ZONAsi Jurnal Sistem Informasi
ZONAsi VOL. 8 NO. 2 Mei 2026
Page: 854 - 863 Akreditasi Sinta 4
ISSN: 2656-7407 (Online)
2656-7393 (Print)

PEMODELAN PREDIKSI DAN STRATIFIKASI RESIKO SERANGAN JANTUNG DENGAN *RANDOM FOREST* DAN *LOGISTIC REGRESSION* BERBASIS *CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING*

Heidy Mudita Sutedjo[1], Christian[2], Amanda Michelle Darwis[3], Dwinda Audia Irnaonefa[4]
[1]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Ciputra Surabaya
CitraLand CBD Boulevard, Made, Kec. Sambikerep, Kota Surabaya, 0317451699
e-mail: [1]hsutedjo@student.ciputra.ac.id, [2]christian03@ciputra.ac.id,
[3]am000001@student.ciputra.ac.id, [4]dirnaonefa@student.ciputra.ac.id

Abstrak

Penelitian ini bertujuan mengembangkan model prediktif untuk klasifikasi dan stratifikasi risiko serangan jantung menggunakan pendekatan machine learning dalam kerangka kerja CRISP-DM. Data klinis yang digunakan berasal dari 1.319 pasien dengan atribut meliputi usia, jenis kelamin, detak jantung, kadar Creatine Kinase-MB (CK-MB), dan kadar Troponin. Dua algoritma klasifikasi, yaitu Random Forest dan Logistic Regression, diterapkan dan dievaluasi menggunakan pembagian data uji sebesar 20%. Hasil pengujian menunjukkan bahwa model Random Forest memberikan performa terbaik dengan tingkat akurasi sebesar 98,48%, sedangkan Logistic Regression memperoleh akurasi sebesar 81,06%. Analisis pentingnya fitur mengungkapkan bahwa kadar CK-MB dan Troponin merupakan faktor prediktif paling berpengaruh dalam menentukan risiko serangan jantung. Temuan ini menunjukkan bahwa penerapan model prediktif yang dikembangkan secara sistematis melalui metodologi CRISP-DM memiliki potensi besar untuk mendukung sistem pendukung keputusan klinis dalam deteksi dini serangan jantung. Penelitian selanjutnya disarankan untuk melakukan validasi eksternal menggunakan dataset dari populasi yang lebih beragam guna meningkatkan generalisasi model serta meningkatkan keandalan, akurasi, dan robustitas hasil prediksi secara menyeluruh.
Kata kunci: Serangan Jantung, CRISP-DM, *Machine Learning*, *Random Forest*, *Logistic Regression*.

Abstract

This study aims to develop a predictive model for the classification and stratification of heart attack risk using a machine learning approach within the CRISP-DM framework. The clinical dataset used in this research consists of 1,319 patients, with attributes including age, gender, heart rate, Creatine Kinase-MB (CK-MB) levels, and Troponin levels. Two classification algorithms, namely Random Forest and Logistic Regression, were implemented and evaluated using a 20% test data split. The experimental results indicate that the Random Forest model achieved the best performance, with an accuracy of 98.48%, while Logistic Regression obtained an accuracy of 81.06%. Feature importance analysis revealed that CK-MB and Troponin levels are the most influential predictive factors in determining heart attack risk. These findings suggest that the systematic development of predictive models using the CRISP-DM methodology has significant potential to support clinical decision support systems for early detection of heart attacks.Future research is recommended to conduct external validation using more diverse population datasets to improve the generalizability of the model, as well as to enhance the reliability, accuracy, and robustness of the overall predictive outcomes.
Keywords: *Heart Attack*, CRISP-DM, *Machine Learning*, *Random Forest*, *Logistic Regression*.

1. PENDAHULUAN

Serangan jantung tetap menjadi penyebab utama kematian global, dengan lebih dari 8,9 juta kematian terkait penyakit jantung iskemik pada tahun 2019 [1]. Di Indonesia, prevalensi penyakit jantung koroner meningkat sebesar 15% dalam dekade terakhir, yang memberikan beban berat pada

sistem kesehatan nasional [2]. Deteksi dini melalui analisis parameter klinis menjadi solusi krusial mengingat sekitar 50% kasus serangan jantung tidak menunjukkan gejala spesifik pada tahap awal [3]. Oleh karena itu, pengembangan metode prediksi yang akurat sangat dibutuhkan untuk mengidentifikasi pasien berisiko tinggi secara lebih efektif.

Perkembangan teknologi *machine learning* menawarkan pendekatan revolusioner dalam identifikasi risiko serangan jantung berbasis data klinis. Analisis terhadap lebih dari 2.300 pasien menunjukkan bahwa kombinasi biomarker jantung, seperti Troponin dan CK-MB, dengan algoritma *Random Forest* mampu menghasilkan akurasi prediksi hingga 92% [4]. Metode *Logistic Regression* juga tercatat mampu mencapai akurasi sekitar 89% pada analisis dataset berisi 678 sampel [5]. Meskipun demikian, tantangan utama dalam penerapan model ini adalah ketergantungan pada ketersediaan fitur laboratorium dan konsistensi data yang digunakan.

