Jurnal ICT : Information Communication & Technology
Vol. 26, N0.1, Juli 2026, pp. 159-167
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI: https://doi.org/10.36054/jict-ikmi.v26i1.360

Prediksi Risiko Serangan Jantung Menggunakan Machine Learning: Pendekatan SMOTE dan Evaluasi Skenario Seleksi Fitur

Wahyu Nugraha[1], Muhamad Syarif[2*]
[1]Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia
Email: [1]wahyu.whn@bsi.ac.id, [2]muhamad.mdx@bsi.ac.id

INFORMASI ARTIKEL

Histori artikel:

Naskah masuk, 18 Mei 2026
Direvisi, 18 Juli 2026
Diiterima, 31 Juli 2026

Kata Kunci:

Machine Learning,
Heart Attack Prediction,
SMOTE,
Feature Selection,
Accuracy Paradox

ABSTRAK

Abstract- *Cardiovascular disease remains one of the leading causes of mortality worldwide. The application of Machine Learning* (ML) *provides a promising approach for the early detection of heart attack risk, yet the predictive modeling process is frequently hindered by imbalanced data and the presence of irrelevant features. This study evaluates the performance of eight classification algorithms in predicting heart attack risk using a dataset of* 8,763 *patient records. To address the bias caused by the majority class dominance* (64.1% *healthy patients*), *the Synthetic Minority Over-sampling Technique* (SMOTE) *was applied to the training dataset. The evaluation was conducted comparatively across three data dimensionality scenarios, that is Baseline Data (all features), the top five features derived from Random Forest Feature Importance, and a subset of five features selected via Recursive Feature Elimination* (RFE). *The results demonstrated that the Gradient Boosting Machine* (GBM) *algorithm achieved the highest functional accuracy of* 60.4% *under the Baseline scenario. The most critical finding of this experiment was the empirical evidence of the Accuracy Paradox in the* RFE *feature scenario. Although linear models such as Logistic Regression and probabilistic models like Naive Bayes achieved the highest "accuracy" at* 64.3%, *further evaluation revealed that their Precision and Recall scores for detecting heart attack risk were exactly* 0.00. *These models completely failed to identify the positive class, merely predicting the majority class for all test samples. In conclusion, demographic and lifestyle features independently possess only moderate predictive power* (~60%), *and relying solely on Accuracy as an evaluation metric in clinical predictive modeling can lead to highly misleading conclusions. As a practical implication, these findings provide a critical evaluation guideline for the development of clinical decision support systems to avoid relying solely on lifestyle attributes and to prioritize Recall-based metrics to prevent the failure of detecting at-risk patients.*

Abstrak- Penyakit kardiovaskular merupakan salah satu penyebab kematian tertinggi di dunia. Pemanfaatan *Machine Learning* (ML) menawarkan solusi prediksi risiko serangan jantung secara dini, namun pemodelannya sering terhambat oleh masalah ketidakseimbangan kelas (*imbalanced data*) dan fitur yang kurang relevan. Penelitian ini mengevaluasi kinerja delapan algoritma klasifikasi dalam memprediksi risiko serangan jantung berdasarkan 8.763 rekam medis pasien. Untuk mengatasi bias akibat dominasi kelas mayoritas (64,1% pasien sehat), menerapkan metode *oversampling Synthetic Minority Over-sampling Technique* (SMOTE) pada data latih. Eksperimen evaluasi dilakukan secara komparatif pada tiga skenario dimensi data, yaitu *Baseline Data* (seluruh fitur), lima fitur teratas berdasarkan *Random Forest Feature Importance*, dan subset lima fitur hasil *Recursive Feature Elimination* (RFE). Hasil pengujian menunjukkan bahwa algoritma *Gradient Boosting Machine* (GBM) pada skenario *Baseline* mencapai akurasi fungsional tertinggi sebesar 60,4%. Temuan paling kritis dari eksperimen ini adalah terbuktinya anomali Paradoks Akurasi (*Accuracy Paradox*) pada skenario penggunaan fitur RFE. Meskipun model linier seperti *Logistic Regression* dan algoritma probabilistik seperti Naive

Bayes mencatatkan angka "akurasi" tertinggi mencapai 64,3%, evaluasi lanjutan menemukan bahwa nilai *Precision* dan *Recall* model tersebut untuk mendeteksi risiko serangan jantung adalah 0,00. Model terbukti gagal total mengidentifikasi kelas positif dan sekadar menebak seluruh sampel uji ke dalam probabilitas kelas mayoritas. Kesimpulannya, fitur demografis dan gaya hidup secara mandiri memiliki daya prediktabilitas yang moderat (~60%), dan penggunaan metrik *Accuracy* secara tunggal untuk evaluasi pemodelan klinis dapat menghasilkan kesimpulan yang sangat menyesatkan. Secara praktis, temuan ini memberikan panduan evaluasi penting bagi pengembangan sistem pendukung keputusan klinis agar tidak hanya bergantung pada atribut gaya hidup serta memprioritaskan evaluasi berbasis metrik Recall guna menghindari kegagalan deteksi pasien berisiko.

