

Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma *Supervised Learning* untuk Penyakit Jantung

Anis Fitri Nur Masruriyah^{1*}, Hilda Yulia Novita², Cici Emilia Sukmawati³, Angga Ramda Ramadhan⁴, Siti Novianti Nuraini Arif⁵, Budi Arif Dermawan⁶

^{1,2,3,4,5}Program Studi Tekni Informatika, Fakultas Ilmu Komputer Universitas Buana Perjuangan Karawang
Jl. HS.Ronggo Waluyo, Karawang, Indonesia

⁶Program Studi Informatika, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Karawang, Indonesia

e-mail: ¹anis.masruriyah@ubpkarawang.ac.id, ²hilda.yulia@ubpkarawang.ac.id,
³cici.emilia@ubpkarawang.ac.id, ⁴if20.anggaramadhan@mhs.ubpkarawang.ac.id,
⁵if20.sitinoviantinurainiarif@mhs.ubpkarawang.ac.id, ⁶budi.arif@staff.unsika.ac.id

(*) Corresponding Author

Artikel Info : Diterima : 04-09-2023 | Direvisi : 19-10-2024 | Disetujui : 30-01-2024

Abstrak - Penyakit jantung tetap menjadi penyebab utama kematian di Indonesia dan dunia. Dalam data mining, ketidakseimbangan kelas antara sampel penyakit jantung dan normal dalam dataset adalah masalah serius. Hal ini dapat mengarah pada bias model ke arah kelas mayoritas, mengakibatkan kinerja buruk dalam mengidentifikasi kasus penyakit jantung yang lebih sedikit. Penelitian ini mengatasi masalah ini dengan menerapkan teknik oversampling, khususnya *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Adaptive Synthetic Sampling* (ADASYN). Hasilnya menunjukkan bahwa model tanpa oversampling mencapai akurasi dan presisi di atas 80%, namun memiliki kinerja yang buruk dalam memisahkan kelas. Model yang memanfaatkan oversampling, meskipun dengan penurunan akurasi dan presisi, berhasil meningkatkan kemampuan dalam membedakan kelas penyakit jantung dan normal. Model terbaik menggunakan algoritma *Random forest* dengan SMOTE mencapai nilai AUC sebesar 0.868, menandakan peningkatan yang signifikan dalam pemisahan kelas. Temuan ini memberikan panduan penting bagi pengembangan model klasifikasi penyakit jantung yang lebih efektif dan akurat. Penggunaan teknik oversampling, seperti SMOTE, terbukti menjadi strategi yang efektif untuk mengatasi ketidakseimbangan kelas dalam data mining penyakit jantung. Meskipun akurasi dan presisi dapat mengalami penurunan, kemampuan model dalam mengidentifikasi penyakit jantung menjadi lebih andal, dengan hasil yang diukur menggunakan AUC yang signifikan. Penelitian ini berkontribusi pada upaya meningkatkan pencegahan dan pengobatan penyakit jantung melalui teknik data mining yang canggih dan berkelanjutan.

Kata Kunci : ADASYN, Klasifikasi, Penyakit Jantung, SMOTE, Supervised Learning

Abstracts - Heart disease remains a leading cause of death in Indonesia and worldwide. In the realm of data mining, class imbalance between heart disease and normal samples within datasets presents a significant challenge. This disparity can lead to model bias toward the majority class, resulting in suboptimal performance in identifying instances of heart disease. This study addresses this issue by implementing oversampling techniques, particularly *Synthetic Minority Over-sampling Technique* (SMOTE) and *Adaptive Synthetic Sampling* (ADASYN). The findings reveal that models without oversampling achieve accuracy and precision exceeding 80%, but exhibit poor class separation performance. In contrast, models employing oversampling, despite experiencing reductions in accuracy and precision, enhance their ability to distinguish between heart disease and normal classes. The top-performing model utilizing the *Random forest* algorithm with SMOTE attains an AUC value of 0.868, signifying a significant improvement in class separation. These discoveries provide essential guidance for the development of more effective and accurate heart disease classification models. The utilization of oversampling techniques, such as SMOTE, proves to be an effective strategy for mitigating class imbalances in heart disease data mining. While accuracy and precision may decrease, the model's capability to identify heart disease becomes more reliable, with notable outcomes assessed using AUC. This research contributes significantly to enhancing efforts in heart disease prevention and treatment through sophisticated and sustainable data mining techniques.

Keywords : ADASYN, Classification, Heart Disease, SMOTE, Supervised Learning



PENDAHULUAN

Penyakit jantung, salah satu penyakit yang tidak menular, namun memiliki prevalensi kematian yang tinggi di Indonesia dan seluruh dunia (BHF, 2019; Braunwald, 2019; National Center for Chronic Disease Prevention and Health Promotion Division for Heart Disease and Stroke Prevention, 2022). Rata-rata pasien yang memiliki riwayat penyakit jantung berusia 30 hingga 50 tahun, perubahan gaya hidup, pola makan dan aktifitas fisik yang kurang menjadi penyebab penyakit jantung (Braunwald, 2019; Pradono & Werdhasari, 2018; Setyaji et al., 2018). Maka, apabila pasien dengan penyakit jantung tidak ditangani dengan baik, maka pasien berusia produktif memiliki risiko kematian yang tinggi (Alkhusari et al., 2020; Kementerian Kesehatan Republik Indonesia, 2021; M. H. Ramadhan, 2022). Sebelumnya, teknik data mining terbukti mampu membantu untuk menganalisa dan membantu paramedis untuk melakukan diagnosis dini penyakit jantung pada pasien (Annisa, 2019; Derisma, 2020; Firdlous, 2022; Handayani, 2021; Mia et al., 2022; Sonjaya et al., 2022).

