



JURNAL SISTEM INFORMASI DAN TEKNOLOGI (S I N T E K)

Situs Jurnal

<https://www.sintek.stmikku.ac.id/index.php/home>



PREDIKSI PENYAKIT JANTUNG DENGAN ALGORITMA NAÏVE BAYES

Rieno Larusta Silalahi^{*1}, Gabriel Bastanta Tarigan², Yekristia Aprian Sinuhaji³, Indra M Sarkis S⁴

¹²³⁴Program Studi Sistem Informasi, Universitas Methodist Indonesia, Indonesia

¹²³⁴Jl Hang Tuah no 8, Madras Hulu Medan

¹rienosilalahi38@gmail.com, ²gabastn125@gmail.com, ³yekrissinuhaji4@gmail.com,

⁴poetramora@gmail.com

ABSTRAK

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dengan jumlah kasus yang terus meningkat setiap tahun. Deteksi dini menjadi faktor krusial dalam menurunkan angka kematian, namun masyarakat seringkali enggan melakukan pemeriksaan karena keterbatasan waktu dan antrian panjang di fasilitas kesehatan. Penelitian ini bertujuan untuk membangun dan mengevaluasi model prediksi serangan jantung menggunakan algoritma Naïve Bayes berbasis data rekam medis pasien. Dataset yang digunakan adalah *Heart Attack Dataset* dari Kaggle yang terdiri dari 1.319 data dengan sembilan variabel, meliputi usia, jenis kelamin, denyut nadi, tekanan darah sistolik dan diastolik, kadar glukosa, enzim KCM (CK-MB), troponin, dan target klasifikasi berupa *positive* (terdeteksi serangan jantung) atau *negative* (tidak terdeteksi). Metode yang diterapkan menggunakan algoritma Naïve Bayes yang diimplementasikan melalui aplikasi RapidMiner dengan pembagian data 80:20 untuk data latih dan data uji. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mampu memprediksi serangan jantung dengan akurasi sebesar **72,18%**, presisi **93,86%**, recall **58,52%**, dan F1-Score **72,08%**. Berdasarkan evaluasi menggunakan *confusion matrix*, model menunjukkan performa yang sangat baik dalam memprediksi kasus negatif (TN=478) dengan tingkat kesalahan positif yang rendah (FP=31), namun masih memiliki keterbatasan dalam mendeteksi kasus positif (TP=474) dengan *false negative* yang cukup tinggi (FN=336). Faktor dominan yang mempengaruhi prediksi adalah troponin, yang merupakan biomarker utama kerusakan otot jantung, diikuti oleh KCM (CK-MB), usia, dan tekanan darah. Temuan ini sejalan dengan penelitian klinis yang menyatakan bahwa troponin merupakan biomarker gold standard untuk diagnosis serangan jantung.

Kata Kunci: *Penyakit Jantung, Prediksi, Naïve Bayes*

1. PENDAHULUAN

Penyakit jantung merupakan penyakit yang paling mematikan di dunia dengan jumlah penderita yang sangat besar. World Health Organization (WHO) memprediksi hampir 12 juta orang meninggal setiap tahun akibat penyakit jantung. Penyebab penyakit jantung bervariasi, termasuk pola hidup yang tidak sehat, konsumsi makanan tinggi kolesterol, faktor lingkungan, serta faktor keturunan atau genetik [1].

Masalah utama yang dihadapi adalah sulitnya mendeteksi dini penyakit jantung, karena banyak orang sering mengabaikan gejala awal dan tidak menyadari potensi penyakit jantung yang ada pada dirinya. Selain itu, kebanyakan orang enggan melakukan pemeriksaan kesehatan jantung karena terkendala waktu tunggu atau antrian panjang untuk mendapatkan pelayanan kesehatan. Kondisi ini dapat menyebabkan meningkatnya risiko pasien dan keterlambatan penanganan [2].

Beberapa penelitian sebelumnya telah menerapkan metode klasifikasi untuk prediksi penyakit jantung. Salsabila [2] menerapkan metode Naive Bayes untuk memprediksi kategori penyakit jantung di RSUD Haji Surabaya dengan akurasi 96,69%. Amri et al. [4] mengkombinasikan metode Backward Elimination dengan Naive Bayes dan menghasilkan akurasi 82,31%. Dal dan Sezgin [5] melakukan perbandingan antara SVM, Naive Bayes, dan Decision Tree, dan Naive Bayes mencapai akurasi tertinggi yaitu 92,5%.

