JERIN: Jurnal Riset Elektro dan Informatika
Vol. 1, No. 1, Mei 2025, pp. 1-9

Prediksi Risiko Kardiovaskular Berdasarkan Faktor Gaya Hidup Menggunakan Algoritma Random Forest

Nazwa Syalaisa Haq[1], Artika Sari Murti[2*], Khayla Alifia Indrawati[3]
[1,2,3]Program Studi Teknik Informatika, Universitas Muslim Indonesia, Kota Makassar

[*]**Email Korespondensi:** 13020230070@student.umi.ac.id

Riwayat Artikel: Diterima: 10/02/2025; Direvisi: 17/04/2025; Disetujui: 02/05/2025

ABSTRAK

Penyakit kardiovaskular merupakan penyebab utama kematian secara global yang sangat dipengaruhi oleh faktor gaya hidup yang dapat dimodifikasi. Penelitian ini bertujuan untuk mengembangkan model prediksi risiko penyakit kardiovaskular yang objektif dengan memanfaatkan algoritma *Random Forest* berdasarkan data gaya hidup dan kesehatan. Metode penelitian mengikuti kerangka kerja CRISP-DM yang meliputi tahap pra-pemrosesan data seperti *label encoding*, standardisasi menggunakan *StandardScaler*, dan penanganan ketidakseimbangan data menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE). Model dikembangkan menggunakan mekanisme *Bootstrap Aggregating* dan dioptimasi melalui *Grid Search Cross-Validation*. Hasil eksperimen menunjukkan bahwa model *Random Forest* mampu mencapai akurasi sebesar 83,40% dengan nilai *precision* sebesar 0,93 pada kelas berisiko. Analisis tingkat kepentingan fitur mengungkapkan bahwa asupan air harian, detak jantung istirahat, dan tekanan darah merupakan prediktor yang paling signifikan dalam menentukan risiko kesehatan individu. Penelitian ini menyimpulkan bahwa model tersebut berpotensi menjadi instrumen deteksi dini guna mendukung intervensi kesehatan yang lebih cepat.

Kata Kunci: gaya hidup; imbalanced handling; prediksi; random forest; risiko kardiovaskular

1. PENDAHULUAN

Penyakit kardiovaskular (*Cardiovascular Disease*/CVD) masih menjadi penyebab utama kematian di dunia dengan lebih dari 17 juta kematian setiap tahunnya [1]. Faktor gaya hidup seperti pola makan tidak sehat, kebiasaan merokok, kurangnya aktivitas fisik, serta konsumsi alkohol berlebih merupakan faktor risiko utama yang dapat dimodifikasi [2]. Deteksi dini terhadap risiko penyakit kardiovaskular sangat penting, karena intervensi berbasis perubahan perilaku dan gaya hidup terbukti dapat menurunkan angka morbiditas dan mortalitas secara signifikan [3].

Beberapa penelitian terdahulu menunjukkan bahwa pendekatan *machine learning* (ML) mampu memberikan hasil yang lebih akurat dibandingkan metode statistik konvensional dalam memprediksi risiko CVD [4]. Algoritma *Random Forest* (RF) menjadi salah satu metode yang populer karena kemampuannya menangani data multivariabel, mengurangi risiko *overfitting*, serta memberikan interpretasi yang baik terhadap pentingnya fitur [5]. Studi [6] menunjukkan bahwa *Random Forest* mampu mengestimasi risiko kejadian kardiovaskular utama (MACE) dengan performa tinggi menggunakan data dunia nyata yang mencakup tekanan darah, kolesterol, dan kepatuhan terhadap pengobatan.

Penelitian [7] membandingkan kinerja *Random Forest* dengan regresi logistik dalam memprediksi *Premature Coronary Artery Disease* (PCAD) dan menemukan bahwa RF memberikan hasil yang lebih baik dalam mengidentifikasi pasien berisiko tinggi berdasarkan data klinis dan gaya hidup. Sementara itu, penelitian [8] menegaskan bahwa integrasi antara variabel gaya hidup dan parameter klinis melalui algoritma RF mampu menghasilkan model prediksi risiko kardiovaskular dengan performa yang konsisten dan dapat diandalkan.