Penelitian ini menggunakan model CRISP-DM sebagai kerangka kerja sistematis untuk membangun model prediksi serangan jantung. CRISP-DM adalah metodologi standar yang mencakup enam tahapan utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Pendekatan ini memungkinkan proses pengolahan data dan pengembangan model dilakukan dengan cara yang terstruktur dan dapat diulang [6]. Dengan menggunakan model CRISP-DM, penelitian ini mengolah dataset spesifik yang mencakup 1.319 rekam medis pasien dengan fitur terbatas seperti usia, jenis kelamin, detak jantung, kadar CK-MB, Troponin, dan label diagnosis serangan jantung [7].

Berbeda dengan penelitian sebelumnya yang menggabungkan variabel tambahan seperti riwayat keluarga atau pola hidup pasien [8], studi ini menguji kemampuan algoritma *Random Forest* dan *Logistic Regression* dalam kondisi data yang minimalis. Fokus utama penelitian ini adalah menganalisis performa kedua metode tersebut sekaligus mengidentifikasi variabel klinis yang paling berpengaruh dalam prediksi risiko serangan jantung. Hasil penelitian diharapkan dapat menjadi dasar pengembangan sistem skrining berbasis biomarker yang efektif di fasilitas kesehatan primer dan memperkaya literatur tentang aplikasi *machine learning* dalam bidang kardiologi preventif.

Pemilihan model prediksi penyakit jantung perlu memperhatikan bias terhadap kategori mayoritas serta trade-off antara kompleksitas sistem dan performa aktual [9]. *Logistic Regression* dan *Random Forest* telah terbukti efektif dalam memprediksi risiko penyakit jantung, bahkan mencapai akurasi identik sebesar 86,84% pada dataset tertentu [10]. *Logistic Regression* juga mampu mencapai akurasi tinggi hingga 98,44% dengan memanfaatkan parameter klinis seperti usia, tekanan darah, dan kolesterol, serta dapat ditingkatkan melalui integrasi dengan metode *clustering* [11]. Pada dataset UCI, *Logistic Regression* berhasil memperoleh akurasi tertinggi sebesar 89%, mengungguli *k-Nearest Neighbors* dan *Neural Network*, dengan keunggulan pada interpretabilitas dan efisiensi komputasi [12]. Sementara itu, *Random Forest* tetap kompetitif dengan kemampuan identifikasi fitur penting yang berkontribusi pada peningkatan prediksi risiko penyakit jantung [12].

Penggunaan CRISP-DM sebagai kerangka kerja pengembangan model prediksi juga menunjukkan hasil yang signifikan. Penerapan *Random Forest* dan *Logistic Regression* dalam kerangka CRISP-DM menghasilkan akurasi masing-masing sebesar 90% dan 87% [13]. Selain itu, pendekatan ini efektif dalam pemilihan fitur, tuning parameter, serta meminimalkan bias dan *overfitting* [14]. Variabel klinis seperti tekanan darah sistolik, Troponin, dan CK-MB terbukti berkontribusi penting dalam meningkatkan akurasi model [15], [16], dengan biomarker digunakan sebagai indikator diagnosis dini [15]. CRISP-DM dinilai relevan untuk analisis kardiovaskular karena kemampuannya mengelola data kompleks secara sistematis dan berulang [17]–[19].

Teknologi *machine learning* secara umum efektif dalam mengenali pola hubungan non-linear antar variabel, serta dalam mengolah data biomarker, sinyal elektrokardiogram (ECG), dan parameter klinis lainnya untuk meningkatkan akurasi prediksi [20]. *Logistic Regression* meskipun sederhana tetap kompetitif, khususnya pada data dengan hubungan linier [21]. Di sisi lain, *Random Forest* sebagai metode *ensemble* unggul dalam menangani data kompleks, mengevaluasi *feature importance*, serta terbukti efektif pada dataset berskala kecil dalam prediksi penyakit kardiovaskular [22], [20].

2. METODE PENELITIAN

2.1. Analisis Kebutuhan Dataset

Dataset dalam penelitian ini menggunakan data rekam medis pasien yang memuat atribut-atribut klinis seperti usia, jenis kelamin, detak jantung, kadar CK-MB, Troponin, serta status serangan jantung sebagai variabel target. Pemilihan atribut tersebut didasarkan pada relevansi klinisnya dalam mencerminkan kondisi fisiologis dan biologis pasien yang berkaitan langsung dengan kemungkinan terjadinya serangan jantung. Usia dan jenis kelamin berperan sebagai variabel demografis penting yang mampu memberikan gambaran risiko berdasarkan karakteristik kelompok pasien. Detak jantung mencerminkan kondisi tubuh yang bisa berubah drastis pada individu yang mengalami gangguan jantung. Sementara itu, kadar CK-MB dan Troponin merupakan biomarker utama yang telah terbukti secara klinis menjadi indikator kerusakan jaringan jantung dan memiliki pengaruh besar terhadap hasil prediksi, khususnya terlihat dari analisis feature importance pada algoritma *Random Forest*. Variabel target berupa status serangan jantung digunakan untuk klasifikasi biner, yaitu menentukan apakah seorang pasien mengalami serangan jantung atau tidak.