Sebelumnya, telah dilakukan analisa komparasi algoritma pembelajaran terbimbing terhadap penyakit jantung (Annisa, 2019). Penelitian tersebut menggunakan data yang diperoleh dari UC Irvine Machine Learning Repository dengan delapan atribut dan satu target. Selanjutnya data dilakukan prapemrosesan dan dibagi menjadi data latih dan data uji menggunakan K-fold Cross Validation. Kemudian algoritma C4.5, K-Nearest Neighbor, Naïve Bayes, *Random forest* dan Decision Stump diterapkan untuk membangun model klasifikasi dengan hasil akurasi tertinggi 80.38% pada algoritma *Random forest*. Menggunakan topik yang sama namun dengan menggunakan 14 atribut dari sumber data UC Irvine Machine Learning juga telah dilakukan (Derisma, 2020). Penelitian tersebut menggunakan algoritma Naïve Bayes, Jaringan Syaraf Tiruan dan *Random forest* dengan nilai akurasi tertinggi diperoleh dari model yang dibangun dengan algoritma Naïve Bayes (83%). Masih dengan *dataset* yang sama, penelitian yang dilakukan oleh Handayani (2021) menggunakan algoritma *Support Vector Machine* (SVM), *Logistic Regression* dan Jaringan Syaraf Tiruan untuk membangun model prediksi. Penelitian tersebut tidak menjelaskan tahapan prapemrosesan dan langsung membagi data berdasarkan perbandingan 90:10, 80:20, 70:30 dan 60:40 sebagai data latih dan data uji. Model dengan akurasi terbaik menggunakan algoritma *Logistic Regression* dengan nilai 86% pada pembagian 80:20.

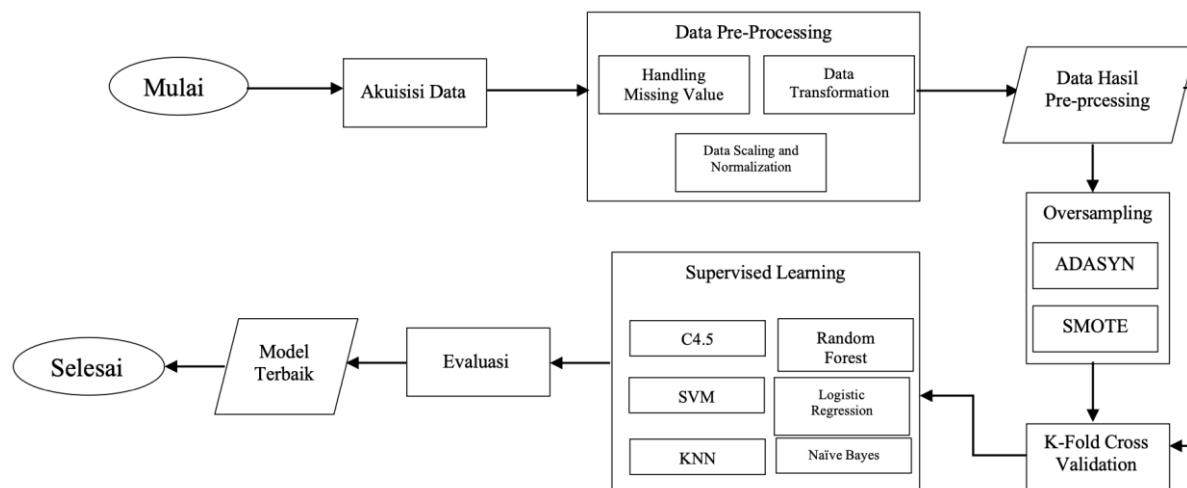
Kemudian, penelitian yang dilakukan oleh Firdlous (2022) menggunakan 12 variabel penyakit jantung membandingkan kinerja algoritma naïve bayes, pohon keputusan, *Random forest* dan SVM. Setelah data melalui tahap prapemrosesan, berikutnya data dibagi menjadi data latih dan data uji dengan perbandingan 80:20 dan dilanjutkan dengan mengimplementasikan algoritma yang telah ditentukan. Hasil akurasi menunjukkan bahwa algoritma *Random forest* memiliki performa yang lebih baik berdasarkan nilai akurasi sebesar 85.7%. Selanjutnya, penelitian yang dilakukan oleh Mia et al. (2022) dan Sonjaya et al. (2022) menerapkan teknik *oversampling Synthetic Minority Oversampling Technique* (SMOTE) dan Adaptive Syntehtic (ADASYN) untuk menangani data penyakit jantung yang tidak seimbang. Kedua penelitian tersebut menggunakan data penyakit jantung yang diperoleh dari *Centers for Disease Control and Prevention* (CDC). Selanjutnya dilakukan prapemrosesan data dan teknik *oversampling*. Setelah data siap, dilakukan pembagian data menggunakan *K-Fold Cross Validation* dengan K-10. Pada penelitian Mia et al. (2022) algoritma yang digunakan adalah pohon keputusan dan *Random forest*, sedangkan Sonjaya et al. (2022) menggunakan SVM dan *Logistic Regression*. Hasil perbandingan menunjukkan pohon keputusan dan *Random forest* mengalami peningkatan akurasi setelah data dilakukan *oversampling* di mana tanpa *oversampling* akurasi pohon keputusan 86.74% meningkat menjadi 91.77% (dengan SMOTE) dan 97.71% (dengan Adasyn). Di sisi lain terjadi penurunan akurasi pada SVM dan *Logistic Regression* namun nilai presisi meningkat. Tanpa teknik *oversampling* akurasi dari SVM 91.69% dengan nilai presisi hanya 57.81% dan *Logistic Regression* memiliki akurasi 91.76 dengan presisi 54.82%. Setelah dilakukan *oversampling* dengan SMOTE akurasi SVM menurun menjadi 75.79% namun presisi meningkat menjadi 74.54%, hal yang sama terjadi pada algoritma *Logistic Regresion* dengan akurasi 75.84% dengan presisi 74.77%. Artinya *oversampling* dapat membantu kinerja dari model yang dibangun.

