MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 8 | No. 2 | Halaman 242 - 253
DOI : https://doi.org/10.26760/mindjournal.v8i2.242-253 Desember 2023

Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling

ANIS FITRI NUR MASRURIYAH, HILDA YULIA NOVITA, CICI EMILIA SUKMAWATI, SITI NOVIANTI NURAINI ARIF, ANGGA RAMDA RAMADHAN

Program Studi Informatika, Universitas Buana Perjuangan Karawangsa
Email: anis.masruriyah@ubpkarawang.ac.id

Received 04 September 2023 | *Revised* 17 Oktober 2023 | *Accepted* 21 November 2023

ABSTRAK

Penyakit jantung mengalami peningkatan setiap tahunnya dan menjadi penyebab kematian tertinggi di Indonesia, terutama pada usia produktif. Pola makan yang tidak seimbang dan gaya hidup tidak sehat menjadi faktor penyebab prevalensi penyakit jantung yang tinggi. Bidang ilmu kedokteran mulai beradaptasi dan mengandalkan model prediksi otomatis berbasis komputer untuk diagnosis secara tepat dan akurat. Data tentang penyakit jantung seringkali memiliki ketidakseimbangan, yaitu jumlah data pada kelas minoritas lebih kecil daripada kelas mayoritas. Oleh karena itu, teknik oversampling seperti SMOTE dan ADASYN digunakan untuk menangani masalah ini. Hasil dari penelitian ini Algoritma Random Forest Classifier menjadi model perbandingan terbaik dengan akurasi sekitar 90,71%. Penerapan teknik oversampling SMOTE + Random Forest, akurasi dapat meningkat hingga sekitar 94,54% dengan kurva ROC sebesar 98,4%. Model diagnosa yang akurat dapat menjadi media bagi tenaga medis untuk mengambil langkah pencegahan yang tepat dan meningkatkan kualitas perawatan pasien.

***Kata kunci**: ADASYN, Klasifikasi, Pohon Keputusan, Regresi, SMOTE*

ABSTRACT

Heart disease is rapidly increasing in Indonesia and has become the primary cause of death, particularly among those in their productive years. The prevalence of heart disease is due to unhealthy lifestyle choices and an imbalanced diet. The medical field is relying more heavily on computer-based automatic prediction models to ensure precise and accurate diagnoses. However, data on heart disease is frequently imbalanced, with fewer cases in the minority class. To resolve this issue, oversampling techniques such as SMOTE and ADASYN have been implemented. The study demonstrates that the Random Forest Classifier Algorithm is the most effective comparison model, with an accuracy rate of approximately 90.71%. By implementing the SMOTE + Random Forest oversampling technique, the accuracy rate increased to around 94.54%, with a ROC curve of 98.4%. A highly accurate diagnostic model is essential for enabling medical personnel to take appropriate preventive measures and enhance the quality of patient care.

***Keywords**: ADASYN, Classification, Decision Tree, Regresi, SMOTE*

1. PENDAHULUAN

Pada dunia nyata, data yang diolah umumnya tidak ideal karena belum dilakukan pra-pemrosesan untuk kesiapan data **(Raja et al., 2022)(Zaki & Wagner Jr, 2020).** Selain itu, ketika menghadapi data yang tidak seimbang, akan memiliki pengaruh yang bias karena hasil pemodelan kurang dapat diandalkan. Maka metode oversampling yang umum digunakan, SMOTE dan ADASYN memiliki peran penting untuk mengatasi data tidak seimbang pada penyakit jantung **(Maldonado et al., 2019)(Ramadhan, 2021)(Satapathy et al., 2021).** Penelitian ini menggunakan data penyakit jantung, karena penyakit ini menjadi salah satu jenis penyakit tidak menular yang peningkatan jumlah pasien dan kematian selalu meningkat setiap **tahun (Braunwald, 2019)(Mohan et al., 2019).** Penyebab utama penyakit jantung di antaranya pola makan yang tidak seimbang dan perubahan gaya hidup yang tidak sehat. Di Indonesia penyakit jantung mengalami peningkatan setiap tahunnya dan menempati peringkat tertinggi penyebab kematian terutama pada usia-usia produktif **(Kementrian Kesehatan Republik Indonesia, 2021)**. Apabila penderita penyakit jantung tidak ditangani dengan baik, maka pasien dengan usia produktif rentan mengalami cacat bahkan meninggal **(Masruriyah, Djatna, Dewi Hardhienata, et al., 2019)**.