***Penulis Korespondensi**:*

Muhamad Syarif
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Universitas Bina Sarana Informatika
Jl. Abdul Rahman Saleh No.18, Kec. Pontianak Tenggara, Kota Pontianak, Kalimantan Barat - Indonesia
Email: muhamad.mdx@bsi.ac.id

1. Pendahuluan

Penyakit kardiovaskular, khususnya serangan jantung, tetap menjadi salah satu penyebab utama mortalitas di seluruh dunia [1][2]. Seiring dengan pesatnya digitalisasi rekam medis, pemanfaatan algoritma *Machine Learning* (ML) telah menjadi pendekatan proaktif yang sangat menjanjikan untuk mendeteksi dini risiko serangan jantung [3]. Model ML dapat memanfaatkan berbagai rekam jejak pasien mulai dari usia, tingkat kolesterol, tekanan darah, hingga faktor gaya hidup seperti kebiasaan merokok dan jam duduk untuk memprediksi probabilitas terjadinya serangan jantung [4]. Meskipun potensinya sangat besar, penerapan algoritma prediktif pada data klinis masih sering dihadapkan pada berbagai tantangan inheren yang apabila tidak dievaluasi secara komprehensif, dapat menghasilkan kesimpulan yang menyesatkan [5].

Tantangan utama pertama adalah fenomena ketidakseimbangan kelas (*imbalanced data*). Dalam praktiknya, jumlah rekam medis pasien yang sehat sering kali jauh lebih besar dibandingkan pasien yang memiliki risiko serangan jantung positif [6]. Sebagai contoh, pada dataset penelitian ini yang terdiri dari 8.763 observasi, distribusi kelas sangat timpang dengan rasio pasien berisiko hanya sebesar 35,8% berbanding 64,1% pasien sehat. Model ML yang dilatih pada data yang tidak seimbang cenderung menghasilkan bias predisposisi terhadap kelas mayoritas, sehingga metrik evaluasi seperti *Accuracy* tidak lagi mampu merepresentasikan keandalan sesungguhnya dalam mengklasifikasikan kelas minoritas yang justru merupakan fokus utama medis [7].

Tantangan kedua berkaitan dengan tingginya dimensi data (*dimensionality*) dan keberadaan fitur yang mengalami *noise* (*noisy features*). Tidak semua atribut pasien memberikan kontribusi signifikan terhadap prediksi risiko. Oleh karena itu, penerapan metode seleksi fitur (*Feature Selection*) seperti *Random Forest Feature Importance* dan *Recursive Feature Elimination* (RFE) sering digunakan untuk mereduksi dimensi data, meningkatkan efisiensi komputasi, dan menghindari *overfitting* [8][9]. Namun, reduksi fitur tanpa metode evaluasi yang komprehensif berpotensi menciptakan Paradoks Akurasi (*Accuracy Paradox*), di mana model seolah-olah menunjukkan tingkat akurasi yang tinggi secara persentase, namun terbukti gagal total secara klinis (nilai *Recall* 0.0) dalam memprediksi kelas positif [10][11].

Berdasarkan latar belakang tersebut, penelitian dilakukan dengan tujuan untuk mengevaluasi dan membandingkan performa prediktif dari delapan algoritma klasifikasi *machine learning*, yaitu *Logistic Regression, K-Nearest Neighbors* (KNN), *Decision Tree, Naive Bayes, Support Vector Machine* (SVM), *Gradient Boosting Machine* (GBM), XGBoost, dan *Multi-Layer Perceptron* (MLP). Untuk mengatasi masalah bias kelas mayoritas, penelitian ini mengimplementasikan *oversampling Synthetic Minority Over-sampling Technique* (SMOTE) untuk menyeimbangkan kelas pada tahap pelatihan [12]. SMOTE bekerja dengan cara menghasilkan sampel sintetis untuk kelas minoritas guna menyeimbangkan distribusi data pada tahap pelatihan, yang terbukti efektif dalam meningkatkan performa prediksi pada berbagai dataset medis yang timpang [13].

Lebih lanjut, penelitian ini juga akan menguji dampak tiga skenario penggunaan dimensi data yakni *Baseline Data* (keseluruhan fitur), *Top* 5 *Feature Importance*, dan data seleksi fitur berbasis RFE terhadap kinerja tiap algoritma. Kontribusi utama dari penelitian ini tidak hanya terletak pada penemuan algoritma terbaik untuk dataset risiko serangan jantung, tetapi juga pada analisis kritis secara empiris mengenai bahaya Paradoks Akurasi pada evaluasi model klasifikasi medis, sehingga memberikan panduan praktis bagi penelitian informatika medis di masa depan.

2. Tinjauan Pustaka

2.1. *Machine Learning*

Machine Learning atau Pembelajaran Mesin adalah teknik pendekatan dari *Artificial Intelligent* yang digunakan untuk meniru dan menggantikan peran manusia dalam melakukan aktivitas serta memecahkan masalah [14]. *Machine Learning* mampu meramalkan dan memahami karakteristik objek yang tidak dikenal melalui pengidentifikasian pola dalam dataset [15].