Selain penelitian internasional, studi lokal juga menunjukkan hasil serupa. Penelitian [9] dalam Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) membuktikan bahwa algoritma *Random Forest* berhasil mengklasifikasikan risiko penyakit jantung berdasarkan *dataset* Cleveland dengan tingkat akurasi lebih dari 90%. Penelitian ini menunjukkan potensi RF dalam analisis kesehatan berbasis data publik. Penelitian dalam *RESTI Journal* menerapkan *Random Forest* pada kasus *Acute Coronary Syndrome* menggunakan data rekam

medis rumah sakit lokal dan melaporkan akurasi model sekitar 83% pada skenario pembagian data 70:30, serta hasil analisis *precision* dan *recall* yang baik. Temuan ini menunjukkan bahwa RF memiliki potensi untuk diterapkan pada analisis kesehatan berbasis data klinis di Indonesia. Firmansyah dan Yulianto [10] juga menemukan bahwa RF mampu memprediksi penyakit jantung dengan akurasi mencapai 91% menggunakan kerangka kerja CRISP-DM, menegaskan pentingnya proses eksplorasi dan pembersihan data sebelum pemodelan dilakukan.

Selain itu, penelitian [1] dan [2] menemukan bahwa algoritma *Random Forest* dan *XGBoost* memiliki performa prediksi yang sangat baik dalam memperkirakan risiko CVD dan mampu mengidentifikasi kontribusi relatif setiap faktor gaya hidup terhadap risiko tersebut. Temuan serupa diperkuat oleh [3], yang menekankan pentingnya *feature selection* dan *data preprocessing* untuk meningkatkan interpretabilitas model prediktif di bidang kesehatan masyarakat.

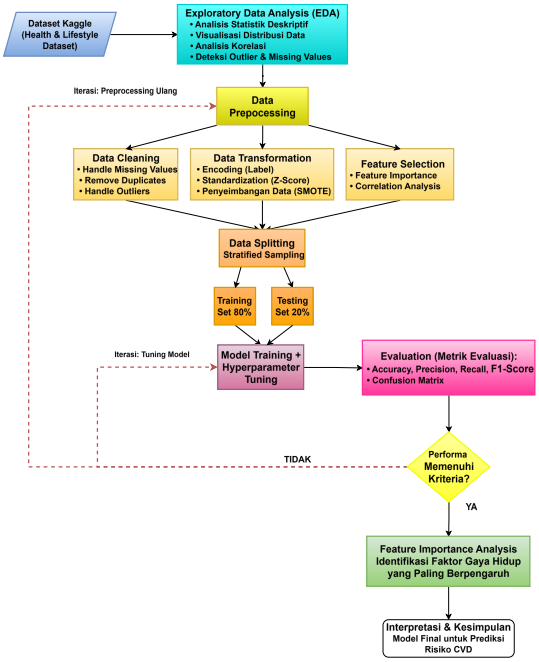
Konsistensi hasil dari berbagai penelitian tersebut menunjukkan bahwa algoritma *Random Forest* unggul karena kemampuannya menangani data yang kompleks, mengurangi risiko *overfitting*, dan memberikan pemahaman yang lebih baik mengenai variabel-variabel yang paling berpengaruh terhadap risiko kardiovaskular. Berdasarkan temuan tersebut, penelitian ini bertujuan untuk membangun model prediksi risiko kardiovaskular berdasarkan faktor gaya hidup menggunakan algoritma *Random Forest*, dengan memanfaatkan *dataset* publik yang mencakup data demografis, kebiasaan hidup, serta parameter kesehatan individu.

2. METODOLOGI PENELITIAN

Penelitian ini mengadopsi prinsip *Cross-Industry Standard Process for Data Mining* (CRISP-DM) sebagai landasan kerja yang sistematis dalam pengembangan model prediksi penyakit kardiovaskular [11]. Pemilihan kerangka kerja ini didasarkan pada efektivitasnya dalam memandu tahapan penambangan data mulai dari pemahaman data hingga evaluasi model yang telah terbukti sebagai standar industri yang andal [10], [11]. Prosedur eksperimen dalam penelitian ini difokuskan pada tahapan teknis yang linear guna memastikan seluruh proses transformasi dan pemodelan dapat terukur serta dapat direproduksi dengan baik [12].

2.1. Desain Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan metode *supervised learning* melalui implementasi algoritma *Random Forest* untuk klasifikasi risiko penyakit kardiovaskular. Desain penelitian mengadopsi kerangka kerja CRISP-DM [11] yang telah diadaptasi ke dalam tahapan teknis yang lebih spesifik untuk konteks prediksi kesehatan berbasis faktor gaya hidup [8], [10]. Tahapan-tahapan tersebut dirancang secara sistematis guna memastikan akurasi model dalam mengenali pola gaya hidup pasien [1]. Alur metodologi penelitian secara menyeluruh ditunjukkan pada Gambar 1. berikut:



Gambar 1. Diagram Alur

2.1.1. Akuisisi *Dataset*

Dataset yang digunakan dalam penelitian ini adalah *Health & Lifestyle Dataset* yang diperoleh dari repositori terbuka Kaggle. Penggunaan *dataset* publik dari repositori tepercaya merupakan praktik standar dalam pengembangan model prediktif untuk menjamin transparansi dan reproduksibilitas penelitian [10]. *Dataset* ini dipilih karena menyediakan cakupan data yang luas mengenai variabel demografis dan perilaku yang relevan dengan risiko kesehatan kronis [1].