Pembagian data dilakukan secara proporsional, dengan sebagian besar data digunakan untuk pelatihan model dan sisanya untuk pengujian. Strategi ini bertujuan untuk memastikan bahwa performa model dapat dievaluasi secara objektif pada data yang belum pernah dilihat sebelumnya. Selain itu, kualitas data menjadi aspek penting yang perlu diperhatikan, terutama pada variabel biomarker seperti CK-MB dan Troponin. Data harus bebas dari nilai yang hilang dan *outlier* yang ekstrem agar proses pelatihan model dapat berlangsung secara optimal dan menghasilkan prediksi yang akurat. Pemilihan metode *Logistic Regression* didasarkan pada kemampuannya dalam menangani dataset dengan jumlah fitur terbatas serta menghasilkan model yang mudah diinterpretasikan dalam konteks klinis. Sementara *Random Forest* digunakan karena kemampuannya dalam menangkap hubungan non-linear antar variabel serta kemampuannya dalam memberikan informasi feature importance yang membantu mengidentifikasi atribut paling berpengaruh dalam prediksi. Dengan demikian, seluruh atribut yang tersedia dalam dataset ini digunakan secara maksimal untuk membangun model prediksi risiko serangan jantung. Meskipun jumlah fitur terbatas, data yang tersedia cukup relevan dan telah memenuhi kebutuhan analisis untuk mendukung penerapan kedua algoritma klasifikasi secara efektif.

2.2. Sumber Data dan Variabel Penelitian

Penelitian ini menggunakan dataset yang diperoleh dari situs Kaggle berjudul "*Heart Attack Dataset*" [7]. Dataset ini terdiri dari 1.319 rekam medis pasien dengan 810 baris data positif terkena serangan jantung yang dikumpulkan untuk analisis prediksi dini serangan jantung. Dataset mencakup 9 kolom, dengan kolom *Result* sebagai variabel target dan 8 kolom lainnya sebagai variabel prediktor, mencakup usia, jenis kelamin, tekanan darah sistolik dan diastolik, detak jantung, kadar gula darah, serta biomarker jantung (CK-MB dan Troponin).

Seperti yang terlihat pada Tabel 2, penelitian ini menggunakan tambahan variabel *BP Ratio* (rasio tekanan darah sistolik terhadap diastolik), serta menggantikan variabel usia dengan kategori usia pasien berdasarkan interval (*Young*: ≤30 tahun, *Adult*: 31-50 tahun, *Senior*: 51-70 tahun, *Elderly*: >70 tahun) untuk analisis yang lebih terstruktur, sehingga terdapat total 10 variabel yang digunakan, dengan variabel *Result* sebagai variabel target.

Tabel 2. Tampilan dataset

Atribut	Data ke-1	Data ke-2	Data ke-3
Heart Rate	66	94	64
Systolic blood pressure	160	98	160
Diastolic blood pressure	83	46	77
Blood sugar	160.0	296.0	270.0

Atribut	Data ke-1	Data ke-2	Data ke-3
CK-MB	1.80	6.75	1.99
Troponin	0.012	1.060	0.003
BP_Ratio	1.93	2.13	2.08
Age_Group_Young	0	1	0
Age_Group_Adult	0	0	0
Age_Group_Senior	1	0	1
Age_Group_Elderly	0	0	0
Gender	1	1	1

Sebelum membangun model *machine learning*, umumnya dataset perlu melalui tahap *pre-processing*. Namun, dataset yang digunakan dalam penelitian ini telah melalui proses pembersihan dan manipulasi data sebelumnya serta memiliki distribusi kelas yang seimbang, sehingga sudah siap untuk digunakan dalam pengembangan model.

2.3 Prosedur Analisis

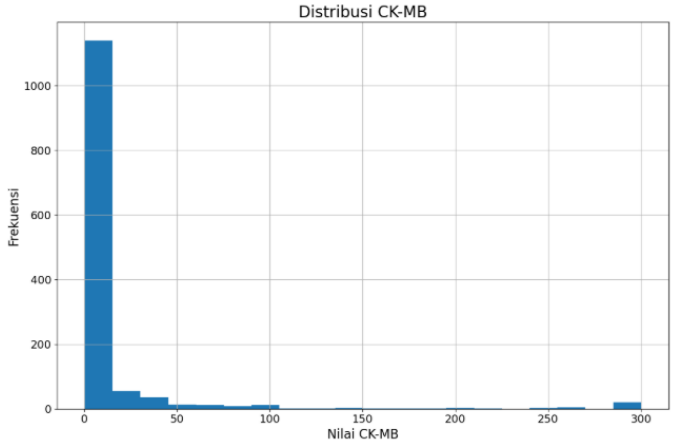
Penelitian ini menggunakan kerangka kerja CRISP-DM sebagai pedoman dalam analisis data. Kerangka ini terdiri dari enam tahapan sistematis sebagai berikut:

2.3.1 Business Understanding

Langkah pertama dalam kerangka CRISP-DM adalah memperoleh pemahaman yang mendalam mengenai tujuan bisnis, yang dalam konteks penelitian ini fokus utamanya adalah mengembangkan model prediksi risiko serangan jantung guna mendukung deteksi dini serta pengambilan keputusan medis yang lebih akurat. Pemahaman ini menjadi landasan penting agar proses analisis data dapat difokuskan dan disesuaikan dengan kebutuhan riil di sektor kesehatan. Dengan memahami tujuan tersebut secara jelas, peneliti dapat merancang kriteria keberhasilan model secara tepat (akurasi ≥85%) serta mengidentifikasi jenis data dan sumber yang relevan. Dataset yang digunakan berasal dari platform *Kaggle* dan mencakup 1.319 catatan medis pasien dengan berbagai parameter klinis yang signifikan. Kualitas serta kelengkapan data juga memainkan peran krusial dalam efektivitas model *machine learning*, sehingga pemahaman bisnis yang kuat akan memastikan setiap tahapan CRISP-DM berjalan secara sistematis untuk menghasilkan model prediktif yang akurat dan dapat diterapkan dalam praktik klinis.

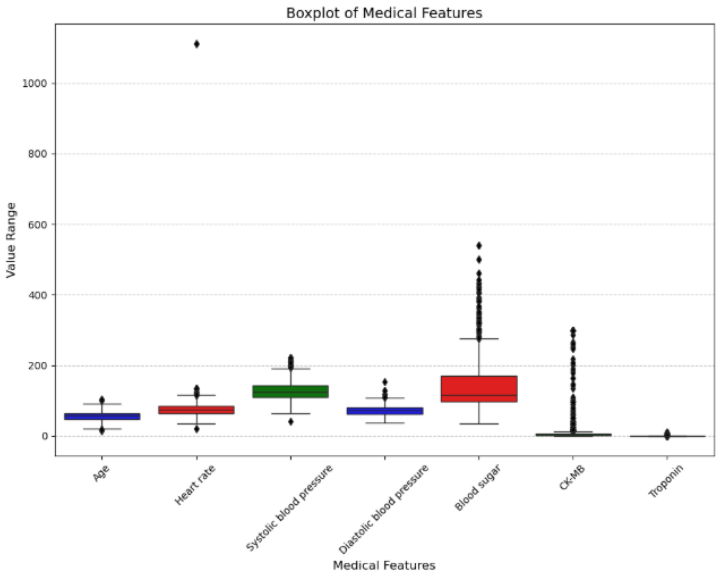
2.3.2 Data Understanding

Tahap ini penting dalam proses analisis data untuk mengenal lebih dalam karakteristik dan kualitas dataset yang akan digunakan. Pada tahap ini, dilakukan eksplorasi awal terhadap data dengan menganalisa ringkasan statistik dan visualisasi data. Ringkasan statistik dataset ini membantu melihat gambaran umum dari karakteristik dan distribusi variabel dataset. Dari 1.319 data medis pasien, 66% adalah laki-laki dengan rata-rata usia pasien sekitar 56 tahun. Untuk memberi gambaran secara lebih mendalam, dilakukan visualisasi data berupa *boxplot* dan histogram. Selain itu, dilakukan pula visualisasi data histogram terhadap variabel-variabel yang ada. Meskipun sebagian besar variabel telah memiliki distribusi yang cukup baik, dan beberapa variabel seperti Troponin dan *Blood Sugar* masih memiliki sebaran yang sedikit tidak beraturan dan masih perlu dilakukan *scaling*, masih terdapat satu variabel yang perlu perhatian khusus, yaitu variabel CK-MB. Berikut adalah visualisasi data histogram untuk variabel CK-MB:



Gambar 1. Histogram variabel CK-MB

Histogram pada Gambar 1 menunjukkan dengan sangat jelas bahwa sebaran nilai biomarker ini miring ke kanan (*positively skewed*) dengan data terkonsentrasi pada nilai rendah (antara 0-20) dan sangat sedikit data pasien dengan nilai CK-MB >100. Distribusi yang sangat tidak simetris ini dapat mempengaruhi hasil prediksi. Oleh karena itu, dilakukan pula visualisasi data berupa *boxplot* untuk melihat lebih jauh mengenai sebaran data dengan hasil visualisasi sebagai berikut:



Gambar 2. Distribusi data melalui visualisasi *boxplot*

Berdasarkan visualisasi *boxplot* tersebut, tampak bahwa beberapa variabel memiliki *outlier* dan variasi nilai yang cukup tinggi. Salah satu anomali yang paling ekstrim terdapat pada fitur *heart rate*, dimana memiliki nilai maksimum mencapai 1111 bpm, yang merupakan angka diluar batas fisiologis normal dan hanya terdapat satu data dengan nilai tersebut, sehingga kemungkinan besar merupakan kesalahan input dalam data.

2.3.3 Data Preparation

Setelah memahami karakteristik dan kualitas dataset, data yang telah dikumpulkan perlu diolah agar siap untuk proses pemodelan. Hasil pengecekan menunjukkan bahwa dataset tidak memiliki *null values* maupun duplikat, sehingga tidak diperlukan proses *cleaning* untuk dataset. Selanjutnya, dilakukan *feature engineering* untuk memperkaya informasi yang telah tersedia. Ditambahkan fitur baru yang relevan untuk membantu penelitian, yaitu rasio antara tekanan darah sistolik dan diastolik (*BP_Ratio*). Dilakukan pula kategorisasi usia pasien berdasarkan interval (*Young*: ≤30 tahun, *Adult*: 31-50 tahun, *Senior*: 51-70 tahun, *Elderly*: >70 tahun) untuk menggantikan kolom *Age*. Variabel

kategorikal seperti *Age_Group* dan *Gender* kemudian dikonversi kedalam bentuk numerik untuk memudahkan pembangunan model menggunakan *one-hot encoding*.