2.2. *Feature Selection*

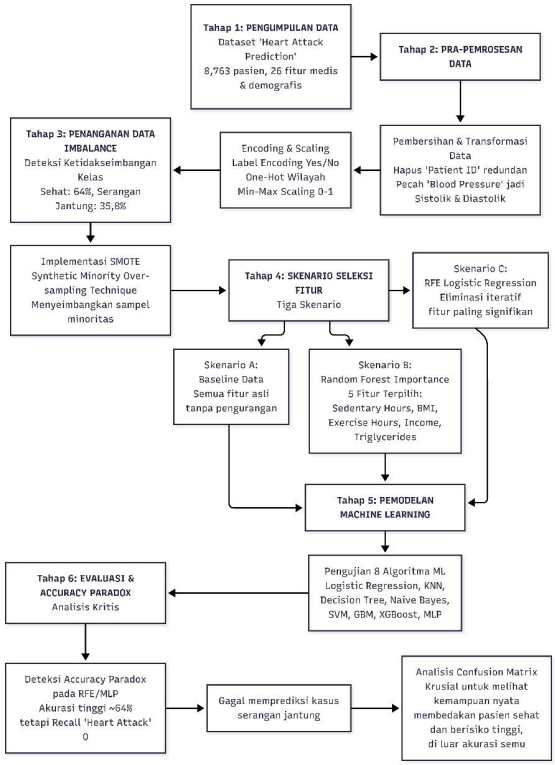
Seleksi fitur merupakan istilah yang digunakan dalam *machine learning* dan statistik untuk melakukan pemilihan subset fitur yang relevan sebelum pembangunan model [16]. Seleksi fitur merupakan salah satu metode dalam tahap *pre-processing data mining*. Seleksi fitur bertujuan untuk mengurangi kompleksitas atribut yang diolah selama proses analisis, dan mengidentifikasi fitur yang paling penting dalam kolom yang terdapat pada dataset [17].

2.3. *Synthetic Minority Over-sampling Technique* (SMOTE)

SMOTE adalah teknik di mana *oversampling* kelas minoritas dilakukan dengan menghasilkan sampel sintetik baru berdasarkan kelas minoritas [18]. SMOTE adalah metode *oversampling* dimana data pada kelas minoritas diperbanyak dengan menggunakan data sintetik yang berasal dari kelas minoritas. *Oversampling* pada SMOTE mengambil *instance* dari kelas minoritas lalu mencari *k-nearest neighbor* dari setiap *instance*, dan menghasilkan *instance* sintetik dari hasil mereplikasi *instance* kelas minoritas [19].

3. Metode Penelitian

Bagian ini menguraikan tahapan sistematis yang dilakukan dalam penelitian, mulai dari pengumpulan data, pra-pemrosesan, penyeimbangan kelas, hingga evaluasi model klasifikasi. Pendekatan yang digunakan mengadopsi standar penambangan data yang komprehensif untuk memastikan validitas hasil eksperimen. Gambar 1 merupakan visualisasikan tahapan penelitian.



Gambar 1. Tahapan Penelitian

3.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder bersumber dari *Kaggle* berupa *Heart Attack Prediction* Dataset. Dataset ini terdiri dari 8.763 observasi (baris) dan 26 atribut awal yang mencakup profil demografis pasien (usia, jenis kelamin), metrik klinis (kolesterol, detak jantung, riwayat diabetes), dan gaya hidup (jam duduk, jam olahraga, pola makan). Variabel target pada penelitian ini adalah *Heart Attack Risk* yang memiliki dua kelas biner, yaitu kelas 0 untuk pasien sehat dan kelas 1 untuk pasien yang memiliki risiko serangan jantung. Pada tahap eksplorasi awal, tidak ditemukan adanya nilai yang hilang (*missing values*) maupun baris data yang terduplikasi dalam dataset.

3.2. Pra-pemrosesan dan Rekayasa Fitur (*Feature Engineering*)

Untuk mempersiapkan data sebelum tahap pemodelan, serangkaian teknik pra-pemrosesan diterapkan:

a. Penghapusan Fitur
   Atribut Patient ID dan *Country* dihapus dari dataset karena dianggap tidak memberikan nilai prediktif secara klinis dan untuk menghindari redundansi.
b. Rekayasa Fitur Tekanan Darah
   Atribut Blood Pressure yang pada awalnya memiliki format teks (misalnya 158/88)

diekstraksi menjadi dua fitur numerik terpisah (*Systolic* dan *Diastolic*). Selanjutnya, dibuat atribut turunan baru bernama BP_Ratio yang dihitung dengan membagi nilai *Systolic* terhadap *Diastolic*. Setelah ekstraksi ini, kolom *Blood Pressure* yang asli dihapus.

3.3. Transformasi dan Penskalaan

a. *Encoding* Variabel Kategorikal
   Fitur biner dan ordinal dikonversi menjadi nilai numberik. Atribut *Sex* dipetakan menjadi *Female* (0) dan *Male* (1), atribut Diet dipetakan menjadi *Unhealthy* (0), *Average* (1), dan *Healthy* (2). Atribut *Continent* dan *Hemisphere* yang merupakan variabel kategorikal nominal maka menerapkan metode *One-Hot Encoding,* tujuannya adalah untuk merepresentasikan atribut dalam bentuk matriks biner, dengan menghapus kolom pertama (*drop_first = True*) untuk menghindari masalah multikolinearitas.
b. Penskalaan Min-Max (MinMaxScaler)
   Transformasi ini menormalisasi seluruh nilai pada fitur numerik sehingga berada pada rentang yang seragam antara 0 dan 1.