Karakteristik *Dataset*:

a. Sumber: Repositori Kaggle (https://www.kaggle.com/datasets/rehan497/health-lifestyle-dataset).
b. Jumlah Observasi: 150.360 sampel data individu.
c. Dimensi Data: Terdiri dari 14 fitur variabel yang mencakup tipe data numerik dan kategorikal.
d. Target Variabel: Status risiko kardiovaskular dalam bentuk klasifikasi biner (0 untuk sehat, 1 untuk berisiko).

2.1.2. *Exploratory Data Analysis* (EDA)

Analisis eksplorasi data (*Exploratory Data Analysis/EDA*) dilakukan untuk memahami karakteristik, distribusi, dan kualitas data sebelum dilakukan pemrosesan lebih lanjut [10]. Tahapan ini krusial untuk memastikan bahwa variabel yang tepat dan bebas dari anomali yang dapat menyesatkan model [6]. Teknik yang digunakan meliputi:

a. Analisis Deskriptif: Menghitung statistik deskriptif (mean, median, standar deviasi) untuk variabel numerik guna memahami profil dasar pasien. Secara khusus, dilakukan tinjauan terhadap distribusi fitur berdasarkan jenis kelamin untuk mendeteksi perbedaan pola risiko yang signifikan antara laki-laki dan perempuan [13].
b. Analisis Distribusi: Mengidentifikasi pola distribusi data menggunakan histogram dan *boxplot* untuk melihat sebaran fitur gaya hidup serta mendeteksi kemiringan (*skewness*) data.
c. Analisis Korelasi: Mengukur keterkaitan antar variabel menggunakan matriks korelasi Pearson untuk menghindari multikolinieritas yang dapat memengaruhi interpretasi model [6].
d. Deteksi *Outlier*: Mengidentifikasi nilai-nilai ekstrim yang berada di luar jangkauan normal menggunakan metode *Interquartile Range* (IQR) atau *Z-Score* karena nilai tersebut dapat memengaruhi performa algoritma berbasis *decision tree* [5].
e. Analisis *Missing Value*: Melakukan identifikasi terhadap nilai yang hilang pada *dataset* untuk menentukan strategi imputasi atau penghapusan untuk menjaga intergitas informasi [9].

2.1.3. Data Preprocessing

Tahap ini merupakan langkah kritis untuk menjamin kualitas data melalui serangkaian transformasi sebelum tahap pemodelan dilakukan [3], [9]. Proses ini bertujuan mengoptimalkan data agar sesuai dengan karakteristik algoritma *Random Forest* sehingga model dapat mengenali pola secara lebih akurat dan efisien [5].

a. Pembersihan Data (*Data Cleaning*)

Proses pra-pemrosesan diawali dengan pembersihan data (*data cleaning*) untuk menjamin integritas *dataset* sebelum tahap pemodelan [9]. Penanganan nilai yang hilang (*missing values*) ditangani melalui imputasi median untuk fitur numerik dan modus untuk fitur kategorikal, atau penghapusan baris jika persentase kehilangan di bawah 5% [5]. Selanjutnya, deteksi *outliers* dilakukan dengan metode *Interquartile Range* (IQR) dan ditangani dengan teknik *winsorization* persentil 1% dan 99% untuk meminimalkan bias [5]. Selain itu, dilakukan penghapusan data duplikat untuk menghindari redudansi yang dapat mengakibatkan *overfitting* dan memengaruhi performa validasi model [3].

b. Transformasi Data (*Data Transformation*)

Transformasi data dilakukan untuk mengubah fitur ke dalam format yang optimal bagi algoritma *Random Forest* [3]. Tahapan ini meliputi:

1) *Encoding* Variabel Kategorikal

Mengubah variabel kategorikal menjadi numerik menggunakan *Label Encoding*. Hal ini dilakukan agar algoritma dapat memproses data non-numerik tanpa menghilangkan informasi hierarkis jika ada [7].