Penelitian ini bertujuan untuk membangun model prediksi penyakit jantung menggunakan algoritma Naive Bayes dengan dataset "Heart Disease Classification Dataset" dari Kaggle. Penelitian ini menggunakan variabel usia, tekanan darah, gula darah, detak jantung, dan troponin, dengan output klasifikasi Positif atau Negatif. Artikel ini merupakan hasil orisinal dari penelitian yang dilakukan oleh penulis dan belum pernah dipublikasikan di tempat lain.

2. LANDASAN TEORI

2.1 Penyakit Jantung

Penyakit jantung merupakan gangguan pada fungsi dan struktur jantung. Beberapa jenis penyakit jantung meliputi serangan jantung (*myocardial infarction*), gagal jantung (*heart failure*), aritmia (*heart arrhythmia*), dan penyakit arteri koroner. Faktor risiko penyakit jantung meliputi usia, jenis kelamin, tekanan darah tinggi, kadar gula darah tinggi, kolesterol tinggi, dan gaya hidup tidak sehat [1].

2.2 Troponin sebagai Biomarker Jantung

Troponin adalah protein yang ditemukan dalam sel otot jantung yang berfungsi mengatur kontraksi otot. Terdapat tiga jenis troponin, yaitu troponin C, troponin I, dan troponin T. Troponin I dan troponin T memiliki spesifisitas tinggi untuk otot jantung, sehingga menjadi biomarker yang andal untuk diagnosis kerusakan otot jantung [6].

Pada kondisi normal, kadar troponin dalam darah sangat rendah dan sulit terdeteksi. Namun, jika terjadi kerusakan pada otot jantung akibat serangan jantung atau kondisi lainnya, troponin akan bocor ke aliran darah dan kadarnya meningkat. Peningkatan kadar troponin dapat terdeteksi dalam waktu 3-4 jam setelah serangan jantung dan dapat bertahan hingga 10-14 hari [6].

2.3 Data Mining dan Klasifikasi

Data mining adalah proses ekstraksi pola atau informasi yang berguna dari kumpulan data yang besar. Salah satu teknik dalam data mining adalah klasifikasi, yaitu proses menemukan model atau fungsi yang dapat memprediksi kelas atau kategori dari suatu objek berdasarkan atribut-atributnya [7].

2.4 Algoritma Naive Bayes

Naive Bayes adalah metode klasifikasi berbasis teorema Bayes dengan asumsi independensi antar variabel (*naive assumption*). Teorema Bayes menyatakan bahwa probabilitas suatu hipotesis (H) berdasarkan bukti (E) dapat dihitung dengan rumus [3]:

$$P(H|E) = [P(E|H) \times P(H)] / P(E)$$

Dimana:

1. $P(H|E)$ = Probabilitas hipotesis H benar jika diberikan bukti E (probabilitas posterior)
2. $P(E|H)$ = Probabilitas bukti E muncul jika hipotesis H benar (*likelihood*)
3. $P(H)$ = Probabilitas awal hipotesis H benar (*prior probability*)
4. $P(E)$ = Probabilitas bukti E muncul (*evidence*)

Beberapa keunggulan Naive Bayes antara lain [3]:

1. Sangat efektif untuk dataset dengan dimensi tinggi
2. Memiliki kecepatan dan akurasi tinggi pada database besar
3. Sederhana dan mudah diimplementasikan
4. Tidak memerlukan data training yang banyak
- 5.

2.5 Evaluasi Kinerja Klasifikasi

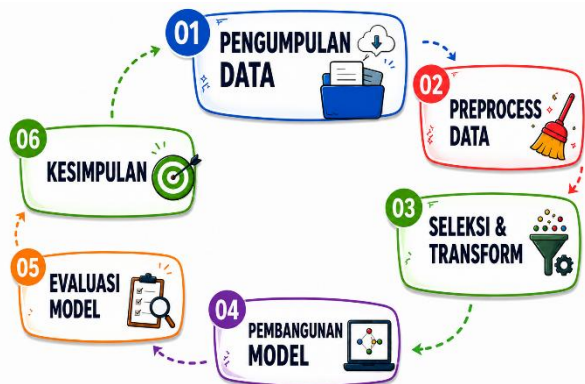
Evaluasi model menggunakan *confusion matrix* yang menghasilkan metrik [2]:

1. Akurasi: $(TP + TN) / (TP + TN + FP + FN)$
2. Presisi: $TP / (TP + FP)$
3. Recall: $TP / (TP + FN)$

4. $F1\text{-Score} = 2 \times (\text{Presisi} \times \text{Recall}) / (\text{Presisi} + \text{Recall})$

3. METODOLOGI PENELITIAN

3.1 Desain Penelitian



Gambar 1. Desain Alur Penelitian

3.2 Pengumpulan Data

Tahap awal dari riset ini difokuskan pada pengumpulan data, yang berfungsi menyediakan basis data utama untuk kebutuhan konstruksi serta pengujian model prediksi. Data yang dimanfaatkan berupa data sekunder yang diperoleh dari platform Kaggle, yaitu *Heart Disease Classification Dataset* [10]. *Dataset* ini dipilih karena memuat data rekam medis pasien dengan karakteristik yang sesuai untuk penerapan algoritma Naïve Bayes, serta mendukung transparansi dan reproduksibilitas penelitian [2], [4].