Sehingga penelitian ini melakukan perbandingan kinerja algoritma supervised learning yang populer digunakan pada penelitian sebelumnya, dan menggunakan teknik *oversampling* untuk mengatasi data tidak seimbang. Selanjutnya evaluasi terukur menggunakan *Receiver Operating Characteristic* (ROC) dan *Area Under the Curve* (AUC).

METODE PENELITIAN

Jenis penelitian pada naskah ini adalah eksperimen, dimana tujuan utamanya mengevaluasi model berdasarkan kinerja algoritma yang terbaik. Adapun tahapan secara umum ditunjukkan pada Gambar 1. Penelitian ini dimulai dengan mengumpulkan data dan melakukan proses preprocessing untuk mempersiapkan data sebelum dilibatkan dalam analisis lebih lanjut. Data hasil preprocessing kemudian dibagi menjadi dua percobaan utama.

Pertama, menggunakan metode K-Fold Cross-Validation untuk memastikan model dapat diuji pada berbagai kombinasi data. Kedua, menggunakan teknik oversampling untuk mengatasi ketidakseimbangan kelas sebelum menerapkan model. Selanjutnya, kedua set data diimplementasikan pada algoritma terpilih, yang kemudian dievaluasi menggunakan metrik akurasi dan ROC-AUC. Akurasi memberikan gambaran sejauh mana model dapat mengklasifikasikan dengan benar, sementara ROC-AUC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif. Evaluasi ini memberikan wawasan mendalam tentang kinerja model dan membantu peneliti memilih pendekatan yang paling efektif untuk mengatasi masalah khusus yang dihadapi dalam penelitian ini.



Gambar 1. Alur Proses Penelitian

Sumber: Penelitian (2023)

A. Data

Penelitian ini menggunakan data yang berasal dari *Center for Disease Control and Prevention* (CDC) yang telah mendapatkan persetujuan dari Organisasi Kesehatan Dunia (Centers for Disease Control and Prevention, 2020). Dataset yang digunakan terdiri dari 18 atribut yang mencerminkan riwayat pasien, dengan dua kelas yaitu 1 sebagai pasien yang mengidap penyakit jantung dan 0 sebagai pasien dalam kondisi normal. Data ini dianggap sebagai sumber informasi yang andal, mengingat kedua lembaga tersebut dikenal sebagai otoritas utama dalam bidang kesehatan global. Pemilihan atribut dan kelas pada dataset ini diharapkan dapat memberikan wawasan mendalam terkait faktor-faktor yang berkontribusi pada kondisi kesehatan jantung, yang pada gilirannya dapat mendukung pemahaman lebih lanjut dan pengembangan model klasifikasi yang akurat.

B. ADASYN

ADASYN merupakan pendekatan yang menggunakan distribusi tertimbang untuk menciptakan contoh sintetik pada kelas minoritas, dengan tingkat kesulitan yang bervariasi sejalan dengan kompleksitas pembelajaran (Haibo He et al., 2008; Nurdian et al., 2022; N. G. Ramadhan, 2021; Satapathy et al., 2021). Proses ADASYN dilaksanakan dengan menghasilkan lebih banyak contoh sintetik untuk contoh-contoh kelas minoritas yang dianggap sulit dipelajari, menjadikannya metode yang adaptif dan responsif terhadap distribusi data yang sebenarnya. Melalui strategi ini, ADASYN memberikan kontribusi signifikan dalam meningkatkan pembelajaran dengan mengurangi bias yang muncul akibat ketidakseimbangan kelas dan menggeser batas keputusan klasifikasi ke arah contoh-contoh yang dianggap sulit. Pendekatan ini dirancang untuk memperbaiki kinerja model pada kelas minoritas, yang seringkali kurang terwakili, dengan tujuan meningkatkan kemampuan model dalam mengatasi ketidakseimbangan kelas dan mengklasifikasikan contoh-contoh yang kompleks.

C. SMOTE

SMOTE adalah sebuah pendekatan terhadap konstruksi pengklasifikasi dari kumpulan data yang tidak seimbang (Chawla et al., 2002; Maldonado et al., 2019; N. G. Ramadhan, 2021). SMOTE menghasilkan data sintesis baru di antara titik data minoritas yang sudah ada. Metode ini bekerja dengan mengambil dua atau lebih titik data dari kelas minoritas, dan kemudian membuat sampel sintesis di antara titik-titik ini di ruang fitur. Tujuan utama SMOTE adalah untuk menciptakan data sintesis untuk membuat kelas minoritas lebih banyak mewakili dan membalansir distribusi kelas.