Penggunaan teknik data mining telah menjadi sangat penting dalam memprediksi penyakit dan membantu profesional medis dalam mengatasi masalah kesehatan **(Masruriyah, Djatna, Dewi Hardhienata, et al., 2019)(Mia et al., 2022)(Sonjaya et al., 2022)**. Studi sebelumnya telah berfokus untuk menemukan model prediksi terbaik, seperti sistem prediksi penyakit jantung yang diusulkan oleh Derisma **(Derisma, 2020)**. *Machine learning* (ML) telah membantu dokter mengklasifikasikan penyakit jantung pada pasien, tetapi penanganan kumpulan data yang besar dapat memengaruhi keakuratan hasil. Keahlian dan pengalaman substansial diperlukan untuk mengolah data yang luas, seperti yang disorot dalam artikel jurnal oleh Muqorobin et al **(Muqorobin et al., 2019)**.

Sebelumnya studi yang dilakukan oleh Pangaribuan et al. **(Pangaribuan et al., 2019)**, algoritma C45 dan machine learning ekstrem menunjukkan akurasi diagnostik yang tinggi hingga 99,05% untuk penyakit jantung. Teknik ekstraksi fitur seperti histogram, haralick, dan momen rona juga digunakan dalam penelitian kanker kulit, dengan algoritma Random Forest mencapai akurasi terbaik sebesar 0,842 menggunakan momen rona **(Khasanah et al., 2021)**.

Selanjutnya, Ath et al. **(Ath et al., 2022)** melakukan penelitian dengan menggunakan algoritma ML untuk memprediksi penyakit jantung sebagai langkah pencegahan dini. Akurasi yang dicapai dengan menggunakan metode *Random Forest* dan *Logistic Regression* adalah 84,48%, menunjukkan peningkatan sebesar 1,32%. El-Hasnony et al. **(El-Hasnony et al., 2022)** mengembangkan model untuk mencegah stroke dan penyakit jantung menggunakan algoritma *machine learning*. Mezzatesta dkk. pada tahun 2019 prediksi angka kematian pada pasien Penyakit Jantung Koroner yang menjalani cuci darah menggunakan algoritma SVM, mencapai akurasi sebesar 95,25%. Demikian pula, Ghosh et al. **(Ghosh et al., 2021)** menerapkan algoritma ekstraksi fitur RELIEF dan LASSO untuk memprediksi penyakit kardiovaskular, memperoleh akurasi sebesar 99,05%. Pembelajaran aktif dilaksanakan untuk mengetahui faktor yang paling berpengaruh pada penyakit jantung.

Profesional medis dapat mengambil tindakan yang tepat untuk mencegah penyakit jantung berdasarkan faktor-faktor yang mempengaruhi ini. Namun, data penelitian mengungkapkan ketidak seimbangan kelas, di mana kelas minoritas lebih kecil dari kelas mayoritas. Berbagai pendekatan, seperti *undersampling* atau *oversampling* data, dapat mengatasi masalah ini. *Oversampling* melibatkan penyeimbangan distribusi kelas dengan mereplikasi *instance* secara

acak di sebagian kecil kelas minoritas. Penelitian ini bertujuan untuk menemukan model terbaik untuk kasus penyakit jantung dengan beberapa algoritma. Mia et al. **(Mia et al., 2022)** dan Sonjaya **(Sonjaya et al., 2022)** membandingkan kinerja klasifikasi metode SMOTE dan ADASYN dalam menangani kasus data yang tidak seimbang. Penelitian ini bertujuan untuk menentukan model terbaik dengan menerapkan beberapa algoritma untuk kasus penyakit jantung. Selain itu, teknik ekstraksi fitur akan diterapkan untuk mengidentifikasi variabel yang paling berpengaruh yang mempengaruhi penyakit jantung. Di mana, susunan artikel pada bagian 2 berisi tentang metodologi yang membahas alur proses penelitian yang telah dilakukan. Kemudian hasil dan pembahasan pada bagian 3 dan Kesimpulan di bagian 4.

2. METODOLOGI

2.1. Data

Data pasien Penyakit Jantung dengan total objek lebih dari 300.000 data dengan delapan belas variabel dan satu kelas target. Sedangkan kumpulan data didapatkan dari catatan medis yang telah dietujui oleh Organisasi Kesehatan Dunia dan dapat diakses pada *Centers for Disease Control and Prevention* (CDC) **(Centers for Disease Control and Prevention, 2020)**. Atribut pada data yang di dan penjelasannya ditunjukkan pada Tabel 1.