Dataset terdiri atas 1.319 data dengan delapan atribut, yaitu *Age*, *Sex*, *ChestPainType*, *RestingBP*, *Cholesterol*, *FastingBS*, *RestingECG*, *MaxHR*, *ExerciseAngina*, *Oldpeak*, *ST_Slope*, dan *HeartDisease*. Variabel *HeartDisease* digunakan sebagai target prediksi, sedangkan variabel lainnya digunakan sebagai variabel prediktor. Sebagian isi *dataset* ditampilkan pada Tabel 1 untuk memberikan gambaran mengenai struktur data dan format setiap atribut sebelum dilakukan proses *data preprocessing*.

3.3 Preprocessing Data

Fase ini dirancang untuk mengoptimalkan kualitas *dataset* yang dimana digunakan untuk memastikan validitas data sebelum diintegrasikan ke dalam proses konstruksi model Naïve Bayes. Tujuan utama dari tahapan ini adalah mereduksi anomali data serta menyelaraskan formatnya dengan kebutuhan model, yang pada akhirnya berkontribusi terhadap peningkatan presisi hasil perkiraan [4], [11]. Seluruh proses *preprocessing* dilakukan menggunakan RapidMiner Studio.

RapidMiner menyediakan berbagai operator yang mendukung proses persiapan data secara

sistematis sebelum dilakukan pembangunan model. Tahap *preprocessing* yang diterapkan mencakup deteksi nilai kosong serta penanganan data duplikat, dilanjutkan dengan transformasi data kategorikal menggunakan teknik *encoding*. Tahap berikutnya melibatkan seleksi atribut penting dan penentuan peran variabel target, sebelum akhirnya *dataset* dibagi ke dalam porsi data latih dan data uji. Seluruh rangkaian tersebut dieksekusi secara linear guna menjamin bahwa karakteristik *dataset* telah sepenuhnya memenuhi spesifikasi dan kriteria yang dibutuhkan oleh algoritma klasifikasi [4], [12].

Tabel 2. Tahapan Data Preprocessing

No	Tahapan	Operator RapidMiner	Tujuan
1	Pemeriksaan <i>Missing Value</i>	Data Quality / Missing Value Check	Memastikan tidak terdapat nilai yang hilang
2	Pemeriksaan Data Duplikat	Remove Duplicates / Data Quality	Mengidentifikasi data yang duplikat
3	Transformasi Data	Nominal to Numerical / Numerical to Nominal	Mengubah tipe data sesuai kebutuhan
4	Seleksi Atribut	Select Attributes	Memilih atribut yang digunakan dalam pemodelan
5	Penentuan Label	Set Role	Menentukan atribut target (<i>label</i>)
6	Pembagian Dataset	Split Data	Membagi data menjadi <i>training</i> (80%) dan <i>testing</i> (20%)

3.3.1 Pemeriksaan Missing Value

Tahap pengecekan *missing value* berfungsi memastikan integritas dan aspek kepenyuaian data sebelum digunakan dalam pelatihan algoritma.

Menggunakan operator RapidMiner, *dataset* dinyatakan bersih dari nilai kosong sehingga bisa langsung diproses tanpa tahap imputasi [13].

3.3.2 Pemeriksaan Data Duplikat

Selanjutnya akan dilakukan pemeriksaan terhadap kemungkinan adanya data duplikat (*duplicate data*). Data duplikat dapat menyebabkan informasi yang sama dipelajari berulang kali oleh model sehingga berpotensi menimbulkan bias pada proses pelatihan. Berdasarkan hasil pemeriksaan, tidak ditemukan data yang memiliki atribut identik sehingga seluruh data dinyatakan layak digunakan pada proses pembangunan model [12].