D. C4.5

Algoritma C4.5 merupakan algoritma pembelajaran pohon keputusan yang mengembangkan struktur pohon keputusan secara rekursif, dimulai dari akar yang terkait dengan atribut pada kumpulan data (Cherfi et al., 2018;

Pangaribuan et al., 2019; Rohman & Rochcham, 2018). Prinsip kerja awal dari algoritma ini melibatkan pemilihan atribut terbaik, yang didasarkan pada perhitungan informasi menggunakan metode entropi (sebagaimana tercantum pada persamaan 1) dan gain rasio (sebagaimana tercantum pada persamaan 2), untuk menghasilkan dua cabang (child node) dari data. Proses tersebut diulangi secara rekursif pada setiap cabang hingga mencapai suatu kondisi penghentian, seperti ketika semua data dalam suatu cabang berasal dari kelas yang seragam atau ketika kedalaman maksimum tercapai. Setelah struktur pohon terbentuk, dilakukan proses pemangkasan (pruning) untuk menghindari overfitting dengan mengeliminasi cabang-cabang yang tidak signifikan atau berpotensi menyebabkan overfitting. Selanjutnya, model diujikan dengan menerapkan data melalui struktur pohon, mengikuti jalur yang sesuai dengan atribut data, hingga mencapai simpul daun yang memiliki label kelas.

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \tag{1}$$

$$Information\ Gain = Entropy(sebelum) - \sum_{j=1}^k Entropy(j, sesudah) \tag{2}$$

E. *Random forest*

Metode pembelajaran ansambel, dikenal sebagai Algoritma *Random forest*, mengintegrasikan sejumlah besar pohon keputusan untuk tujuan klasifikasi (Hartshorn, 2020; Khasanah et al., 2021; Primajaya & Sari, 2018). Algoritma ini mengadopsi teknik *Bootstrap Aggregating* (Bagging), di mana setiap pohon yang dibangun dalam *Random forest* dipelajari dari sampel acak yang diambil dengan penggantian (bootstrap) dari kumpulan data pelatihan. Oleh karena itu, setiap pohon dalam ansambel mendapatkan pelatihan pada dataset yang mengalami sedikit variasi. Selanjutnya, dilaksanakan seleksi atribut secara acak melalui proses Random Feature Selection, di mana atribut dipilih dari keseluruhan himpunan atribut yang ada. Langkah ini dirancang untuk menghindari kemungkinan overfitting dan menghasilkan variasi di antara pohon-pohon yang membentuk ansambel. Pada tahap akhir, dilakukan proses prediksi di mana setiap pohon dalam ansambel memberikan prediksi kelasnya. Dalam konteks klasifikasi, kelas yang paling umum diambil sebagai hasil akhir.

F. SVM

Algoritma SVM secara efektif menentukan *hyperplane* dengan margin terbesar antara dua kelas, sebagaimana dikemukakan dalam studi sebelumnya (Febriawan et al., 2019; Koda et al., 2018; Noor et al., 2019; Pisner & Schnyer, 2020; N. G. Ramadhan, 2021). Margin didefinisikan sebagai jarak antara *hyperplane* dan titik data terdekat dari masing-masing kelas, yang dikenal sebagai vektor pendukung. Keunggulan SVM dapat diperluas melalui penggunaan kernel. Kernel, sebagai suatu fungsi, mentransformasikan data awal ke dimensi yang lebih tinggi, mempermudah pemisahan data. Tabel 1 menunjukkan berbagai jenis fungsi kernel yang dapat digunakan. Pendekatan ini memungkinkan SVM untuk mengatasi permasalahan klasifikasi non-linier pada dimensi aslinya. Setelah *hyperplane* ditemukan, SVM mampu memprediksi kelas data pengujian berdasarkan posisinya terhadap *hyperplane* tersebut.

Tabel 1. Jenis Kernel

SVM	Kernel	Persamaan
Linear	Linier	$K(x, y) = x \times y$
	<i>Polynomial</i>	$K(x, y) = (x \times y + 1)^p$
Non-Linear	RBF	$K(x, y) = \exp\left(\frac{\ x - y\ ^2}{2\sigma^2}\right)$
	<i>Sigmoid</i>	$K(x, y) = \tanh(cx^T y + h)$

Sumber : Penelitian (2023)

G. K-Nearest Neighbor (KNN)

K-NN adalah algoritma klasifikasi yang berdasarkan konsep bahwa data dengan atribut yang mirip cenderung berada dalam kelas yang sama (Noor et al., 2019; Rahayu et al., 2017; Yana & Nafi'iyah, 2021). Ketika memiliki dataset pelatihan, K-NN mempertahankan dataset tersebut sebagai referensi. Ketika ingin mengklasifikasikan data baru, algoritma ini mencari K (jumlah tetangga terdekat yang Anda tentukan sebelumnya) data pelatihan yang paling mirip dengan data baru berdasarkan jarak (umumnya jarak Euclidean atau lainnya). Setelah K tetangga terdekat ditemukan, K-NN mengambil mayoritas kelas dari tetangga-tetangga tersebut sebagai prediksi untuk data baru.