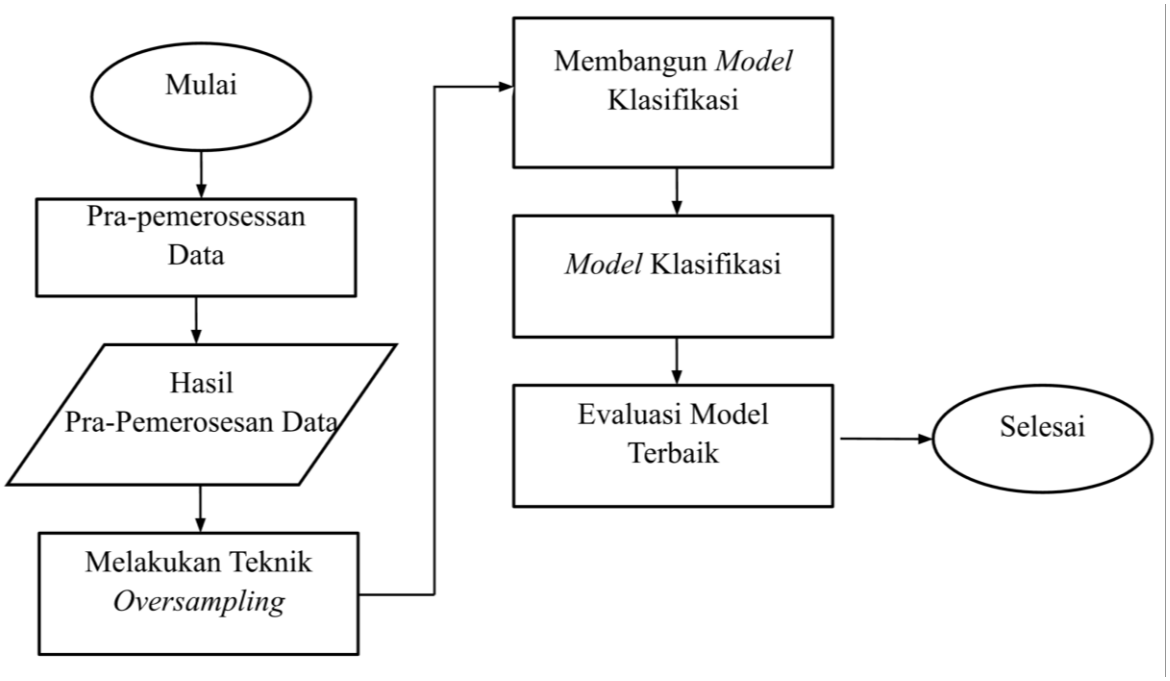
Tabel 1. Atribut Data Penelitian

Atribut	Keterangan
BMI	Indeks massa tubuh(*body mass indeks*)
Smoking	Kondisi di mana pasien merokok ditandai dengan yes atau bukan ditandai dengan no.
AlcoholDrinking	Kondisi di mana pasien peminum alcohol ditandai dengan *yes* atau bukan ditandai dengan *no*.
Stroke	Kondisi di mana pasien pengidap penyakit *stroke* ditandai dengan *yes* atau bukan ditandai dengan *no*.
PhysicalHealth	Kondisi di mana Kesehatan fisik pasien pernah mengalami penyakit atau cedera, dalam 30 hari terakhir tidak baik. Dihitung berdasarkan hari (0-30 hari).
MentalHealth	Kondisi di mana Kesehatan mental pasien pernah mengalami penyakit mental atau trauma, dalam 30 hari terakhir tidak baik. Dihitung berdasarkan hari (0-30 hari)
DiffWalking	Kondisi di mana pasien mengalami kesusahan berjalan atau menaiki/menuruni tangga.
Sex	Jenis kelamin pasien dengan *female* sebagai perempuan dan *male* sebagai pria.
AgeCategory	Rentang usia pasien dengan jarak usia 5 tahun
Race	Perbedaan ras pasien.
Diabetic	Kondisi di mana pasien pengidap penyakit diabetes ditandai dengan *yes* atau bukan ditandai dengan *no*.
PhysicalActivity	Laporan pasien melakukan aktivitas olahraga dalam 30 hari terakhir selain dari pekerjaan rutinnya.
GenHealth	Kondisi di mana kesehatan pasien secara umum sehat ditandai dengan *yes* atau tidak ditandai dengan *no*.
SleepTime	Lama waktu tidur pasien dalam satuan jam.
Asthma	Kondisi di mana pasien pengidap penyakit asma ditandai dengan *yes* atau bukan ditandai dengan *no*.
KidneyDisease	Kondisi di mana pasien pengidap penyakit ginjal kronik ditandai dengan *yes* atau bukan ditandai dengan no.
SkinCancer	Kondisi di mana pasien pengidap kanker kulit ditandai dengan *yes* atau bukan ditandai dengan *no*.

Atribut	Keterangan
HeartDisease	Hasil Berdasarkan variabel *independent* dengan kriteria 1 sebagai pengidap penyakit jantung dan 0 sehat atau normal.