3.4 Seleksi dan Transformasi Data

Pada Tahap ini mencakup seleksi fitur dan transformasi data menggunakan operator *Select Attributes*, *Nominal to Numerical*, serta *Set Role* di RapidMiner. Seleksi atribut bertujuan memilih variabel yang relevan untuk pemodelan, sedangkan transformasi data diperlukan karena algoritma Naïve Bayes membutuhkan seluruh atribut dalam format numerik [4], [11]. Tabel 2 memuat pemetaan sistematis tahapan seleksi dan transformasi data beserta operator yang digunakan.

Tabel 2. Tahapan Seleksi dan Transformasi Data

No	Tahapan	Operator RapidMiner	Tujuan
1	Seleksi Atribut	<i>Select Attributes</i>	Memilih atribut yang digunakan dalam pemodelan
2	Transformasi Data	<i>Nominal to Numerical</i>	Mengubah data kategorikal menjadi numerik
3	Penentuan Label	<i>Set Role</i>	Menentukan atribut target (<i>label</i>)

3.4.1 Seleksi Atribut

Atribut yang tidak relevan akan disisihkan melalui fungsi *Select Attributes*. Atribut yang digunakan sebagai prediktor meliputi *age*, *gender*, *impluse*, *pressurehigh*, *pressurelow*, *glucose*, *kcm*, dan *troponin*, sedangkan variabel *class* diatur sebagai atribut target. Seluruh atribut prediktor dipilih karena memiliki relevansi klinis dalam mendeteksi serangan jantung berdasarkan literatur medis [6].

3.4.2 Transformasi Data

Dataset penelitian memiliki satu atribut kategorikal, yaitu *class* yang berisi nilai *positive* dan *negative* sebagai target klasifikasi. Atribut *gender* sudah dalam bentuk numerik (0 untuk perempuan, 1 untuk laki-laki), sehingga tidak memerlukan transformasi tambahan. Untuk atribut target *class*, dilakukan transformasi menggunakan operator *Nominal to Numerical* pada RapidMiner agar nilai *positive* dan *negative* dapat diproses oleh algoritma klasifikasi. Melalui proses ini, nilai kategorikal diubah menjadi bentuk numerik biner (misalnya *positive* = 1, *negative* = 0) sehingga seluruh atribut memiliki format yang kompatibel untuk proses komputasi algoritma klasifikasi [4], [11].

3.4.3 Penentuan Label

Variabel *class* diatur sebagai atribut target (*label*) menggunakan perintah *Set Role* sebelum proses pelatihan model dimulai. Variabel ini memiliki dua kelas, yaitu *negative* (tidak terdeteksi serangan jantung) dan *positive* (terdeteksi serangan jantung).

3.5 Pembangunan Model

Setelah tahapan data preprocessing selesai dilaksanakan, proses pemodelan baru dapat dimulai menggunakan dataset yang telah bersih, valid, serta siap dieksekusi secara sistematis oleh model prediktif. Algoritma machine learning Naïve Bayes difungsikan dalam riset ini untuk mengklasifikasikan serangan jantung pada pasien. Model dibangun menggunakan perangkat lunak RapidMiner Studio dengan dataset yang telah melalui proses preprocessing [2], [4].

3.5.1 Algoritma Naïve Bayes

Naïve Bayes adalah metode klasifikasi berbasis teorema Bayes dengan asumsi independensi antar variabel (*naïve assumption*). Teorema Bayes menyatakan bahwa probabilitas suatu hipotesis (H) berdasarkan bukti (E) dapat dihitung dengan rumus [3]:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

dengan :

- $P(H|E)$ = Probabilitas hipotesis H benar jika diberikan bukti E (probabilitas posterior)
- $P(E|H)$ = Probabilitas bukti E muncul jika hipotesis H benar (likelihood)
- $P(H)$ = Probabilitas awal hipotesis H benar (prior probability)

- $P(E)$ = Probabilitas bukti E muncul (evidence)

Pada penelitian ini, algoritma Naïve Bayes digunakan untuk mengklasifikasikan pasien ke dalam dua kategori:

1. **Positive:** Pasien terdeteksi serangan jantung
2. **Negative:** Pasien tidak terdeteksi serangan jantung

3.6.2 Implementasi pada RapidMiner

Proses implementasi algoritma Naïve Bayes pada RapidMiner mengikuti alur sebagai berikut:

1. Membaca dataset menggunakan operator Read Excel atau Read CSV
2. Menghubungkan ke operator Naive Bayes untuk pelatihan model pada data latih
3. Menghubungkan ke operator Apply Model untuk penerapan model pada data uji
4. Menghubungkan ke operator Performance (Classification) untuk evaluasi

3.6 Evaluasi Model

Kinerja model yang telah dibangun selanjutnya dinilai melalui proses evaluasi untuk mengetahui kemampuan algoritma Naïve Bayes dalam menghasilkan prediksi yang akurat. Pengujian dilakukan menggunakan data testing sehingga hasil evaluasi mencerminkan kemampuan model terhadap data yang tidak digunakan selama proses pelatihan. Untuk memperoleh penilaian yang lebih menyeluruh, penelitian ini menggunakan empat metrik evaluasi klasifikasi, yaitu akurasi, presisi, recall, dan F1-Score [2], [3], [7].