2) Normalisasi/Standardisasi

Melakukan penskalaan data menggunakan *StandarScaler* agar seluruh fitur memiliki rentang nilai yang seragam dengan rata-rata (μ) dan standar deviasi (σ) = 1 [3]. Rumus yang digunakan adalah:

$$z = \frac{(x - \mu)}{\sigma} \quad (1)$$

Di mana:
z: nilai terstandarisasi
x: nilai asli
μ: rata-rata
σ: standar deviasi

3) Penyeimbangan Data (*Data Imbalance Treatment*)

Untuk mengatasi ketidakseimbangan proporsi antara kelas individu sehat dan berisiko, penelitian ini menerapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) [14]. Teknik ini bekerja dengan mensintesis sampel baru pada kelas minoritas melalui pendekatan *k-nearest neighbors* sehingga model dapat mempelajari pola secara objektif tanpa mengalami bias terhadap kelas mayoritas [12].

Implementasi SMOTE ini krusial untuk meningkatkan sensitivitas model terhadap kelas berisiko yang jumlahnya lebih sedikit dibandingkan kelas sehat [14]. Sampel sintesis (x_{new}) dihasilkan menggunakan persamaan berikut:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \quad (2)$$

Di mana:
x_i: sampel dari kelas minoritas
x_{zi}: salah satu dari *k-nearest neighbors* terdekat
λ: angka acak antara 0 dan 1.

c. Seleksi Fitur (*Feature Selection*)

Seleksi fitur dilakukan untuk mengidentifikasi variabel yang paling relevan untuk meningkatkan efisiensi proses komputasi dan akurasi prediksi model dalam mendeteksi risiko kesehatan masyarakat [3]. Proses ini diawali dengan menerapkan teknik *Feature Importance* pada algoritma *Random Forest* untuk mengukur kontribusi relatif setiap fitur gaya hidup berdasarkan nilai penurunan rata-rata *Gini Impurity* pada seluruh pohon keputusan dalam model [10]. Melalui mekanisme ini, fitur-fitur yang memiliki skor kepentingan rendah dan tidak signifikan dalam klasifikasi dapat dieliminasi untuk menyederhanakan kompleksitas model tanpa mengurangi performanya secara drastis [5].

Selanjutnya, dilakukan *Correlation Analysis* menggunakan koefisien Pearson untuk memeriksa keterkaitan antar variabel dan mendeteksi adanya multikolinieritas. Langkah ini sangat penting untuk memastikan bahwa tidak terdapat variabel dengan korelasi terlalu tinggi yang berpotensi mengaburkan interpretabilitas hasil prediksi, sehingga model tetap memberikan informasi objektif mengenai faktor risiko kardiovaskular [6].

2.1.4. Pembagian Data (*Data Splitting*)

Dataset dibagi menjadi dua subset utama untuk menjamin validitas proses pelatihan dan pengujian model sesuai dengan standar desain studi risiko kardiovaskular [15]. Proses pembagian ini menggunakan metode *stratified sampling* untuk menjaga proporsi kelas target tetap konsisten pada kedua subset agar tidak terjadi bias distribusi data. Sebesar 80% data dialokasikan sebagai *training set* yang berfungsi sebagai basis pengetahuan bagi algoritma dalam mempelajari pola risiko gaya hidup maupun kesehatan [12]. Sementara itu, 20% data sisanya digunakan sebagai *testing set* atau data independen untuk mengevaluasi kemampuan generalisasi model terhadap data baru demi meminimalkan potensi kesalahan diagnosis dalam klasifikasi model [12], [16].

2.1.5. Model Training & Hyperparameter

Tahap pemodelan dilakukan dengan mengimplementasikan algoritma *Random Forest*, sebuah metode *ensemble learning* yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang akurat dan stabil [6]. Algoritma ini dipilih karena kemampuannya menangani data medis yang bersifat non-linear,

robust terhadap *overfitting*, serta mampu memberikan informasi tingkat kepentingan *feature importance* yang krusial bagi analisis medis [10], [17]. Proses pengembangan model ini dilakukan melalui tahapan berikut:

a. Prinsip Kerja Model: Algoritma bekerja menggunakan mekanisme *Bootstrap Aggregating* (Bagging) dengan membuat subset data secara acak serta melakukan seleksi fitur acak pada setiap *node*. Prediksi akhir ($\hat{y}$) ditentukan melalui *Majority Voting* yaitu mengambil kelas yang paling sering muncul (modus) dari seluruh pohon keputusan [17].