H. Naïve Bayes

Naive Bayes adalah algoritma klasifikasi probabilistik yang berdasarkan pada Teorema Bayes (Dimsyiar

M Al Hafiz et al., 2021; Ratnawati, 2018). Naive Bayes menghitung probabilitas bahwa data baru termasuk dalam setiap kelas berdasarkan atribut yang ada dalam data tersebut. Ini melibatkan penggunaan asumsi "naive" bahwa atribut-atribut tersebut adalah independen satu sama lain. Dengan kata lain, Naive Bayes menganggap bahwa atribut-atribut ini tidak memiliki ketergantungan satu sama lain. Dengan menggunakan Teorema Bayes, Naive Bayes menghitung probabilitas posterior untuk setiap kelas dan memilih kelas dengan probabilitas posterior tertinggi sebagai prediksi.

I. K-Fold Cross Validation

K-Fold Cross-Validation adalah teknik yang umum digunakan dalam pembelajaran mesin dan statistik untuk mengevaluasi kinerja model atau algoritma (Djatna et al., 2018; Masruriyah et al., 2019; Mia et al., 2022; Nur Masruriyah et al., 2021; Sonjaya et al., 2022). Ini digunakan untuk mengukur sejauh mana model yang dikembangkan mampu melakukan generalisasi pada data yang belum pernah dilihat sebelumnya. K-Fold Cross-Validation memecah dataset menjadi beberapa "lipatan" (folds) dan melakukan beberapa iterasi pelatihan dan pengujian. Setiap hasil prediksi selanjutnya *Confusion Matrix* digunakan untuk menghitung metrik evaluasi kinerja model untuk iterasi tertentu. Metrik-metrik ini termasuk akurasi, presisi, recall (sensitivitas) dan spesifisitas. Isi dari Confusion Matrix terdapat pada Tabel 2. *Confusion Matrix* terdiri dari empat metrik utama yang dihasilkan dari hasil prediksi model, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP berisi jumlah data yang diprediksi dengan benar oleh model sebagai positif dan positif pada data sebenarnya. Selanjutnya, TN adalah total data yang diprediksi secara tepat negatif oleh model dan negatif pada data sebenarnya. Kemudian, FP memuat jumlah data yang salah diprediksi oleh model sebagai positif padahal dalam data sebenarnya negatif. Juga dikenal sebagai kesalahan Tipe I atau Alarm Palsu. Terakhir, FN adalah jumlah data yang salah diprediksi oleh model sebagai negatif padahal dalam data sebenarnya positif.

Kombinasi K-Fold Cross-Validation dengan Confusion Matrix membantu menghindari bias dalam evaluasi model, karena model diuji pada beberapa subset data yang berbeda (Agtira et al., 2023; Djatna et al., 2018; Masruriyah et al., 2019). Ini memberikan gambaran yang lebih obyektif tentang seberapa baik model yang dibangun dapat menggeneralisasi dan menghindari overfitting. Selain itu, dengan menggunakan Confusion Matrix dapat memahami lebih dalam tentang jenis kesalahan yang dibuat oleh model.

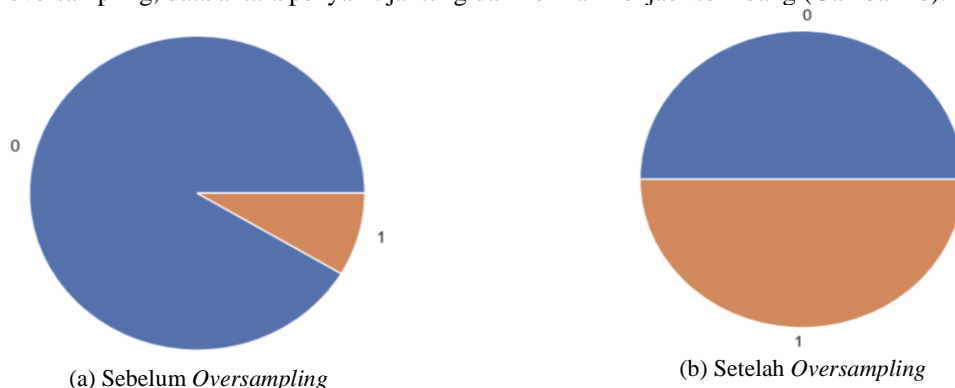
Tabel 2. *Confusion Matrix*

		Prediksi	
		Positive	Negative
Sebenarnya	Positive	True Positive (TP)	True Negative (TN)
	Negative	False Positive (FP)	False Negative (FN)

Sumber: Penelitian (2023)

HASIL DAN PEMBAHASAN

Hasil dari *pre-processing* berupa data tanpa *missing value*, dengan jenis data seragam berupa integer atau numerik. Setelah data siap perbedaan hasil teknik *oversampling* ditunjukkan pada Gambar 2. Di mana, pada Gambar 2a data target penyakit jantung belum seimbang, mayoritas adalah data pasien normal. Namun setelah dilakukan *oversampling*, data antara penyakit jantung dan normal menjadi seimbang (Gambar 2b).