3.4. Penanganan Ketidakseimbangan Data (*Data Imbalance*)

Berdasarkan distribusi kelas pada data latih, teridentifikasi masalah ketidakseimbangan data di mana proporsi pasien sehat (kelas 0) berjumlah 3.933 sampel (64,1%) berbanding pasien berisiko (kelas 1) berjumlah 2.201 sampel (35,8%). Untuk mengatasi agar model tidak bias terhadap kelas mayoritas, metode SMOTE diaplikasikan pada data latih. SMOTE berhasil menyeimbangkan distribusi data latih menjadi 3.933 sampel untuk masing-masing kelas. Tabel 1 menunjukkan distribusi kelas sebelum dan sesudah proses *oversampling* dari tiga skenario data *training.*

Tabel 1. Distribusi Kelas Sebelum dan Sesudah SMOTE

	Kelas Target	Sebelum SMOTE	Sesudah SMOTE
Baseline Data	Sehat (0.0)	3.933	3.933
	Berisiko (1.0)	2.201	3.933
Feature Importance Data	Sehat (0.0)	3.933	3.933
	Berisiko (1.0)	2.201	3.933
RFE Data	Sehat (0.0)	3.933	3.933
	Berisiko (1.0)	2.201	3.933

3.5. Skenario Seleksi Fitur

Pembagian data dilakukan dengan proporsi 70% data latih dan 30% data uji (*test_size=0.3*) menggunakan *random state* sebesar 42. Evaluasi algoritma diuji pada tiga skenario dimensi data yang berbeda, yaitu:

a. Skenario 1 (*Baseline Data*)
   Menggunakan seluruh atribut yang telah melewati proses pembersihan dan penskalaan tanpa reduksi fitur.
b. Skenario 2 (*Feature Importance Data*)
   Mengekstraksi 5 atribut terpenting berdasarkan pembobotan algoritma *Random Forest Regressor*. Kelima fitur terpilih adalah *Sedentary Hours Per Day*, *Body Mass Index* (BMI), *Exercise Hours Per Week*, *Income*, dan *Triglycerides*.
c. Skenario 3 (*Recursive Feature Elimination*)
   Menggunakan teknik RFE dengan algoritma dasar *Logistic Regression* untuk mengambil 5 subset fitur terbaik secara rekursif. Atribut yang diseleksi oleh pendekatan linier ini meliputi *Diabetes*, *Continent_Asia*, *Continent_Europe*, *Hemisphere_Southern*, dan BP_Ratio.

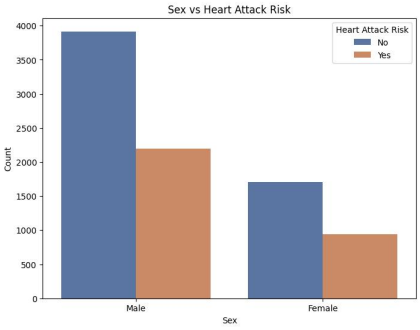
3.6. Algoritma Klasifikasi dan Evaluasi Model

Pengujian eksperimen menggunakan delapan algoritma *machine learning*: *Logistic Regression, K-Nearest Neighbors* (KNN)*, Decision Tree, Naive Bayes, Support Vector Machine* (SVC)*, Gradient Boosting Machine* (GBM)*, XGBoost*, dan *Multi-Layer Perceptron* (MLP). Untuk memastikan metodologi yang komprehensif, mencegah terjadinya overfitting, dan mendukung replikasi penelitian, eksperimen ini menerapkan prosedur validasi model menggunakan *K-Fold Cross-Validation* (dengan k=5) yang dikombinasikan dengan pencarian *GridSearchCV* pada data latih guna menemukan konfigurasi parameter (*hyperparameter*) yang optimal. Pemilihan rentang parameter didasarkan pada karakteristik dataset medis yang kompleks serta menjaga efisiensi komputasi. Setelah model divalidasi dan diimplementasikan dengan parameter optimal tersebut, kinerja setiap model diukur menggunakan matriks konfusi (*Confusion Matrix*) dan metrik klasifikasi yang difokuskan pada skor *Accuracy, Precision, Recall*, serta *F1-Score*.

4. Hasil dan Pembahasan

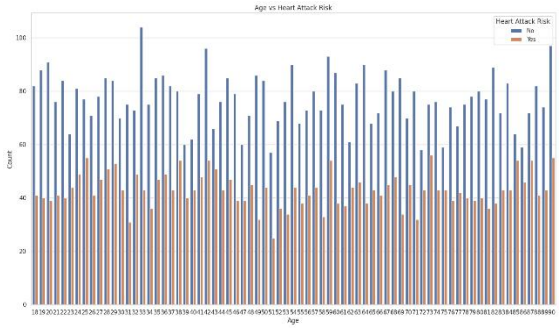
4.1. Analisis Data Eksploratif (EDA)

Analisis Data Eksploratif (EDA) diterapkan sebelum tahap pemodelan guna memahami karakteristik dan distribusi risiko serangan jantung di dalam dataset.



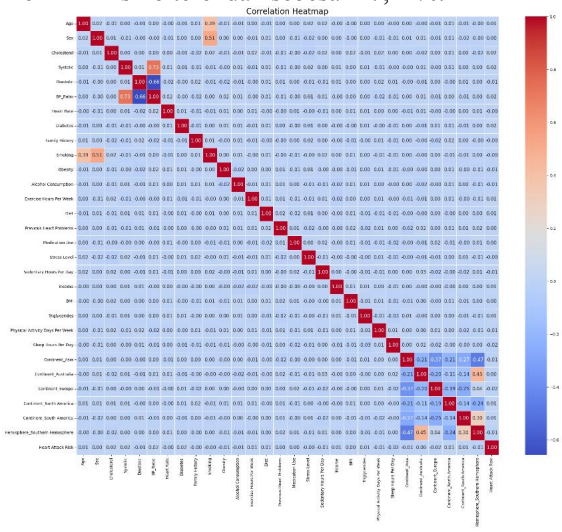
Gambar 2. *Sex vs Heart Attack Risk*

Hasil analisis pada gambar 2, menunjukan persentase risiko serangan jantung berdasarkan jenis kelamin ditemukan relatif seimbang antara pria (*male*) dan wanita (*female*), di mana 35,9% pasien pria memiliki risiko positif dibandingkan dengan 35,6% pada wanita.