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_n(x)\} \quad (3)$$

Di mana:
$\hat{y}$ = prediksi kelas
$h_1(x)$ = prediksi dari pohon ke-i
n = jumlah pohon dalam forest

b. *Hyperparameter Tuning*: Untuk mencapai performa optimal, dilakukan pencarian kombinasi parameter kombinasi terbaik menggunakan metode *Grid Search Cross-Validation* [12]. Teknik ini bekerja dengan mengevaluasi secara sistematis seluruh kombinasi parameter yang ditentukan untuk menemukan konfigurasi yang menghasilkan akurasi tertinggi dan meminimalkan bias [2]. Parameter yang dioptimasi meliputi jumlah pohon (*n_estimators*), kedalaman maksimal (*max_depth*), dan jumlah fitur yang dipertimbangkan (*max_features*) sebagaimana dirinci pada tabel di bawah ini:

Tabel 1. Tabel Parameter

Parameter	**Rentang Nilai**	**Rentang Nilai**
n_estimators	[100, 200, 300, 500]	Jumlah pohon dalam forest
max_depth	[10, 20, 30, None]	Kedalaman maksimum pohon
min_samples_split	[2, 5, 10]	Jumlah minimum sampel untuk split
min_samples_leaf	[1, 2, 4]	Jumlah minimum sampel pada leaf node
max_features	['sqrt', 'log2', None]	Jumlah fitur untuk split terbaik

c. Proses Pelatihan: Model diinisialisasi dengan konfigurasi parameter terbaik dan dilatih menggunakan *training set*. Untuk memastikan kemampuan generalisasi yang optimal, diterapkan *5-fold cross-validation* selama proses pelatihan guna memantau metrik performa secara konsisten [12]. Teknik ini membagi data menjadi lima bagian secara bergiliran untuk validasi sehingga mendeteksi potensi *overfitting* secara dini dan menjamin model tetap reliabel saat menghadapi data pasien baru [3], [17].

2.1.6. Evaluation Model

Tahap evaluasi bertujuan untuk mengukur performa model dari berbagai perspektif guna memastikan keandalan prediksi risiko kardiovaskular pada tingkat klinis [16]. Pengujian dilakukan menggunakan data yang tidak terlibat dalam proses pelatihan (*testing set*) dengan mengacu pada metrik evaluasi standar dalam klasifikasi biner untuk meminimalkan kesalahn diagnosis [5], [16]. Metrik yang digunakan dalam penelitian ini meliputi:

a. *Confusion Matrix*: Tabel kontingensi yang digunakan untuk memetakan hasil prediksi model dibandingkan dengan nilai aktual [5].

$$Confusion\ Matrix = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \quad (4)$$

b. *Accuracy*: Mengukur sejauh mana model mampu mengklasifikasikan seluruh data secara benar.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

c. *Precision*: Mengukur ketepatan model dalam memprediksi kelas positif dari seluruh hasil yang diprediksi positif untuk mengurangi kesalahan *false positive*.

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

d. *Recall* (*Sensitivity*): Mengukur kemampuan model dalam mengidentifikasi seluruh sampel yang benar-benar berisiko kardiovaskular. Metrik ini sangat krusial dalam dominan kesehatan untuk menghindari kegagalan deteksi pasian sakti [16].

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

e. F1-Score: Rata-rata harmonik antara *precision* dan *recall* yang memberikan gambaran keseimbangan performa model, terutama setelah penerapan teknik SMOTE untuk menangani ketidakseimbangan kelas [12].

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (8)$$

Di mana:
TP (*True Positive*): Prediksi benar untuk kelas berisiko.
TN (*True Negative*): Prediksi benar untuk kelas sehat.
FP (*False Positive*): Prediksi salah (model menebak berisiko, aslinya sehat).
FN (*False Negative*): Prediksi salah (model menebak sehat, aslinya berisiko).

2.1.7. Feature Importance Analysis

Setelah model terbaik terbentuk, dilakukan analisis *feature importance* untuk menentukan tingkat kontribusi relatif setiap variabel gaya hidup terhadap prediksi risiko kardiovaskular [6]. Proses ini menggunakan perhitungan penurunan rata-rata *Gini Impurity* pada setiap *node* di seluruh pohon keputusan dalam *Random Forest* [10].

Hasil dari tahapan ini memberikan gambaran variabel mana yang paling dominan dalam membedakan antara individu yang berisiko dan tidak berisiko sehingga dapat memberikan wawasan tambahan bagi pengambilan keputusan klinis terkait faktor gaya hidup yang paling kritis [17].