2.2. Metode Penelitian

Selanjutnya, data di analisis dengan teknologi komputasi yang menerapkan aturan statistik untuk Data Mining. Secara umum, proses analisis data penelitian ini diawali dengan pengumpulan data, proses analitika dan selanjutnya ditampilkan pada Gambar 1. Penelitian ini menggunakan empat tahap analitika (*data quality analytics, descpritive analytics, diagnostic analytics dan predictive analytics*). Pada tahap pertama pra-pemrosesan data sudah termasuk *data quality analytics* dan *descpritive analytics*. Selanjutnya hasil pra-pemrosesan data diolah dan diseimbangkan menggunakan teknik *oversampling* SMOTE dan ADASYN. Teknik SMOTE mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan kumpulan data dengan melakukan *resampling* sampel kelas minoritas **(Maldonado et al., 2019)(Nurdian et al., 2022)(Siringoringo, 2018)**. Sedangkan, teknik ADASYN menggunakan bobot distribusi untuk data kelas minoritas berdasarkan kesulitan pelatihan data dengan model **(Li et al., 2021)(Nurdian et al., 2022)(Qing et al., 2022)(Satapathy et al., 2021)**. Data sintesis dihasilkan dari kelas minoritas yang sulit dipelajari dan data minoritas yang mudah dipelajari.



Gambar 1. Alur Penelitian

Agar tujuan penelitian dapat dicapai beberapa teknik klasifikasi pohon keputusan, *logistic regresi* dan *support vector machine* (SVM) digunakan untuk membuat model prediksi. Jenis algoritma pohon keputusan yang digunakan pada peneltian ini adalah *Random Forest* dan C45. Algoritma Random Forest adalah metode *ensemble learning* yang memanfaatkan banyak pohon keputusan (decision trees) dalam proses klasifikasi. Cara kerja dari *Random Forest* dimulai dengan Bootstrap Aggregating (*Bagging*) di mana *Random Forest* memilih sampel acak dengan penggantian (*bootstrap*) dari dataset pelatihan untuk setiap pohon yang akan dibuat **(Hartshorn, 2020)(Khasanah et al., 2021)(Primajaya & Sari, 2018)**. Hal ini berarti setiap pohon dalam hutan akan melihat *dataset* yang sedikit berbeda. Kemudian, *Random Feature Selection* dilakukan dengan cara memilih atribut secara acak dari seluruh atribut yang

tersedia. Ini membantu menghindari *overfitting* dan memperkenalkan variasi di antara pohon-pohon dalam hutan. Tahap terakhir, prediksi di mana setiap pohon dalam hutan memberikan prediksi kelas. Pada klasifikasi, kelas yang paling umum diambil sebagai hasil akhir.

Kemudian algoritma C45, algoritma ini bekerja dengan cara membangun pohon keputusan secara rekursif berdasarkan atribut-atribut (fitur) pada *dataset* yang diberikan **(Cherfi et al., 2018)(Mia et al., 2022)(Pangaribuan et al., 2019)(Rohman & Rochcham, 2018)**. Proses algoritma ini dimulai dengan semua data yang ada di akar pohon. Pada setiap langkah, algoritma memilih atribut yang paling baik berdasarkan pengukuran informasi berdasarkan *entropy* Persamaan (1) dan *gain information* atau *gain ratio* Persamaan (2) untuk membagi data menjadi dua cabang (*node* anak). Proses ini diulang secara rekursif pada setiap cabang hingga mencapai kriteria berhenti (misalnya, saat semua data dalam satu cabang adalah dari kelas yang sama atau saat kedalaman maksimum dicapai). Setelah pohon dibentuk, *pruning* (pemangkasan) dilakukan untuk menghindari *overfitting*. Beberapa cabang (*node*) yang tidak signifikan atau yang menyebabkan *overfitting* dapat dihapus. Selanjutnya, pengujian model dilakukan dengan meneruskan data melalui pohon, mengikuti cabang yang sesuai berdasarkan atribut-atribut data, hingga mencapai daun (*leaf*) yang memiliki label kelas.

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

$$Information\ Gain = Entropy(before) - \sum_{j=1}^{k} Entropy(j, after) \quad (2)$$

Selanjutnya algoritma *logistic regression*, bagian dari metode statistik yang umum digunakan dalam masalah klasifikasi **(Ath et al., 2022)**. *Logistic regression* menggunakan fungsi logistik (sigmoid) untuk memodelkan probabilitas bahwa data termasuk dalam kelas tertentu. Fungsi sigmoid mengubah nilai linear menjadi nilai probabilitas yang berada dalam rentang 0 hingga 1. Berdasarkan nilai probabilitas yang diperoleh dari transformasi sigmoid, kita dapat mengambil keputusan klasifikasi dengan memilih ambang batas (*threshold*). Misalnya, jika probabilitas lebih besar dari 0.5, kita bisa memprediksi sebagai kelas 1; jika probabilitas lebih kecil atau sama dengan 0.5, kita bisa memprediksi sebagai kelas 0. *Logistic regression* mencari parameter yang terbaik menggunakan klasifikasi entropi maksimum menggunakan persamaan 3. Di mana probabilitas dilambangkan dengan p dan a, b akan menjadi parameter model, a adalah faktor **(Sonjaya et al., 2022)**.