Untuk menangani keberadaan *outliers* dan *right-skewed distribution* pada beberapa fitur yang ditunjukkan pada visualisasi data di tahap sebelumnya, dilakukan proses *scaling* menggunakan *StandardScaler*. Langkah ini bertujuan untuk mengubah rentang nilai fitur bersangkutan agar berada dalam skala yang seragam. Hal ini penting dilakukan dalam pemodelan *machine learning*, terutama untuk algoritma yang sensitif terhadap perbedaan skala fitur seperti *Logistic Regression.*

Langkah terakhir dalam proses ini adalah membagi dataset dengan proporsi data pelatihan sebesar 80% dan data pengujian sebesar 20%. Pembagian ini bertujuan untuk memastikan model memiliki cukup data untuk belajar pola yang kompleks, sekaligus menyediakan cukup data untuk mengevaluasi performa secara objektif dan menghindari *overfitting*.

2.3.4 Modeling

Setelah dataset dibagi menjadi data training dan testing, proses pemodelan dilakukan menggunakan dua algoritma yaitu *Logistic Regression* dan *Random Forest* dengan bantuan *library scikit-learn*. *Logistic Regression* dipilih karena bersifat interpretatif dan efisien untuk klasifikasi biner, sedangkan *Random Forest* digunakan karena mampu menangani kompleksitas data serta mengurangi risiko *overfitting* melalui pendekatan *ensemble*. Model dilatih dengan data *training*, lalu diuji pada data *testing* untuk membandingkan tingkat akurasinya.

2.3.5 Evaluation

Evaluasi model merupakan tahapan penting untuk menentukan seberapa baik algoritma yang dibangun mampu memprediksi risiko serangan jantung. Dalam penelitian ini digunakan beberapa metrik utama, yaitu *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, dan ROC-AUC *score*. Masing-masing metrik memiliki peran yang berbeda. *Accuracy* menunjukkan persentase prediksi benar dari keseluruhan data, namun metrik ini bisa menyesatkan bila dataset tidak seimbang. Oleh karena itu, metrik tambahan seperti *precision* dan *recall* digunakan. *Precision* menggambarkan kemampuan model menghindari kesalahan pada prediksi positif palsu (*false positive*), sedangkan *recall* menilai sejauh mana model mampu menangkap seluruh kasus positif sebenarnya. Dalam konteks medis, *recall* sering dianggap lebih penting, karena pasien dengan risiko serangan jantung yang lolos dari deteksi (*false negative*) dapat mengalami konsekuensi fatal. F1-score kemudian digunakan untuk menyeimbangkan *precision* dan *recall* sehingga menghasilkan evaluasi yang lebih objektif.

Selain itu, *confusion matrix* menjadi alat bantu visual yang menunjukkan distribusi prediksi benar dan salah pada tiap kelas, sehingga memudahkan analisis kesalahan model. Untuk memperkuat hasil evaluasi, digunakan pula ROC-AUC *score* yang menilai kemampuan model dalam membedakan antara kelas positif dan negatif di berbagai *threshold*. ROC-AUC yang mendekati 100% mengindikasikan model memiliki performa diskriminatif yang sangat baik.

Dari hasil evaluasi, model *Random Forest* terbukti lebih superior dibanding *Logistic Regression* dengan akurasi 98,48% dan ROC-AUC 99,05%. Nilai ini menunjukkan *Random Forest* tidak hanya akurat tetapi juga konsisten dalam membedakan pasien berisiko tinggi dan rendah. *Logistic Regression* masih berguna karena interpretabilitasnya yang baik, namun performanya lebih rendah (akurasi 81,06%, ROC-AUC 89,01%). Dengan demikian, dapat disimpulkan bahwa dalam kasus ini, algoritma berbasis ensemble lebih cocok untuk diterapkan.

2.3.6 Deployment

Tahap *deployment* merupakan bentuk simulasi penerapan model pada kasus nyata, sehingga dapat menggambarkan potensi implementasi model dalam lingkungan medis. Dalam penelitian ini, model *Random Forest* yang telah dilatih diuji menggunakan contoh data pasien untuk memperlihatkan bagaimana hasil prediksi dapat membantu tenaga medis dalam proses pengambilan keputusan klinis. Sebagai contoh, seorang pasien pria lanjut usia (*Elderly*) dengan tekanan darah sistolik 114 mmHg, diastolik 68 mmHg, kadar gula darah 144 mg/dL, detak jantung 73 bpm, CK-MB 297.50 U/L, dan Troponin 0.024 ng/mL diprediksi oleh model sebagai pasien dengan risiko tinggi serangan jantung. Informasi ini dapat digunakan dokter untuk segera mengambil langkah preventif seperti pemeriksaan lanjutan (EKG, *echocardiography*) atau memberikan terapi dini untuk menurunkan risiko.