2.1.8. Interpretasi dan Kesimpulan

Tahap akhir adalah interpretasi hasil untuk memberikan wawasan klinis dari model yang telah diuji. Hal ini penting untuk memastikan model tidak bersifat *black-box* dan dapat dipahami secara logis dalam konteks medis [5]. Teknik yang digunakan meliputi:

a. *Partial Dependence Plot* (PDP): Memvisualisasikan hubungan fungsional antara fitur tertentu dengan prediksi risiko untuk melihat pola kecenderungan data secara grafis [6].
b. SHAP (*Shapley Additive Explanations*) *Values*: Menggunakan pendekatan *game theory* untuk mengukur kontribusi setiap fitur terhadap prediksi pada level individu, sehingga memberikan transparansi pada keputusan model [3].
c. Penarikan Kesimpulan: Merangkum seluruh temuan eksperimen untuk menjawab tujuan penelitian mengenai efektivitas model *Random Forest* dalam mendeteksi risiko kardiovaskular secara akurat dan objektif [10], [17].

3. HASIL DAN PEMBAHASAN

Pada bagian ini dipaparkan temuan utama dari eksperimen yang telah dilakukan, meliput hasil pemrosesan data, evaluasi kinerja model *Random Forest*, serta analisis variabel yang paling berpengaruh terhadap prediksi risiko kardiovaskular.

3.1. Hasil Penelitian

3.1.1. Performa Model Klasifikasi

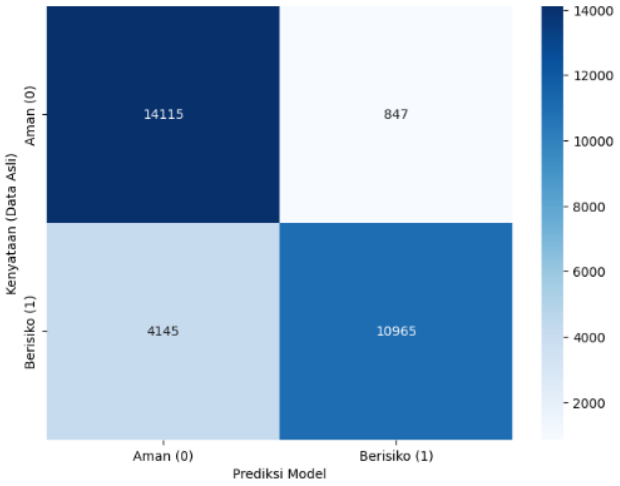
Bagian ini menyajikan hasil evaluasi performa model *Random Forest* dalam mengklasifikasikan risiko penyakit kardiovaskular berdasarkan data gaya hidup. Pengujian dilakukan menggunakan data uji (*testing set*) setelah melalui tahap pra-pemrosesan dan penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE).

Evaluasi performa model dilakukan menggunakan metrik accuracy, precision, dan recall. Berdasarkan hasil pengujian, model menghasilkan nilai akurasi sebesar 83,40%. Ringkasan hasil evaluasi performa model berdasarkan laporan klasifikasi disajikan pada Tabel 2.

Tabel 2. Tabel Evaluasi Performa Model

Kelas	Kategori	Precision	Recall	F1-Score	Accuracy
0	Aman	0.77	**0.94**	**0.85**	0.834
1	Berisiko	**0.93**	0.73	0.81	0.834

Distribusi hasil prediksi model terhadap data uji dianalisis lebih lanjut menggunakan *confusion matrix* yang ditampilkan pada Gambar 2.

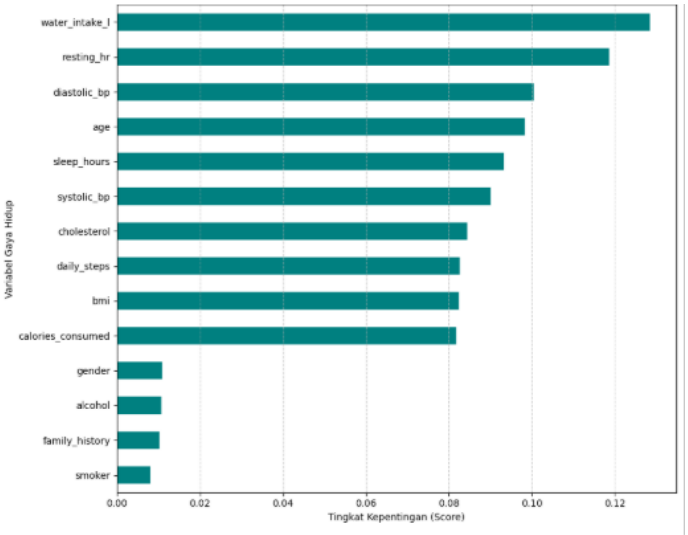


Gambar 2. *Confusion Matrix*

Confusion matrix pada Gambar 2. menunjukkan jumlah prediksi yang benar dan salah pada masing-masing kelas, yaitu kelas tidak berisiko dan kelas berisiko, berdasarkan hasil klasifikasi model terhadap data uji.