Gambar 2. Perbandingan Hasil *Oversampling*

Sumber: Penelitian (2023)

Hasil penelitian disajikan dalam Tabel 3. Tabel ini menunjukkan data akurasi dan presisi dari model-model yang dikembangkan tanpa menerapkan teknik oversampling, sebelumnya diproses dengan K-Fold Cross Validation K=10. Dalam konteks ini, *Random forest* menonjol dengan konsistensi yang luar biasa, mencapai akurasi sebesar 89% dan presisi sebesar 87%, menjadikannya pilihan utama. Sementara itu, C4.5 mengungguli algoritma lainnya dengan akurasi sebesar 91%. Terdapat perbedaan signifikan dalam performa SVM yang menunjukkan akurasi terendah, hanya sekitar 18%.

Pada tahap berikutnya, data yang telah mengalami oversampling juga diproses dan hasilnya dicantumkan dalam Tabel 3. Ditemukan bahwa secara umum, terjadi penurunan akurasi dan presisi antara 5-15% pada berbagai algoritma, kecuali untuk SVM yang mengalami peningkatan yang mencolok, melebihi 30%. Tabel ini menjadi suatu panduan komprehensif untuk memahami dampak oversampling pada hasil eksperimen dan memberikan gambaran holistik tentang performa relatif dari berbagai algoritma yang diterapkan dalam penelitian ini.

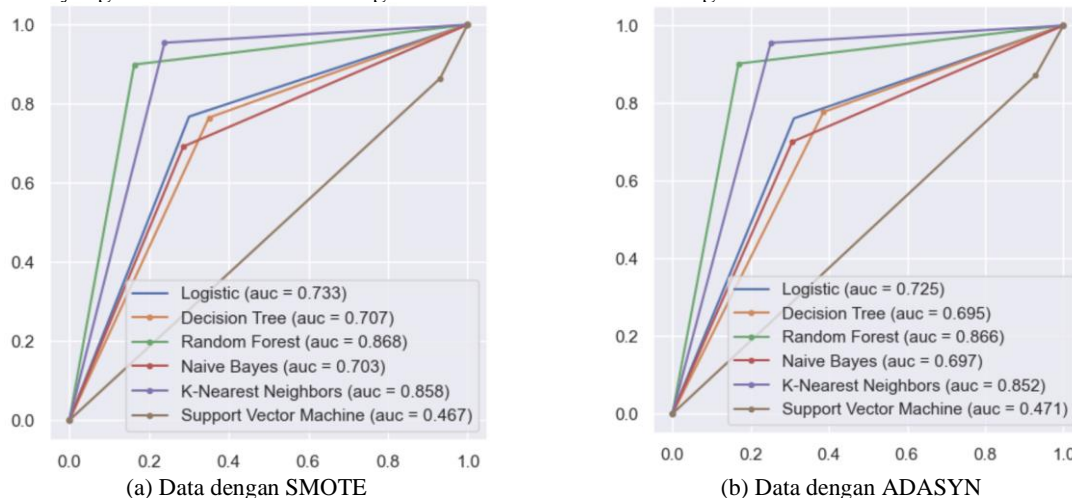
Tabel 3. Hasil Perbandingan

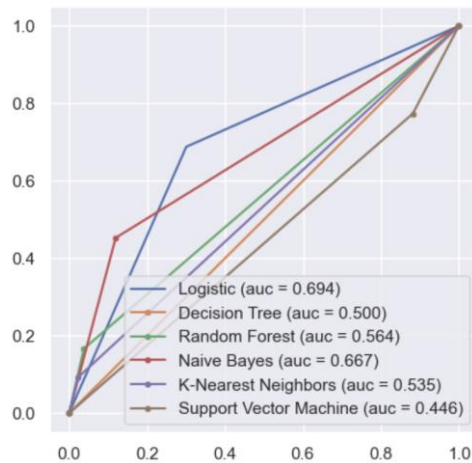
Algoritma	Tanpa Oversampling		SMOTE		ADASYN	
	Akurasi (%)	Presisi (%)	Akurasi (%)	Presisi (%)	Akurasi (%)	Presisi (%)
C4.5	91	83	70	70	69	70
<i>Random forest</i>	89	87	87	87	81	81
SVM	18	78	52	56	52	55
Logistic Regression	72	72	73	73	72	72
KNN	90	86	86	87	82	82
Naïve Bayes	84	88	70	70	69	69

Sumber: Penelitian (2023)

Meskipun nilai akurasi dan presisi memberikan indikasi awal tentang kinerja model, penelitian ini mengevaluasi secara menyeluruh melalui penggunaan kurva *Receiver Operating Characteristic* (ROC) dan *Area Under the Curve* (AUC) sebagai alat evaluasi tambahan. Gambar 3 mempresentasikan hasil evaluasi ini. Meskipun model tanpa penerapan teknik oversampling menunjukkan tingkat akurasi dan presisi yang melebihi 80%, perhatian khusus diberikan terhadap nilai AUC yang berada di bawah 0.7. Penurunan ini mengindikasikan perlunya peningkatan atau revisi model untuk meningkatkan kemampuannya dalam membedakan kelas.

Sebaliknya, penerapan teknik oversampling, meskipun disertai dengan penurunan nilai akurasi, memberikan hasil yang lebih memuaskan ketika meninjau nilai AUC yang melampaui 0.8. Hasil ini memberikan indikasi bahwa model yang memanfaatkan teknik oversampling memiliki kemampuan yang baik dalam membedakan antara kelas. Temuan ini menyajikan perspektif yang lebih holistik dan informatif terkait keandalan model, khususnya dalam konteks penanganan ketidakseimbangan kelas. Pentingnya evaluasi menggunakan ROC dalam konteks penelitian ini terletak pada kemampuannya memberikan wawasan tambahan yang belum dieksplorasi sebelumnya. Hasil ini dapat memberikan arahan bagi pengembangan dan peningkatan model klasifikasi yang lebih efektif dalam mengatasi masalah ketidakseimbangan kelas.