$$p = \frac{1}{1 + e^{-(a+bx)}} \quad (3)$$

Terakhir algoritma SVM yang mencari *hiperplane* yang memiliki margin terbesar antara dua kelas . Margin adalah jarak antara *hiperplane* dan titik-titik data terdekat dari masing-masing kelas (disebut vektor pendukung, *support vectors*). SVM dapat diperluas dengan menggunakan kernel. Kernel adalah fungsi yang mengubah data asli ke dalam dimensi yang lebih tinggi, di mana data tersebut lebih mudah dipisahkan, jenis fungsi kernel ditunjukkan pada Tabel 2. Ini memungkinkan SVM untuk mengatasi masalah klasifikasi yang tidak linear di dalam dimensi asli. Setelah *hiperplane* ditemukan, SVM dapat memprediksi kelas dari data uji berdasarkan posisinya terhadap *hiperplane*.

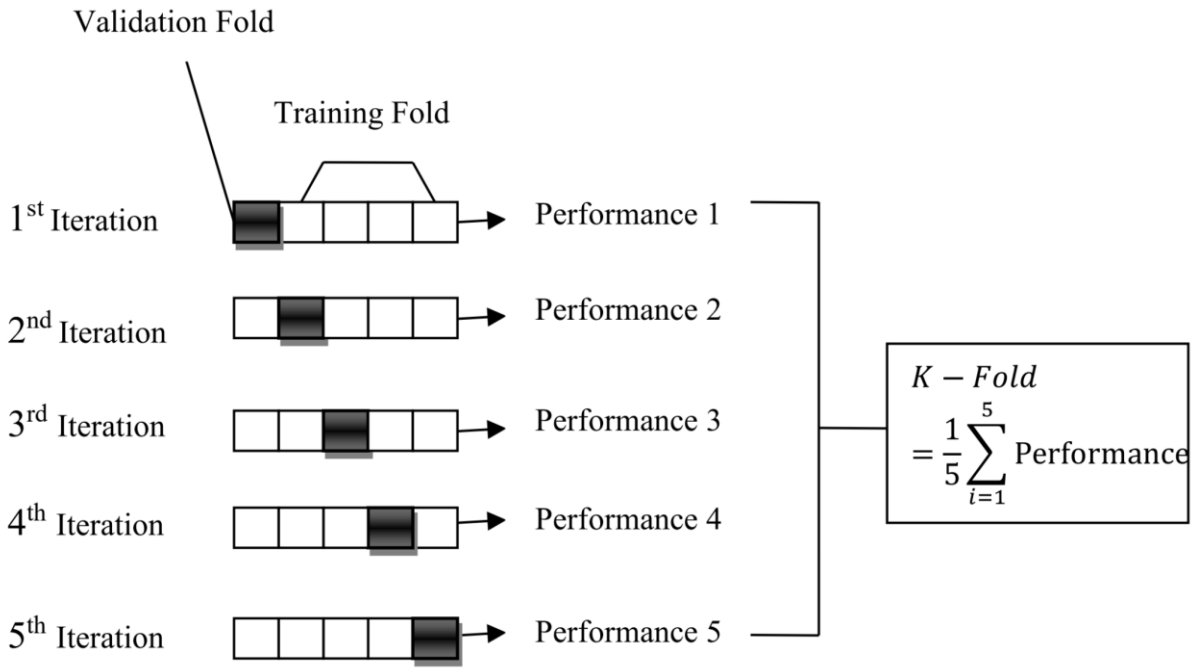
Tabel 2. Fungsi Kernel

SVM	Jenis Kernel	Persamaan
Linier	Linier	$K(x,y) = x \times y$
Non linier	*Polynomial*	$K(x,y) = (x \times y + 1)^p$
	RBF	$K(x,y) = exp\left(\frac{\|x-y\|^2}{2\sigma^2}\right)$
	Sigmoid	$K(x,y) = \tanh(cx^T y + h)$

2.3. Evaluasi

Selanjutnya, evaluasi kinerja dilakukan terhadap model dengan tujuan untuk mengetahui seberapa baik kinerja model dengan menggunakan data uji. Evaluasi didasarkan pada akurasi, presisi, sensitivitas, dan spesifisitas **(Djatna et al., 2018)(Masruriyah, Djatna, Hardhienata, et al., 2019)(Mia et al., 2022)(Nugraha et al., 2021)(Sonjaya et al., 2022)**. Akurasi digunakan untuk mengukur seberapa baik model atau sistem klasifikasi dalam melakukan klasifikasi secara benar secara keseluruhan. Kemudian, presisi mengukur seberapa baik model dalam mengidentifikasi kelas positif dengan benar dari semua prediksi yang diklasifikasikan sebagai kelas positif. Selanjutnya, sensitivitas mengukur seberapa baik model dalam mengidentifikasi semua *instance* dari kelas positif secara benar. Terakhir, spesifisitas mengukur seberapa baik model dalam mengidentifikasi semua *instance* dari kelas negatif secara benar.