Gambar 3. *Age vs Heart Attack Risk*

Hasil analisis pada gambar 3, menunjukan persentase risiko serangan jantung pada kelompok usia 85 tahun memiliki risiko tertinggi yaitu sebesar 45,76%, sementara kelompok usia 49 tahun memiliki risiko terendah sebesar 27,12%.



Gambar 4. *Correlation Heatmap*

Hasil evaluasi matriks korelasi pada Gambar 4 *Correlation Heatmap* mengungkapkan temuan penting terkait hubungan antar variabel. Tidak ada satupun fitur klinis maupun demografis yang memiliki korelasi linier tunggal yang kuat terhadap target probabilitas Risiko Serangan Jantung. Fitur dengan korelasi positif maupun negatif tertinggi seperti *Cholesterol* (0,019), *Systolic* (0,018), BP_Ratio (0,017), serta *Sleep Hours Per Day* (-0,018) dan letak *Continent_Europe* (-0,015) berada pada rentang koefisien yang amat rendah mendekati batas nol. Lemahnya korelasi ini menegaskan bahwa risiko serangan jantung merupakan kondisi kompleks yang dipengaruhi oleh interaksi multivariat yang non-linier, sekaligus menjelaskan alasan di balik kegagalan algoritma berbasis linier dalam mengklasifikasikan kelas minoritas pada skenario RFE. Meskipun demikian, matriks tersebut secara akurat memvalidasi korelasi logis antar-fitur independen, terbukti dari korelasi matematis yang kuat antara fitur turunan BP_Ratio dengan metrik *Systolic* (0,726) dan *Diastolic* (-0,656), serta adanya korelasi positif antara kebiasaan merokok (*Smoking*) dengan jenis kelamin (*Sex*) sebesar 0,514 dan faktor usia (*Age*) sebesar 0,394.

4.2. Kinerja Model pada Data Dasar (*Baseline*)

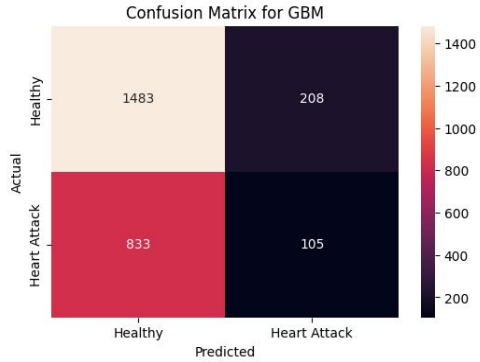
Pengujian tahap pertama dilakukan menggunakan seluruh fitur hasil pra-pemrosesan (*baseline data*). Evaluasi menggunakan data uji (setelah SMOTE) menunjukkan bahwa algoritma *Gradient Boosting Machine* (GBM) memberikan performa prediksi terbaik dengan tingkat akurasi mencapai 60,4% seperti yang terlihat pada Gambar 5. Algoritma terbaik kedua adalah XGBoost dengan akurasi 58,3%. Sebaliknya, model berbasis linier dan jarak seperti *Logistic Regression* (49,7%) dan KNN (47,3%) memiliki kinerja terendah.

Keunggulan pada algoritma GBM dan XGBoost mengindikasikan bahwa data rekam medis pasien memiliki karakteristik dan interaksi fitur yang kompleks dan non-linier, lebih mampu ditangani secara efektif oleh algoritma berbasis *ensemble tree*. Untuk memvalidasi signifikansi perbedaan performa ini, dilakukan pengujian statistik menggunakan Analisis Varians *(Analysis of Variance*/ANOVA) satu arah terhadap metrik performa hasil *k-fold cross-validation*. Hasil uji statistik menunjukkan nilai *p-value* < 0,05, yang secara empiris mengonfirmasi bahwa keunggulan GBM dan XGBoost atas model linier dan probabilistik memiliki perbedaan yang signifikan secara statistik. Temuan kinerja superior dari model ensemble pada studi ini sejalan dengan penelitian terdahulu juga menyimpulkan bahwa algoritma seperti *Gradient Boosting* secara konsisten mengungguli model tunggal konvensional dalam menavigasi dataset kardiovaskular berdimensi kompleks.. Keseluruhan hasil kinerja model prediksi pada skenario *baseline* dapat dilihat pada Gambar 6.