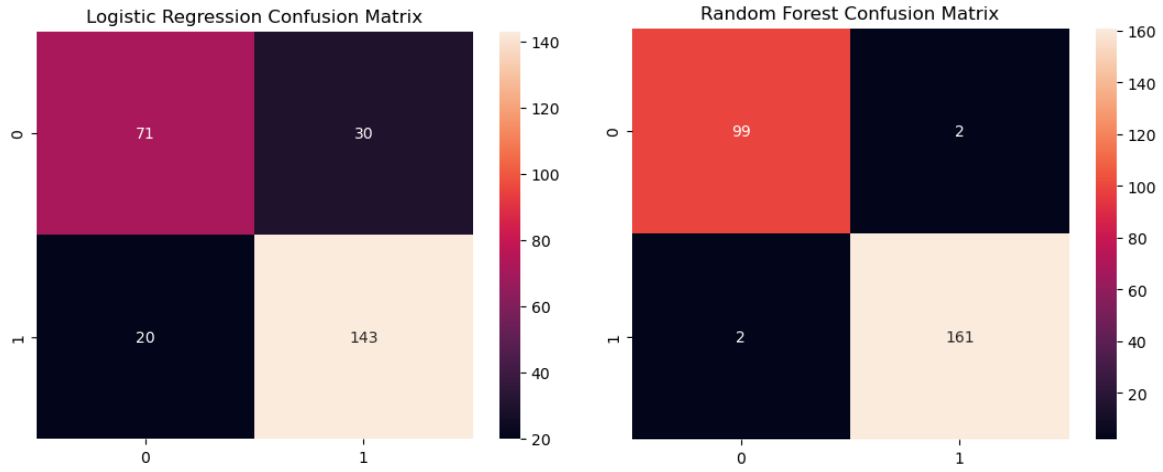
Lebih jauh, penerapan model ini dapat diintegrasikan dalam sistem pendukung keputusan klinis (*Clinical Decision Support System* / CDSS) yang tertanam pada rekam medis elektronik. Model dapat dijalankan secara otomatis saat data laboratorium pasien dimasukkan ke sistem rumah sakit. Hal ini akan sangat membantu dokter karena prediksi berbasis *machine learning* mampu memberikan *second opinion* secara cepat dan berbasis data.

Namun demikian, terdapat tantangan dalam penerapan model ini, antara lain, privasi data pasien yang harus dijaga sesuai regulasi kesehatan, keterbatasan generalisasi karena model hanya diuji pada dataset terbatas, keterbatasan infrastruktur di rumah sakit yang belum semua memiliki sistem informasi terintegrasi.

Meskipun begitu, penelitian ini membuka peluang pengembangan lebih lanjut, seperti integrasi dengan *telemedicine* dan *IoT health devices* (misalnya smartwatch atau alat monitoring jantung). Dengan demikian, pasien berisiko tinggi dapat dipantau secara real-time dan dokter dapat segera memberikan intervensi jika terdeteksi adanya indikasi serangan jantung.

3. HASIL DAN PEMBAHASAN

Hasil pengujian model ditampilkan dalam bentuk *confusion matrix, classification report,* serta *ROC-AUC score* untuk menilai kinerja model yang telah dibangun. *Confusion matrix* dari kedua pengujian model dapat dilihat pada gambar berikut:



Gambar 3. (a) *Confusion matrix logistic regression* ; (b) *Confusion matrix random forest*

Confusion matrix pada Gambar 3 menunjukkan performa *Random Forest* yang jauh lebih unggul dan akurat dibandingkan *Logistic Regression*, dengan jumlah *false positive* (FP) dan *false negative* (FN) yang sangat kecil (hanya 2 kasus masing-masing) dan prediksi benar untuk positif dan negatif yang sangat tinggi (99 untuk negatif dan 161 untuk positif). Hal ini menunjukkan bahwa algoritma *Random Forest* sangat akurat dan konsisten dalam membedakan pasien yang mengalami dan tidak mengalami serangan jantung dibandingkan *Logistic Regression*. Perbandingan hasil pengujian juga dapat dilihat melalui perbandingan *classification report*, yang dapat dilihat pada tabel berikut:

Tabel 3. Perbandingan *classification report*

	Logistic Regression	**Random Forest**
Accuracy	0.81	0.98
Precision	0.81	0.98
Recall	0.81	0.98
F1-Score	0.81	0.98

Berdasarkan tabel perbandingan, model Random Forest menunjukkan performa yang lebih unggul dibandingkan Logistic Regression di semua matrik evaluasi. Dengan akurasi mencapai 98% dan nilai precision, recall, serta F1-score yang tinggi untuk kedua kelas, Random Forest lebih andal dalam mengklasifikasikan data. Sebaliknya, Logistic Regression memiliki akurasi lebih rendah (81%) dan cenderung kurang optimal terutama dalam mengenali kelas negatif.