Pada ML, umumnya data dibagi menjadi data latih dan data uji. Pembagian ini bertujuan untuk menguji performa dan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya. Sehingga, keandalan dan ketepatan evaluasi model, yang pada akhirnya membantu dalam membangun sistem pembelajaran mesin yang lebih baik dan lebih kuat. Evaluasi kinerja pada penelitian ini menggunakan teknik K-*Fold Validation*. Menggunakan K-Fold n_split 10 merupakan proses K-Fold 10 terbaik berdasarkan hasil riset para peneliti yang telah dilakukan peneliti dengan membandingkan hasil penelitian K-1 sampai dengan K-10 **(Djatna et al., 2018)(Zhang & Yang, 2015)**. Cara kerja dari Teknik K-*Fold Validation* adalah dengan membagi data menjadi data uji dan data latih sebanyak K. Tahapan penerapan K-*Fold* yang terlampir pada Gambar 2. Pertama, acak data dan split data menjadi K-*Fold* Kedua, pada iterasi 1 ambil sebagian lekukan menjadi data latih dan sebagian menjadi data uji, lakukan sampai semua lekukan K-*Fold* telah dilakukan. Proses tahapan kedua tersebut dilakukan sebanyak K kali hingga setiap kelompok dilakukan sebagai validasi dan tersisa sebagai data latih. Ketiga, hitung akurasi pada masing – masing lekukan. Keempat, rata – ratakan akurasi setiap iterasi dan menghasilkan akurasi terbaik.



Gambar 2. Cara Kerja K-*Fold Cross Validation*
(Mia et al., 2022)(Sonjaya et al., 2022)

3. HASIL DAN PEMBAHASAN

Hasil dari tahapan prapemrosesan data pada penelitian ini dilakukan dengan cara menghapus data dengan komponen yang tidak lengkap untuk menghindari manipulasi pengisian data lebih jauh karena data yang digunakan lebih dari 1000. Hal ini dilakukan agar data lebih ideal untuk digunakan pada tahap selanjutnya, Selanjutnya setelah data dengan komponen tidak lengkap telah dihapus maka dilakukan normalisasi pada data-data yang memiliki lebih dari 4 kategori. Data dibersihkan agar tidak terdapat data duplikat ataupun bernilai null. Data yang berjenis kategori ditransformasikan menjadi angka dengan 1 = *yes/male*, lalu 0 = *no/female*. Tujuannya agar mesin dapat melakukan pengujian, karena mesin hanya dapat membaca numerik untuk melakukan pengujian.

Pada tahapan modelling, sebelum membentuk aturan klasifikasi dan pemodelan, perlu dilakukan pembagian data menjadi dua kelompok yaitu data latih (*train model*) dan data *test* (menguji performa dari model tersebut). Berbagi data ini bertujuan untuk menganalisis apakah aturan klasifikasi yang dihasilkan oleh algoritma *Random Forest*, *Support Vector Machine*, *Logistic Regresion* dan algoritma C45, yang dapat digunakan untuk memprediksi *Heart Desease*. Dikarenakan data target penyakit jantung yang tidak seimbang, maka perlu dilakukan teknik *oversampling* dengan parameter *default*. Sehingga, data yang tidak seimbang bisa menjadi seimbang berdasarkan proses *oversampling* yang sudah dilakukan. Selanjutnya, pada tahapan ini menggunakan teknik pengujian atau perhitungan K-*Fold* terbaik menggunakan K-*Fold* 10. Pada ML akan dilakukan *split* data latih dan uji, untuk mengoptimalkannya menggunakan *Cross Validation* (K-*Fold*). Sebuah metode prosedur pengambilan sampling untuk mendapatkan eveluasi model ML dengan tujuan *Best Validation* dan *Best Learning Result*.

Partisi data K-*Fold Validation* digunakan pada pengujian karena metode ini dapat digunakan untuk mengurangi bias yang terdapat pada data random. Dalam penelitian ini digunakan 10-

fold cross validation dengan membagi *dataset* menjadi 10-*fold* yang berbeda dengan ukuran yang sama dan melakukan pelatihan dan pengujian pada model sebanyak 10 kali. Kemudian, *dataset* yang telah melewati tahapan *pre-processing* dilanjutkan dengan proses pembelajaran (*learning*) menggunakan metode klasifikasi *supervised learning* yaitu SVM, *Random Forest*, *Decision Tree*, dan *Logistic Regression*. Evaluasi didasarkan pada akurasi, presisi, sensitivitas, dan spesifisitas. Proses evaluasi menggunakan data uji yang telah dipisahkan pada proses sebelumnya dan hasil evaluasi menggunakan matriks konfusi yang ditunjukkan pada Tabel 3. Sehingga hasilnya nanti bisa dibandingkan sebagai data terbaik.