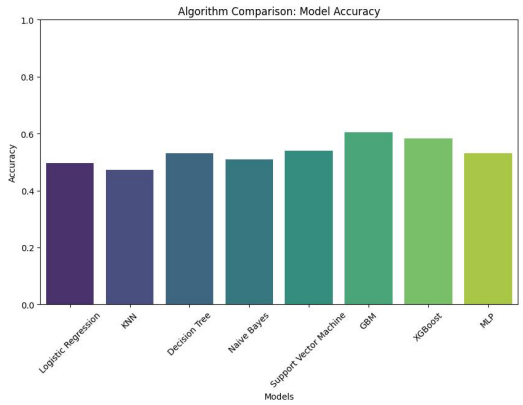
```
GBM: Accuracy: 0.604
Classification Report for GBM:
              precision    recall  f1-score   support

         0.0       0.64      0.88      0.74      1691
         1.0       0.34      0.11      0.17       938

    accuracy                           0.60      2629
   macro avg       0.49      0.49      0.45      2629
weighted avg       0.53      0.60      0.54      2629
```



Gambar 5. Hasil Prediksi GBM (*Baseline*)



Gambar 6. Hasil Kinerja Seluruh Model (*Baseline*)

4.3. Dampak Seleksi Fitur (*Feature Importance*) Terhadap Kinerja Model

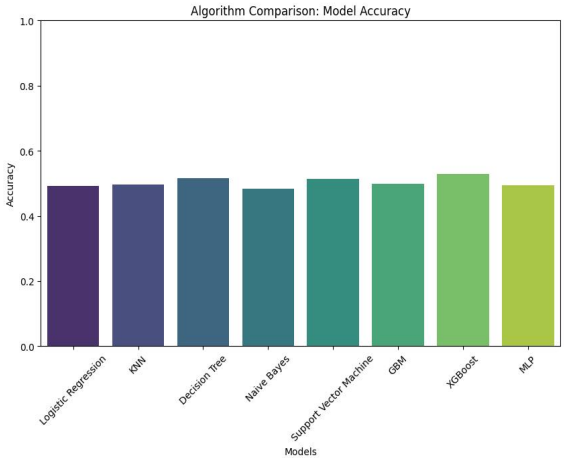
Pada skenario kedua, dimensi data direduksi dan difokuskan hanya pada 5 fitur teratas hasil seleksi metode *Random Forest Feature Importance* yang dapat dilihat pada Gambar 7.

11 rows × 8763 rows × 6 cols

	Sedentary Hours Per…	BMI	Exercise Hours …	Income	Triglycerides	Heart A…
0	6.615001	31.251233	4.168189	261404	286	0
1	4.963459	27.194973	1.813242	285768	235	0
2	9.463426	28.176571	2.078353	235282	587	0
3	7.648981	36.464704	9.828130	125640	378	0
4	1.514821	21.809144	5.804299	160555	231	0
...	...	...	...	...	...	...
8758	10.806373	19.655895	7.917342	235420	67	0
8759	3.833038	23.993866	16.558426	217881	617	0
8760	2.375214	35.406146	3.148438	36998	527	1
8761	0.029104	27.294020	3.789950	209943	114	0
8762	9.005234	32.914151	18.081748	247338	180	1

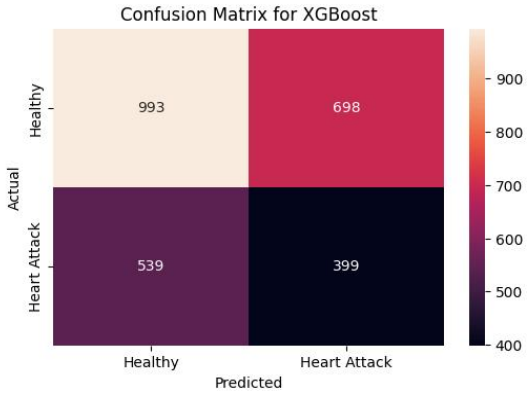
Gambar 7. Fitur Terpilih Melalui Metode *Random Forest Feature Importance*

Pengurangan fitur ternyata menurunkan kinerja secara keseluruhan pada hampir semua model algoritma. Penurunan skor ini menyiratkan bahwa kelima metrik gaya hidup tersebut secara terisolasi tidak memiliki daya diskriminatif yang cukup kuat untuk memprediksi serangan jantung, sehingga pemangkasan fitur lain justru menghilangkan informasi (*varians*) penting dari data. Anomali performa akibat reduksi dimensi ini membedakan hasil studi ini dengan teori umum seleksi fitur seperti yang dikemukakan oleh Guyon dan Elisseeff, di mana penyusutan fitur idealnya mempertahankan atau bahkan meningkatkan metrik klasifikasi. Pengujian statistik lanjutan (*paired t-test*) terhadap penurunan akurasi model sebelum dan sesudah reduksi fitur juga menunjukkan nilai signifikan (p-*value* < 0,05), yang menegaskan bahwa seluruh variabel fisiologis dan demografis pasien sebenarnya memiliki bobot laten yang terikat secara kolektif. Hasil kinerja kesuluruhan model prediksi pada skenario *feature importance* dapat dilihat pada Gambar 8.



Gambar 8. Kinerja Seluruh Model (*Feature Importance*)

Performa tertinggi pada skenario ini dicapai oleh *XGBoost* dengan akurasi 52,9% seperti pada Gambar 9.



Gambar 9. Hasil Prediksi *XGBoost* (*Feature Importance*)

4.4. Temuan Kritis (Paradox Akurasi pada Evaluasi RFE)

Skenario pengujian ketiga, menggunakan subset 5 fitur geografis dan medis diskrit hasil seleksi linier RFE. Hasil seleksi dapat dilihat pada Gambar 10.