3.1.2. Faktor Gaya Hidup Dominan (*Feature Importance*)

Peringkat kepentingan fitur diperoleh dari hasil ekstraksi model *Random Forest* untuk menentukan variabel yang paling berkontribusi dalam klasifikasi risiko kesehatan. Tingkat kepentingan setiap variabel diukur menggunakan skor numerik yang merepresentasikan pengaruh fitur tersebut terhadap keputusan model.



Gambar 3. *Feature Importance*

Berdasarkan grafik visualisasi pada Gambar 3., ditemukan bahwa asupan air (*water intake*) merupakan fitur dengan kontribusi paling dominan terhadap prediksi model dengan skor *importance* tertinggi, yang kemudian diikuti oleh detak jantung istirahat (*resting heart rate*) di posisi kedua. Faktor tekanan darah, baik diastolik maupun sistolik (*diastolic & systolic BP*), serta usia juga menunjukkan nilai kepentingan yang signifikan dan masuk ke dalam kategori faktor risiko utama dalam model ini. Sementara itu, variabel terkait

aktivitas fisik dan kebiasaan sehari-hari seperti durasi tidur (*sleep hours*), jumlah langkah harian (*daily steps*), konsumsi kalori (*calories consumed*), serta indeks massa tubuh (*BMI*) memiliki tingkat kepentingan pada level menengah. Adapun fitur-fitur seperti jenis kelamin (*gender*), konsumsi alkohol (*alcohol*), riwayat keluarga (*family history*), dan status perokok (*smoker*) tercatat memiliki nilai kontribusi paling rendah dibandingkan fitur lainnya dalam dataset ini.

3.2. Pembahasan

Capaian Akurasi sebesar 83,40% dan F1-*Score* yang stabil menunjukkan bahwa model *Random Forest* berhasil memetakan profil gaya hidup dengan sangat baik. Poin krusial dalam penelitian ini adalah peningkatan nilai *Recall* pada kelas berisiko mencapai 73% yang membuktikan bahwa penerapan SMOTE berhasil menghilangkan bias model terhadap kelas mayoritas. Dengan menyeimbangkan dataset, model mampu mempelajari pola unik dari kelompok minoritas sehingga risiko kegagalan deteksi pada individu yang sebenarnya membutuhkan perhatian medis dapat diminimalisir secara signifikan dibandingkan dengan model tanpa penyeimbangan data.

Di sisi lain, analisis terhadap *Precision* memberikan gambaran mengenai efisiensi model dalam menekan angka *False Positive* sehingga meminimalisir kesalahan interpretasi bagi individu sehat. Meskipun terdapat 4.145 kesalahan *False Negative*, hal ini merupakan konsekuensi dari tumpang tindih (*overlap*) fitur gaya hidup yang sangat mirip antar individu. Dominasi fitur asupan air harian dan detak jantung istirahat sebagai prediktor utama mengonfirmasi bahwa indikator fisiologis harian memiliki bobot informasi yang besar dalam skirining awal risiko kardiovaskular meskipun verifikasi klinik tetap diperlukan untuk mengatasi keterbatasan dataset ini.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma *Random Forest* (RF) menggunakan kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM) untuk memprediksi risiko penyakit kardiovaskular dengan akurasi 83,40%. Penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) terbukti efektif mengatasi ketidakseimbangan data (*data imbalance*) sehingga menghasilkan nilai *precision* 0,93 dalam mendeteksi kelas berisiko. Melalui analisis *feature importance*, diidentifikasi bahwa asupan air harian, detak jantung istirahat, dan tekanan darah merupakan prediktor gaya hidup paling signifikan. Secara keseluruhan, model yang dioptimasi dengan *Grid Search Cross-Validation* (GSCV) ini valid sebagai instrumen deteksi dini risiko kesehatan. Pengembangan selanjutnya disarankan untuk memperluas dataset dengan variabel klinis spesifik guna meningkatkan generalisasi model pada populasi yang lebih luas.

DAFTAR PUSTAKA

[1] S. A. Kissi, M. G. M. Talukder, and M. Z. Iqbal, "Data-Driven Predictive Modelling of Lifestyle Risk Factors for Cardiovascular Health," *Electron.*, vol. 14, no. 14, 2025, doi: 10.3390/electronics14142906.

[2] B. E. Sianga, M. C. Mbago, and A. S. Msengwa, "Predicting the prevalence of cardiovascular diseases using machine learning algorithms," *Intell. Med.*, vol. 11, no. January, p. 100199, 2025, doi: 10.1016/j.ibmed.2025.100199.