Tabel 3. Hasil Akurasi Model Klasifikasi

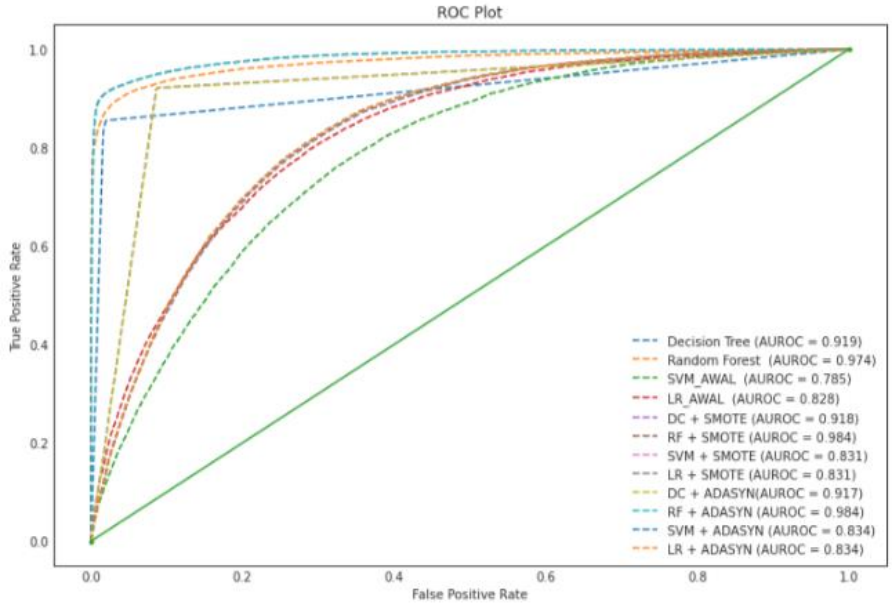
Machine Learning	**Akurasi (%)**
C45	86,65
RANDOM FOREST (RF)	90,71
SUPPORT VECTOR MACHINE (SVM)	91,68
LOGISTIC REGRESSION (LG)	91,76

Hasil penelitian menerapkan teknik *oversampling* (Tabel 4) diperoleh nilai akurasi terbaik pada perhitungan K-*Fold* 10 untuk SMOTE+RF 94,54% dan ADASYN+RF 94,49%. Sehingga, teknik gabungan RF+SMOTE dan RF + ADASYN mempunyai kinerja lebih baik dari pada C45 + ADASYN, C45 + SMOTE . Di sisi lain *Support Vector Machine* akurasi dengan SMOTE 76,24% dan ADASYN 74,47%, serta *Logistic Regression* akurasi dengan SMOTE 76,27% dan ADASYN 74,51% mengalami penurunan akurasi. Terbukti algoritma pohon keputusan *Random Forest* maupun *Decision Tree* dengan adanya *oversampling* dapat meningkatkan hasil akurasi.

Tabel 4. Hasil Akurasi Setelah Teknik *Oversampling*

Machine Learning	**Dengan SMOTE (%)**	**Dengan ADASYN (%)**
C45	91,74	91,74
RANDOM FOREST (RF)	94,54	94,49
SUPPORT VECTOR MACHINE (SVM)	76,24	74,47
LOGISTIC REGRESSION (LG)	76,27	74,51

Penurunan akurasi setelah dilakukan SMOTE dan ADASYN pada *Logistic Regression* dan SVM dikarenakan rentang nilai numerik yang cukup signifikan. Teknik *oversampling* pada penelitian ini menyebabkan penyebaran data minoritas dalam numerik yang lebih luas di ruang fitur, sehingga menyulitkan algoritma seperti *logistic regression* dan SVM untuk menemukan batas keputusan yang optimal dan akurat. Sebaliknya, peningkatan akurasi yang signifikan terjadi setelah menerapkan teknik oversampling pada algoritma C45 dan Random Forest disebabkan oleh perubahan keseimbangan data target dalam dataset. Dengan mengatasi ketidakseimbangan ini, teknik oversampling memungkinkan algoritma-algoritma ini untuk lebih efektif dan menyeluruh dalam memahami dan memodelkan setiap kelas target, menghasilkan prediksi yang lebih akurat dan dapat diandalkan. Teknik oversampling memberikan bobot yang lebih seimbang pada setiap kelas, mengurangi bias yang mungkin timbul dari ketidakseimbangan data, dan secara keseluruhan meningkatkan kinerja model dengan mengoptimalkan pembelajaran dari berbagai kelas yang ada dalam dataset.