11 rows × 8763 rows × 6 cols

	BP_Ratio	Diabetes	Continent_As…	Continent_E…	Hemisphere_So…	Heart Atta…
0	1.80	0	0	0	1	0
1	1.77	1	0	0	0	0
2	1.76	1	0	1	0	0
3	1.63	1	0	0	0	0
4	1.03	1	1	0	0	0
...	...	...	...	...	...	...
8758	1.24	1	1	0	0	0
8759	1.54	1	0	0	0	0
8760	2.15	0	0	0	1	1
8761	1.78	1	0	0	1	0
8762	2.06	1	0	1	0	1

Gambar 10. Fitur Terpilih Melalui Seleksi Linier RFE

Hasil evaluasi metrik tunggal menunjukkan peningkatan akurasi yang signifikan, di mana model *Logistic Regression*, *Naive Bayes*, *Support Vector Machine* (SVM), dan *Multi-Layer Perceptron* (MLP) mencapai akurasi tertinggi di angka 64,3%, dapat dilihat pada Gambar 11 berikut.

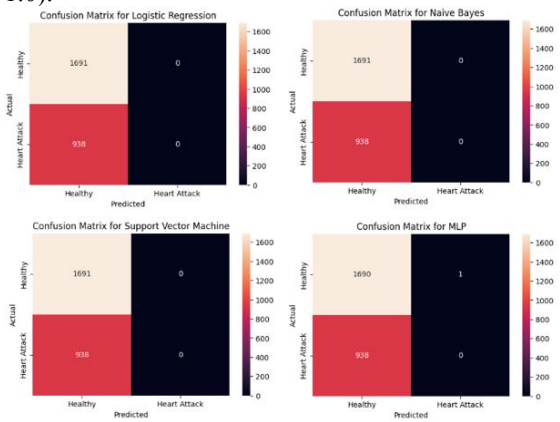
```
Logistic Regression: Accuracy: 0.643
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.64      1.00      0.78      1691
         1.0       0.00      0.00      0.00       938

    accuracy                           0.64      2629
   macro avg       0.32      0.50      0.39      2629
weighted avg       0.41      0.64      0.50      2629
```

Gambar 11. Hasil Prediksi *Logistic Regression* (RFE)

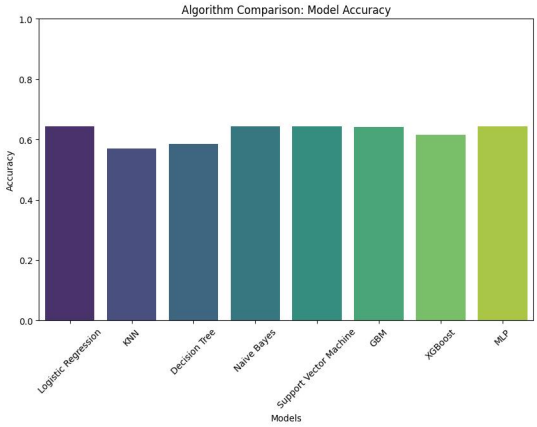
Namun, hasil pengujian lebih mendalam pada metrik *Classification Report* mengungkapkan adanya kegagalan prediktif yang disebut sebagai Paradoks Akurasi (*Accuracy Paradox*). Diketahui bahwa data uji terdiri dari 2.629 sampel observasi (1.691 kelas sehat dan 938 kelas berisiko). Model-model dengan akurasi 64,3% tersebut memiliki nilai presisi, *recall*, dan *F1-Score* sebesar 0.00 untuk memprediksi pasien yang sakit (kelas positif 1.0).



Gambar 12. Model Gagal Melakukan Prediksi (Kelas Positif 1.0)

Analisa dari Gambar 12 model-model tersebut sekadar memprediksi seluruh sampel ke kelas mayoritas (Sehat/0.0), menghasilkan nilai akurasi secara teoretis tetapi gagal total dalam mendeteksi satu pun pasien dengan risiko serangan jantung sesungguhnya. Kegagalan ini menunjukkan bahwa subset fitur yang dipilih oleh RFE dengan estimator regresi logistik linier menyebabkan model terjebak dalam probabilitas kelas mayoritas meskipun data latih telah diseimbangkan menggunakan SMOTE. Hasil tersebut menjadi bukti empiris yang kuat bahwa mengandalkan metrik *Accuracy* semata di ranah diagnostik medis dapat menyebabkan kesimpulan yang sangat menyesatkan, di mana model yang tampak akurat sama sekali tidak memiliki nilai klinis (*recall* = 0) [10], [11].

Keseluruhan hasil kinerja model prediksi pada skenario RFE dapat dilihat pada Gambar 13.



Gambar 13. Hasil Kinerja Seluruh Model (RFE)

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan serangkaian eksperimen dan analisis komparatif yang telah dilakukan terhadap delapan algoritma *machine learning* dan tiga skenario pemilihan fitur untuk prediksi risiko serangan jantung, dapat ditarik beberapa kesimpulan utama:

a. Keterbatasan Prediktif Data Gaya Hidup
   Pengujian pada skenario *Baseline Data* menunjukkan bahwa algoritma *Gradient Boosting Machine* (GBM) menghasilkan akurasi tertinggi sebesar 60,4%. Batas performa yang berkisar di angka 60% ini mengindikasikan bahwa penggunaan atribut demografis dan gaya hidup saja secara mandiri memiliki daya prediktif yang terbatas untuk mendeteksi serangan jantung tanpa didampingi data uji klinis mendalam.
b. Dampak Reduksi Fitur Berbasis Pohon Keputusan
   Skenario *Feature Importance* yang mereduksi dimensi menjadi lima fitur teratas justru menurunkan performa prediktif seluruh model. Hal ini mengonfirmasi bahwa prediksi kondisi medis yang kompleks memerlukan interaksi dari berbagai fitur secara holistik, bukan sekadar variabel terisolasi.
c. Bahaya Evaluasi Tunggal dan Paradoks Akurasi