[3] G. I. Al Jowf and M. Kolhar, "Key factors in predictive analysis of cardiovascular risks in public health," *Sci. Rep.*, vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-025-07874-x.

[4] A. Dogan, Y. Li, C. Peter Odo, K. Sonawane, Y. Lin, and C. Liu, "A utility-based machine learning-driven personalized lifestyle recommendation for cardiovascular disease prevention," *J. Biomed. Inform.*, vol. 141, no. March, p. 104342, 2023, doi: 10.1016/j.jbi.2023.104342.

[5] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Informatics Med. Unlocked*, vol. 41, no. July, p. 101316, 2023, doi: 10.1016/j.imu.2023.101316.

[6] S. Castel-Feced, I. Aguilar-Palacio, S. Malo, J. González-García, L. Maldonado, and M. J. Rabanaque-Hernández, "Prediction of cardiovascular risk using machine-learning methods. Sex-specific differences," *Front. Cardiovasc. Med.*, vol. 12, no. June, pp. 1–11, 2025, doi: 10.3389/fcvm.2025.1579947.

[7] J. Wang, Y. Xu, L. Liu, W. Wu, C. Shen, H. Huang, Z. Zhen, J. Meng, C. Li, and Z. Qu, "Comparison of LASSO and random forest models for predicting the risk of premature coronary artery disease," BMC Medical Informatics and Decision Making, vol. 23, no. 1, pp. 1–10, 2023, doi: 10.1186/s12911-023-02407-w.

[8] H. Wang, W. J. Tucker, J. Jonnagaddala, A. E. Schutte, B. Jalaludin, S. T. Liaw, K. A. Rye, R. K. Wong, and K. L. Ong, "Using machine learning to predict cardiovascular risk using self-reported questionnaires: Findings from the 45 and Up Study," International Journal of Cardiology, vol. 386, pp. 149–156, 2023, doi: 10.1016/j.ijcard.2023.05.030.

[9] E. P. Cynthia, M. A. Rizky, A. Nazir, and F. Syafria, "Random Forest Algorithm to Investigate the Case of Acute Coronary Syndrome," RESTI J. (Syst. Eng. Inf. Technol.), vol. 5, no. 2, pp. 369–378, 2021, doi: 10.29207/resti.v5i2.3000.

[10] A. Y. Firmansyah, "Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest," *J. Inf. dan Komput.*, vol. 11, no. 02, pp. 184–189, 2023, doi: 10.35959/jik.v11i02.499.

[11] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[12] A. E. Pranajaya and E. R. Susanto, "Optimasi Random Forest untuk Prediksi Penyakit Jantung Menggunakan SMOTEENN dan Grid Search," JPTI, vol. 5, no. 7, pp. 1965–1979, 2025.

[13] L. B. Elvas, M. Nunes, J. C. Ferreira, M. S. Dias, and L. B. Rosário, "AI-Driven Decision Support for Early Detection of Cardiac Events: Unveiling Patterns and Predicting Myocardial Ischemia," *J. Pers. Med.*, vol. 13, no. 9, 2023, doi: 10.3390/jpm13091421.

[14] A. Rahim, I. Pratiwi, and M. Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique dan Random Forest," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 2995–3011, 2023.

[15] D. Pella, S. Toth, J. Paralic, J. Gonsorcik, J. Fedacko, P. Jarcuska, D. Pella, Z. Pella, F. Sabol, M. Jankajova, G. Valocik, A. Putrya, A. Kirschova, L. Plachy, M. Rabajdova, M. Hunavy, B. Kafkova, I. Doci, S. Timkova, M. Dvoroznakova, F. Babic, P. Butka, L. Dimunova, M. Marekova, Z. Paralicova, J. Majernik, J. Luczy, J. Janosik, and M. Kmec, "The possible role of machine learning in detection of increased cardiovascular risk patients – KSC MR Study (design)," Archives of Medical Science, vol. 18, no. 4, pp. 991–997, 2022, doi: 10.5114/aoms.2020.99156.

[16] M. Ahmed, M. H. Sulaiman, M. M. Hassan, and T. Bhuiyan, "Predicting the Classification of Heart Failure Patients Using Optimized Machine Learning Algorithms," *IEEE Access*, vol. 13, pp. 30555–30569, 2025, doi: 10.1109/ACCESS.2025.3541069.

[17] L. Yang, H. Wu, X. Jin, P. Zheng, S. Hu, X. Xu, W. Yu, and J. Yan, "Study of cardiovascular disease prediction model based on random forest in eastern China," Scientific Reports, vol. 10, no. 1, p. 5245, 2020.