Gambar 3. Kurva *Receiver Operating Characteristic* (ROC)

Agar dapat mengetahui kinerja ML dalam melakukan klasifikasi maka divisualisasikan menggunakan kurva *Receiver Operating Characteristic* (ROC). Berdasarkan Gambar 3, dapat dilihat bahwa algoritma RF dengan SOMTE dan dengan ADASYN memiliki nilai ROC yang sama tinggi dan paling tinggi.

KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Random Forest Classifier menonjol sebagai model terbaik dalam perbandingan algoritma, dengan tingkat akurasi awal mencapai 90,71% berdasarkan perhitungan Confusion Matrix. Namun, hasil ini dapat ditingkatkan lebih lanjut dengan penerapan teknik oversampling. Dalam konteks pengujian K-Fold 10 dan penerapan teknik oversampling, SMOTE + Random Forest ternyata menjadi salah satu pilihan terbaik dengan akurasi yang signifikan, yakni 94,54%. Kurva ROC juga menunjukkan kinerja yang sangat baik, mencapai 98,4%. Hasil ini jelas mengungguli penerapan Random Forest tanpa SMOTE dan penerapan oversampling dengan ADASYN.

Penting untuk dicatat bahwa dalam mengatasi potensi penurunan akurasi akibat oversampling, langkah-langkah berhati-hati harus diambil. Ini termasuk pemilihan teknik oversampling yang tepat, menghindari oversampling yang berlebihan, dan melibatkan validasi silang untuk mengevaluasi performa model secara lebih andal. Selain teknik oversampling, ada berbagai metode lain untuk menangani ketidakseimbangan kelas, seperti undersampling dan teknik berbasis ensemble, yang harus dievaluasi sesuai dengan karakteristik dataset dan tujuan akhir tugas pembelajaran mesin. Kesimpulannya, pemilihan dan penyesuaian metode oversampling yang tepat sangat penting dalam meningkatkan kinerja model dan hasil akhir penelitian.

DAFTAR RUJUKAN

Ath, S., Al, T., Darmawan, D., Fahmi, N., Hakim, A., Qibtiya, M. Al, & Syafei, N. S. (2022). *Jurnal Teknologi Terpadu HYBRID MACHINE LEARNING MODEL UNTUK MEMPREDIKSI PENYAKIT JANTUNG DENGAN METODE LOGISTIC REGRESSION DAN RANDOM*. *8*(1), 40–46.

Braunwald, E. (2019). Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. In *Elsivier* (Vol. 7, Issue 2).

Centers for Disease Control and Prevention. (2020). *BRFSS Survey Data and Documentation*. Centers for Disease Control and Prevention.

Cherfi, A., Nouira, 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

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

El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2022). Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors*, *22*(3). https://doi.org/10.3390/s22031184

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, *9*, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

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

Li, Y., Xu, W., Li, W., Li, A., & Liu, Z. (2021). Research on hybrid intrusion detection method based on the ADASYN and ID3 algorithms. *Mathematical Biosciences and Engineering*, *19*(2). https://doi.org/10.3934/MBE.2022095

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

Masruriyah, A. F. N., Djatna, T., Hardhienata, M. K. D., Handayani, H. H., & Wahiddin, D. (2019). *Predictive Analytics For Stroke Disease*. 13–16.

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

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Muqorobin, M., Utomo, P. B., Nafi'Uddin, M., & Kusrini, K. (2019). Implementasi Metode Certainty Factor pada Sistem Pakar Diagnosa Penyakit Ayam Berbasis Android. *Creative Information Technology Journal*, *5*(3), 185. https://doi.org/10.24076/citec.2018v5i3.198

Nugraha, R. G., Yoga Wibowo, M., Ajie, P., Handayani, H. H., Fauzi, A., & Nur Masruriyah, A. F. (2021). Implementation of Deep Learning in Order to Detect Inapposite Mask User. *2021 6th International Conference on Informatics and Computing, ICIC 2021*, 4–9. https://doi.org/10.1109/ICIC54025.2021.9632994

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*.

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

Qing, Z., Zeng, Q., Wang, H., Liu, Y., Xiong, T., & Zhang, S. (2022). ADASYN-LOF Algorithm for Imbalanced Tornado Samples. *Atmosphere*, *13*(4). https://doi.org/10.3390/atmos13040544

Raja, R., Nagwanshi, K. K., Kumar, S., & Laxmi, K. R. (2022). *Data Mining and Machine Learning Applications*.

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

Rohman, A., & Rochcham, D. M. (2018). MODEL ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT JANTUNG. In *Jurnal Neo Teknika* (Vol. 4, Issue 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

Siringoringo, R. (2018). *KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN ALGORITMA SMOTE DAN k-NEAREST NEIGHBOR* (Vol. 3, Issue 1).

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

Zaki, M. J., & Wagner Jr, M. (2020). *Data Mining and Machine Learning Fundamental Concepts and Algorithms*.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006