairul Reza Bakara, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *J. Penelit. Rumpun Ilmu Tek.*, vol. 2, no. 4, pp. 155–164, 2023, doi: 10.55606/juprit.v2i4.3039.
- [21] D. T. Wilujeng, M. Fatekurohman, and I. M. Tirta, "Analisis Risiko Kredit Perbankan Menggunakan Algoritma K-Nearest Neighbor dan Nearest Weighted K-Nearest Neighbor," *Indones. J. Appl. Stat.*, vol. 5, no. 2, p. 142, 2023, doi: 10.13057/ijas.v5i2.58426.