Temuan paling kritis dalam penelitian ini terjadi pada skenario subset fitur *Recursive Feature Elimination* (RFE). Evaluasi dengan akurasi tunggal memperlihatkan angka tertinggi (64,3%) pada model linier dan probabilistik. Namun, analisis lanjutan melalui matriks klasifikasi membuktikan terjadinya Paradoks Akurasi, di mana model sama sekali tidak mampu mendeteksi kelas minoritas positif pasien berisiko (*Recall* dan *Precision* bernilai 0.00). Hal ini membuktikan secara empiris bahwa mengandalkan metrik *Accuracy* sebagai tolok ukur tunggal pada data medis sangat rentan menghasilkan bias klinis yang fatal.

d. Keterbatasan Dataset dan Generalisasi Model
Terlepas dari temuan komparatif yang diperoleh, penelitian ini memiliki keterbatasan eksplisit pada penggunaan dataset sekunder yang sangat bertumpu pada indikator demografis dan gaya hidup, tanpa menyertakan variabel rekam medis klinis yang lebih mendalam (seperti hasil EKG atau parameter biokimia spesifik). Keterbatasan fitur ini menyebabkan batas fungsional model tertahan di angka ~60%, sehingga kemampuan generalisasi model jika diimplementasikan pada lingkungan klinis dunia nyata yang lebih kompleks masih terbatas dan harus diinterpretasikan secara hati-hati. Kelemahan pada kelengkapan dataset ini sekaligus menegaskan ruang lingkup bagi pengembangan penelitian selanjutnya.

5.2. Saran

Berdasarkan dari temuan dalam eksperimen ini, beberapa rekomendasi yang dapat diajukan untuk penelitian selanjutnya adalah:

a. Peningkatan Kualitas Data Klinis
Disarankan untuk tidak hanya bertumpu pada wawancara gaya hidup pasien, tetapi mengintegrasikan variabel medis yang definitif seperti hasil EKG atau level enzim jantung spesifik untuk meningkatkan performa melampaui batas 60%.

b. Fokus pada Metrik *Recall*
Peneliti selanjutnya disarankan mengganti metrik pengoptimalan utama dari *Accuracy* menjadi *Recall* atau *F1-Score.* Tujuannya adalah meminimalisasi *False Negatives* (pasien berisiko yang salah didiagnosis sehat). Eksplorasi dengan *Cost-Sensitive Learning* patut dipertimbangkan selain metode SMOTE.

c. Metode Seleksi Fitur Non-Linier
Disarankan bereksperimen dengan teknik reduksi dimensi non-linier seperti *Kernel* PCA atau algoritma *Wrapper* berbasis *Ensemble* untuk mempertahankan varians informasi prediktif yang seringkali bersifat non-linier.

Daftar Pustaka

[1] "Cardiovascular diseases (CVDs)". Accessed: May 11 2026. [*Online*]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study," *Journal of the American College of Cardiology.*, vol. 76, no. 25, pp. 2982-3021, 2020.

[3] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine (Poland)*, vol. 17, no. 1, pp. 1100-1113, 2022.

[4] A. R. Ilyas, S. Javaid, and I. L. Kharisma, "Heart Disease Prediction Using ML," *Engineering Proceedings*, vol. 107, no. 1, 2025.

[5] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, p. 160035, 2016.

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220-239, 2017.

[7] L. A. Jeni, J. F. Cohn, and F. De la Torre, "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics," *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 245-251, 2013.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research.*, vol. 3, pp. 1157-1182, 2003.

[9] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Medical Genetics.*, vol. 19, no. 1, p. 65, 2018.

[10] C. Valverde, A. Francisco, J. Peláez Moreno, "100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox," *PLoS One*, vol. 9, no. 1, pp. 1-10, 2014.

[11] M. Kuhn, "Applied predictive modeling," 2013, *Springer*.

[12] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[13] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project," *PLoS One*, vol. 12, no. 7, 2017.

[14] A. Wijoyo, A. Y. Saputra, S. Ristanti, S. R. Sya'Ban, M. Amalia, and R. Febriansyah, "Pembelajaran Machine Learning," *OKTAL: Jurnal Ilmu Komputer dan Science*, vol. 3, no. 2, pp. 375-380, 2024.

[15] R. S. Nurhalizah, R. Ardianto, and Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *Jurnal Ilmu Komputer dan Informatika (JIKI)*, vol. 4, no. 1, pp. 61-72, 2024.

[16] S. R. Azizah, R. Herteno, A. Farmadi, D. Kartini, and I. Budiman, "Kombinasi Seleksi Fitur Berbasis Filter dan Wrapper Menggunakan Naive Bayes pada Klasifikasi Penyakit Jantung," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 10, no. 6, pp. 1361-1368, 2023.

[17] E. S. Septiany, H. H. Handayani, T. A. Mudzakir, and A. F. N. Masruriyah, "Optimasi Metode Support Vector Machine Menggunakan Seleksi Fitur Recursive Feature Elimination dan Forward Selection untuk Klasifikasi Kanker Payudara," *TIN: Terapan Informatika Nusantara*, vol. 5, no. 2, pp. 144-154, 2024.

[18] Ridwan, E. H. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, 2024.

[19] M. I. C. Rachmatullah, S. Armiati, and Mubassiran, "Menerapkan SMOTE pada Klasifikasi Data Penyakit Stroke," *IMPROVE*, vol. 17, no. 1, 2025.