(c) Data Tanpa Teknik Oversampling

Gambar 3. Perbandingan Hasil Evaluasi Model

Sumber: Penelitian (2023)

KESIMPULAN

Berdasarkan hasil penelitian yang disajikan, penelitian ini menyimpulkan bahwa teknik oversampling, secara umum, membuktikan diri sebagai alat yang efektif dalam meningkatkan kinerja algoritma Supervised Learning dalam konteks klasifikasi penyakit jantung. Hasil penelitian secara khusus menunjukkan peningkatan kualitas pemisahan kelas antara penyakit jantung dan kategori normal, terutama pada model terbaik yang mengadopsi algoritma *Random forest* dengan penerapan metode oversampling SMOTE, yang mencapai nilai Area Under the Curve (AUC) yang mengesankan sebesar 0.868. Namun, penelitian ini juga mengungkapkan adanya penurunan akurasi setelah penerapan oversampling. Ini memunculkan pertanyaan tentang hubungan antara oversampling dan tahap pre-processing. Diperlukan analisis lebih lanjut yang mendalam untuk memahami dampak oversampling pada berbagai jenis data yang digunakan dalam klasifikasi penyakit jantung.

Dalam konteks kelemahan penelitian ini, ada sejumlah pertimbangan yang harus diperhatikan. Pertama, hasil penelitian ini mungkin tergantung pada pemilihan algoritma klasifikasi yang digunakan, dan oleh karena itu, penelitian lebih lanjut dapat menginvestigasi penggunaan algoritma lain. Kedua, penting untuk menyadari bahwa hasil ini bersifat khusus untuk klasifikasi penyakit jantung dan mungkin tidak langsung berlaku untuk masalah klasifikasi lainnya. Saran penelitian selanjutnya melibatkan penelitian lanjutan dalam beberapa aspek. Pertama, penelitian mendatang harus mempertimbangkan pengembangan teknik pre-processing yang dapat mengatasi penurunan akurasi setelah oversampling. Kedua, untuk memperluas generalisasi temuan, penelitian dapat mencoba menerapkan teknik oversampling dalam berbagai konteks klasifikasi penyakit.

REFERENSI

- Agtira, B. H., Handayani, H. H., & Masruriyah, A. F. N. (2023). Perbandingan Algoritma NBC dan Decision Tree pada Sentimen Analisis Mengenai Vaksinasi Covid-19 Di Indonesia. *Remik*, 7(1). <https://doi.org/10.33395/remik.v7i1.12151>
- Alkhusari, Handayani, M., Saputra, M. A. S., & Romadhon, M. (2020). Analisis Kejadian Penyakit Jantung Koroner Di Poliklinik Jantung. *Jurnal 'Aisyiyah Medika*, 5, 99–110.
- Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung. *Jurnal Teknik Informatika Kaputama (JTIK)*, 3(1).
- BHF. (2019). UK Factsheet. *British Heart Foundation*, April, 1–21.
- Braunwald, E. (2019). Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. In *Elsivier* (Vol. 7, Number 2).
- Centers for Disease Control and Prevention. (2020). *BRFSS Survey Data and Documentation*. Centers for Disease Control and Prevention.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: *Synthetic Minority Over-sampling Technique*. In *Journal of Artificial Intelligence Research* (Vol. 16).
- Cherfi, A., Nouria, K., & Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. *Applied Artificial Intelligence*, 32(2), 119–137. <https://doi.org/10.1080/08839514.2018.1447479>