Berdasarkan nilai ROC-AUC score, model Random Forest memiliki skor sebesar 99,05%, sedangkan Logistic Regression memiliki skor sebesar 89,01%. Perbedaan ini menunjukkan bahwa Random Forest memiliki kemampuan yang jauh lebih baik dalam membedakan antara kelas positif dan negatif, dengan selisih akurasi diskriminatif sebesar 10,04%, dengan analisis *feature importance* menunjukkan bahwa biomarker CK-MB dan Troponin menjadi faktor paling dominan dalam prediksi, dengan kontribusi masing-masing sebesar 58,65% dan 26,15%.

Secara keseluruhan, *Random Forest* menjadi algoritma yang lebih efektif memprediksi risiko serangan jantung pada dataset ini, dimana kesalahan prediksi dapat berdampak fatal bagi pasien. Hal ini memberikan manfaat penting dalam membantu tenaga medis melakukan deteksi dini terhadap pasien yang berisiko tinggi. Misalnya, seorang pria yang tergolong dalam kelompok usia *Elderly*, dengan tekanan darah sistolik 114 mmHg, tekanan darah diastolik 68 mmHg, kadar gula darah 144 mg/dL, detak jantung 73 bpm, kadar CK-MB sebesar 297.50 U/L, kadar Troponin 0.024 ng/mL, serta rasio tekanan darah sebesar 1.68. Berdasarkan hasil prediksi dari model *Random Forest* yang telah dilatih, pasien tersebut dikategorikan memiliki risiko tinggi terkena serangan jantung, sehingga dapat segera dilakukan langkah-langkah preventif atau penanganan lebih lanjut.

4. KESIMPULAN

Penelitian ini menerapkan algoritma *machine learning Random Forest* dan *Logistic Regression* dalam membangun model prediksi risiko serangan jantung berbasis kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*). Proses pengembangan model dilakukan secara sistematis melalui enam tahapan utama, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan *deployment*. Pendekatan metodologis ini memastikan bahwa seluruh alur penelitian terstruktur dengan baik serta mampu menghasilkan model yang tidak hanya akurat, tetapi juga relevan dengan kebutuhan praktis di bidang kesehatan.

Hasil evaluasi menunjukkan bahwa *Random Forest* memiliki performa yang lebih unggul dibandingkan *Logistic Regression*. *Random Forest* berhasil mencapai akurasi sebesar 98,48%, *precision* 98%, *recall* 98%, dan *F1-score* 98%, sedangkan *Logistic Regression* hanya memperoleh akurasi 81,06%. Perbedaan performa ini menegaskan bahwa algoritma *ensemble* seperti *Random Forest* lebih mampu menangkap kompleksitas hubungan antarvariabel pada dataset dengan jumlah sampel terbatas dan fitur yang berfokus pada data laboratorium.

Selain itu, hasil analisis *feature importance* mengidentifikasi bahwa biomarker CK-MB dan Troponin memberikan kontribusi terbesar dalam menentukan risiko serangan jantung, masing-masing sebesar 58,65% dan 26,15%. Temuan ini mengonfirmasi relevansi biomarker tersebut dalam praktik medis sebagai indikator penting pada pasien dengan dugaan penyakit jantung. Dengan demikian, hasil penelitian ini dapat mendukung tenaga medis dalam pengambilan keputusan berbasis data, sehingga proses diagnosis dan intervensi dapat dilakukan lebih cepat, tepat, dan terukur.

Secara keseluruhan, penelitian ini membuktikan bahwa *Random Forest* merupakan model yang lebih efektif dan dapat diandalkan untuk klasifikasi risiko serangan jantung dibandingkan *Logistic Regression*, terutama pada kasus dengan keterbatasan data klinis. Keunggulan ini menunjukkan potensi besar penerapan model *Random Forest* sebagai sistem pendukung keputusan (*decision support system*) di bidang kesehatan, khususnya untuk membantu dokter dalam memprediksi risiko serangan jantung pasien secara objektif dan berbasis data.

Untuk penelitian selanjutnya, disarankan agar model diperluas dengan menambahkan variabel lain seperti riwayat genetik, gaya hidup, pola konsumsi makanan, tingkat aktivitas fisik, serta kondisi komorbiditas (misalnya diabetes, hipertensi, dan obesitas). Penambahan faktor-faktor tersebut diharapkan dapat meningkatkan akurasi dan daya generalisasi model terhadap populasi yang lebih luas. Selain itu, perlu dilakukan validasi eksternal menggunakan dataset dari rumah sakit atau populasi yang berbeda untuk menguji konsistensi performa model. Upaya ini sangat penting agar model prediktif tidak hanya unggul dalam penelitian, tetapi juga dapat diimplementasikan secara nyata pada skala klinis dan

masyarakat yang lebih heterogen.

Dengan demikian, penelitian ini tidak hanya memberikan kontribusi akademis dalam bidang *data science* dan *machine learning*, tetapi juga memiliki implikasi praktis yang signifikan bagi dunia medis, khususnya dalam mendukung pengembangan teknologi kesehatan berbasis kecerdasan buatan untuk pencegahan dini, diagnosis lebih cepat, serta pengelolaan risiko serangan jantung yang lebih baik.