- Derisma, D. (2020). Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining. *Journal of Applied Informatics and Computing*, 4(1). <https://doi.org/10.30871/jaic.v4i1.2152>
- Dimsyar M Al Hafiz, Khoirul Amaly, Javen Jonathan, M Teranggono Rachmatullah, & Rosidi. (2021). Sistem Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes. *Jurnal Rekayasa Elektro Sriwijaya*, 2(2). <https://doi.org/10.36706/jres.v2i2.29>
- Djatna, T., Hardhienata, M. K. D., & Masruriyah, A. F. N. (2018). An intuitionistic fuzzy diagnosis analytics for stroke disease. *Journal of Big Data*, 5(1). <https://doi.org/10.1186/s40537-018-0142-7>
- Febriawan, H. K., Helmholz, P., & Parnum, I. M. (2019). Support vector machine and decision tree based classification of side-scan sonar mosaics using textural features. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(2/W13), 27–34. <https://doi.org/10.5194/isprs-archives-XLII-2-W13-27-2019>
- Firdlous, D. A. (2022). Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Penyakit Jantung. *Infoman's : Jurnal Ilmu-Ilmu Manajemen Dan Informatika*, 16(1).
- Haibo He, Yang Bai, Eduardo A. Garcia, & Shutao Li. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*.
- Handayani, F. (2021). Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, 7(3). <https://doi.org/10.26418/jp.v7i3.48053>
- Hartshorn, S. (2020). *Machine Learning with Random forest and Decision Tree*.
- Kementrian Kesehatan Republik Indonesia. (2021). *Penyakit Jantung Koroner Didominasi Masyarakat Kota*. <https://www.kemkes.go.id/article/view/21093000002/penyakit-jantung-koroner-didominasi-masyarakat-kota.html>
- Khasanah, N., Komarudin, R., Afni, N., Maulana, Y. I., & Salim, A. (2021). Skin Cancer Classification Using Random forest Algorithm. *Sisfotenika*, 11(2), 137. <https://doi.org/10.30700/jst.v11i2.1122>
- Koda, S., Zeggada, A., Melgani, F., & Nishii, R. (2018). Spatial and Structured SVM for Multilabel. *Ieee Transactions on Geoscience and Remote Sensing*, 1–13.
- Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing Journal*, 76, 380–389. <https://doi.org/10.1016/j.asoc.2018.12.024>
- Masruriyah, A. F. N., Djatna, T., Dewi Hardhienata, M. K., Handayani, H. H., & Wahiddin, D. (2019). Predictive Analytics For Stroke Disease. *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*. <https://doi.org/10.1109/ICIC47613.2019.8985716>
- Mia, M., Masruriyah, A. F. N., & Pratama, A. R. (2022). The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease. *JURNAL SISFOTEK GLOBAL*, 12(2), 138. <https://doi.org/10.38101/sisfotek.v12i2.551>
- National Center for Chronic Disease Prevention and Health Promotion Division for Heart Disease and Stroke Prevention. (2022, June). *Heart Disease Maps and Data Sources*. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/statistical_reports.htm
- Noor, N. Bin, Anwar, M. S., & Dey, M. (2019). Comparative Study between Decision Tree, SVM and KNN to Predict Anaemic Condition. *BECITHCON 2019 - 2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, December*, 24–28. <https://doi.org/10.1109/BECITHCON48839.2019.9063188>
- Nur Masruriyah, A. F., Basri, H., Handayani, H. H., Fauzi, A., Juwita, A. R., & Wahiddin, D. (2021). The Rise Efficiency of Coronavirus Disease Classification Employing Feature Extraction. *2021 6th International Conference on Informatics and Computing, ICIC 2021*. <https://doi.org/10.1109/ICIC54025.2021.9632914>
- Nurdian, R. A., Mujib Ridwan, & Ahmad Yusuf. (2022). Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru. *Jurnal Teknik Informatika Dan Sistem Informasi*, 8(1). <https://doi.org/10.28932/jutisi.v8i1.4004>
- Pangaribuan, J. J., Tedja, C., & Wibowo, S. (2019). Perbandingan Metode Algoritma C4.5 dan Extreme Learning Machine untuk Mendiagnosis Penyakit Jantung Koroner. In *PSDKU Medan Jurusan Teknik Informatika INFORMATICS ENGINEERING RESEARCH AND TECHNOLOGY*.
- Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning: Methods and Applications to Brain Disorders*. Elsevier Inc. <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>
- Pradono, J., & Werdhasari, A. (2018). Faktor Determinan Penyakit Jantung Koroner pada Kelompok Umur 25-65 tahun di Kota Bogor, Data Kohor 2011-2012. *Buletin Penelitian Kesehatan*, 46(1), 23–34. <https://doi.org/10.22435/bpk.v46i1.48>
- Primajaya, A., & Sari, B. N. (2018). Random forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, 1(1), 27. <https://doi.org/10.24014/ijaidm.v1i1.4903>

- Rahayu, S., Bharata Adji, T., Akhmad Setiawan, N., & Teknik Elektro dan Teknologi Informasi, D. (2017). Penghitungan k-NN pada Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADASYN-kNN) untuk Data Nominal-Multi Kategori. *Ktrl.Inst (J.Auto.Ctrl.Inst)*, 9(2), 2017.
- Ramadhan, M. H. (2022). Faktor Risiko Penyakit Jantung Koroner (PJK). *Jurnal Kedokteran Syariah Kuala*.
- Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, 8(2). <https://doi.org/10.15294/sji.v8i2.32484>
- Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *INOVTEK Polbeng - Seri Informatika*, 3(1), 50. <https://doi.org/10.35314/isi.v3i1.335>
- Rohman, A., & Rochcham, D. M. (2018). MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG. In *Jurnal Neo Teknika* (Vol. 4, Number 2).
- Satapathy, S. K., Mishra, S., Mallick, P. K., & Chae, G. S. (2021). ADASYN and ABC-optimized RBF convergence network for classification of electroencephalograph signal. *Personal and Ubiquitous Computing*. <https://doi.org/10.1007/s00779-021-01533-4>
- Setyaji, D. Y., Prabandari, Y. S., & Gunawan, I. M. A. (2018). Aktivitas fisik dengan penyakit jantung koroner di Indonesia The relationships of physical activity with coronary heart disease in Indonesia. *Jurnal Gizi Klinik Indonesia*, 14(3), 115–121. <https://jurnal.ugm.ac.id/jgki>
- Sonjaya, C. B., Masruriyah, A. F. N., Kusumaningrum, D. S., & Pratama, A. R. (2022). The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease. *INTERNAL (Information System Journal)*, 5(2), 166–175. <https://doi.org/10.32627>
- Yana, Y. E., & Nafi'iyah, N. (2021). Klasifikasi Jenis Pisang Berdasarkan Fitur Warna, Tekstur, Bentuk Citra Menggunakan SVM dan KNN. *RESEARCH: Journal of Computer, Information System & Technology Management*, 4(1), 28. <https://doi.org/10.25273/research.v4i1.6687>