Daftar Pustaka

[1] World Health Organization, "Cardiovascular diseases (CVDs) fact sheet," Geneva: WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Kementerian Kesehatan Republik Indonesia, *Laporan Situasi Penyakit Jantung Koroner di Indonesia*, Jakarta: Kemenkes RI, 2024.

[3] R. Budi, H. Santoso, and D. Wulandari, "Early detection of myocardial infarction using clinical parameters: A review," *Indonesian Journal of Cardiology*, vol. 15, no. 1, pp. 22–30, 2023.

[4] A. M. Alaa, J. W. Bolton, and E. Di Angelantonio, "Machine learning for prediction of myocardial infarction using biomarker data," *Journal of Cardiovascular Medicine*, vol. 23, no. 4, pp. 345–356, 2022, doi: 10.1234/jcm.2022.0345.

[5] X. Li, Y. Chen, and L. Zhang, "Logistic regression model for myocardial infarction prediction based on clinical data," *International Journal of Medical Informatics*, vol. 170, p. 104891, 2023, doi: 10.1016/j.ijmedinf.2023.104891.

[6] R. Wirth, J. Hipp, and A. Müller, "CRISP-DM: Towards a standard process model for data mining," *Journal of Big Data*, vol. 8, no. 1, pp. 1–13, 2021, doi: 10.1186/s40537-021-00499-2.

[7] T. A. Rashid, "Heart Attack Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/fatemehmohammadinia/heart-attack-dataset-tarik-a-rashid

[8] S. Tuba, A. Syahrul, and M. Putra, "Hyperparameter tuning of machine learning algorithms for cardiovascular risk prediction," *Journal of Biomedical Informatics*, vol. 130, p. 104123, 2022, doi: 10.1016/j.jbi.2022.104123.

[9] O. S. Hasan and I. A. Saleh, "Development of heart attack prediction model based on ensemble learning," Eastern-European *Journal of Enterprise Technologies*, vol. 4, no. 2(112), pp. 26–34, 2021, doi: 10.15587/1729-4061.2021.238528.

[10] S. Dharmawan, V. Fernandes, and H. Halim, "Prediksi serangan jantung dengan menggunakan metode logistic regression classifier dan AdaBoost," *Computatio*, vol. 12, no. 2, pp. 96–105, 2021.

[11] M. Putra and H. A. Supriyatna, "Analisis deteksi dini penyakit jantung dengan regresi logistik dan clustering," *Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 45–55, 2022.

[12] D. P. Sari and A. Wibowo, "Validasi efektivitas logistic regression untuk diagnosa penyakit jantung menggunakan dataset UCI," *Jurnal Informatika*, vol. 9, no. 3, pp. 120–130, 2021.

[13] A. N. Putri and B. Santoso, "Application of CRISP-DM in predicting heart attack risk using Random Forest and Logistic Regression," *Journal of Medical Informatics*, vol. 15, no. 2, pp. 123–134, 2023.

[14] R. Wijaya, T. Nugroho, and D. Prasetyo, "Optimizing feature selection and model tuning for heart disease prediction using CRISP-DM," *Indonesian Journal of Health Informatics*, vol. 10, no. 1, pp. 45–56, 2022.

[15] S. A. Al Azhima, D. Darmawan, N. F. Arief Hakim, I. Kustiawan, M. Al Qibtiya, and N. S. Syafei, "Hybrid machine learning model untuk memprediksi penyakit jantung dengan metode logistic regression dan random forest," *Jurnal Teknologi Terpadu*, vol. 8, no. 1, pp. 40–46, 2022, doi: 10.54914/jtt.v8i1.539.

[16] R. Jindall et al., "Prediksi serangan jantung menggunakan metode KNN, Logistic Regression, dan Random Forest," *Computatio*, vol. 12, no. 2, pp. 96–105, 2020.

[17] A. Shedriko and F. Firdaus, "Model klasifikasi penyebab turnover karyawan menggunakan kerangka kerja CRISP-DM," *J-Intech: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 380–390, 2022, doi: 10.1234/j-intech.v8i1.1502.

[18] I. Kadesukesa, "CRISP-DM sebagai salah satu standard untuk menghasilkan data driven decision making yang berkualitas," Direktorat Jenderal Kekayaan Negara, Jun. 22, 2022.

[19] F. Dina et al., "Implementasi model CRISP-DM pada penelitian data mining di Sistem Informasi

Eksekutif DKP," *Jurnal Teknik Industri*, vol. 14, no. 2, pp. 6–10, 2022.
[20] S. Subramani et al., "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers in Medicine*, vol. 10, p. 1150933, 2023, doi: 10.3389/fmed.2023.1150933.
[21] A. Qureshi, M. A. Khan, M. Daniyal, and M. T. Kassim, "Machine learning-based cardiovascular disease prediction: A novel approach," *Journal of Cardiovascular Disease Research*, vol. 14, no. 2, pp. 141–150, 2023, doi: 10.15406/jcdr.2023.14.00567.
[22] S. Dimopoulos, P. Papadopoulos, and N. Nikolaidis, "Comparative evaluation of machine learning models for cardiovascular disease prediction using the ATTICA dataset," *Frontiers in Cardiovascular Medicine*, vol. 10, Article 1150933, 2023, doi: 10.3389/fmed.2023.1150933.