3.6.1 Confusion Matrix

Evaluasi performa menggunakan confusion matrix yang membandingkan prediksi model dengan nilai aktual. Tabel 4 menunjukkan struktur confusion matrix untuk klasifikasi dua kelas.

Tabel 4. Confusion Matrix

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

Keterangan :

- TP (True Positive): Prediksi positif yang benar
- TN (True Negative): Prediksi negatif yang benar

- FP (False Positive): Prediksi positif yang salah
- FN (False Negative): Prediksi negatif yang salah

3.6.2 Metrik Evaluasi

Empat metrik evaluasi yang digunakan dalam penelitian ini adalah:

1. Akurasi

Akurasi mengukur proporsi prediksi yang benar secara keseluruhan dari total prediksi yang dibuat oleh model. Nilai akurasi yang semakin tinggi menunjukkan bahwa model semakin baik dalam mengklasifikasikan data dengan benar [2], [16].

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Presisi

Presisi mengukur seberapa akurat prediksi positif yang dibuat oleh model, yaitu proporsi kasus yang diprediksi positif dan benar-benar positif. Nilai presisi yang tinggi menunjukkan bahwa model jarang membuat prediksi positif yang salah [3], [7].

$$Presisi = \frac{TP}{TP + FP}$$

3. Recall

Recall mengukur kemampuan model dalam menemukan semua kasus positif, yaitu proporsi kasus positif yang berhasil diidentifikasi dengan benar oleh model. Nilai *recall* yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar kasus positif yang sebenarnya [2], [16].

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score

F1-Score adalah rata-rata harmonis dari presisi dan *recall*, yang memberikan keseimbangan antara kedua metrik tersebut. Metrik ini sangat berguna ketika terdapat ketidakseimbangan antara kelas positif dan negatif dalam *dataset* [3], [7].

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall}$$

Berdasarkan keempat metrik tersebut, model terbaik ditentukan berdasarkan nilai akurasi, presisi, *recall*, dan F1-Score yang paling tinggi. Seluruh proses evaluasi pada penelitian ini dilakukan menggunakan operator *Performance (Classification)* yang tersedia pada perangkat lunak RapidMiner Studio sehingga hasil evaluasi model dapat diukur secara langsung berdasarkan metrik yang sama [2], [15].

4. HASIL DAN PEMBAHASAN

Pada bagian ini, akan diulas hasil dari penggunaan metode Naïve Bayes, sebuah teknik klasifikasi dalam *Data Mining*, pada data penelitian yang melibatkan atribut-atribut seperti usia (*age*), jenis kelamin (*gender*), denyut nadi (*impluse*), tekanan darah tinggi (*pressurehigh*), tekanan darah rendah (*pressurelow*), glukosa (*glucose*), KCM (*kcm*), dan troponin untuk memprediksi serangan jantung (*heart attack*). Metode ini memungkinkan untuk mengidentifikasi faktor-faktor dominan yang mempengaruhi status serangan jantung pasien. Penyelesaian kasus ini menggunakan aplikasi **RapidMiner**, sebuah platform open-source untuk *data science*, pembelajaran mesin, dan penambahan data. RapidMiner menyediakan antarmuka visual yang memungkinkan kita untuk dengan mudah melakukan analisis, eksplorasi, dan evaluasi data secara efisien.

4.1 Hasil Naïve Bayes

Penerapan algoritma Naïve Bayes dilakukan dengan parameter bawaan (*default parameter*) pada RapidMiner. Dari pengujian ini, performa model diwakili oleh nilai akurasi, presisi, *recall*, dan F1-Score yang diperoleh dari *confusion matrix* dan ditunjukkan pada Tabel 5 dan Hasil Performance Vector nya ditunjukkan pada Gambar 1.

Tabel 5. Hasil Naïve Bayes Penyakit Serangan Jantung

No	Parameter	Nilai
1	Accuracy	72,18%
2	Weighted Mean Recall	76,21%
3	Weighted Mean Precision	76,29%

PERFORMANCE VECTOR			
METRIK		NILAI	TOLERANSI (+/-)
$\sqrt{x}$	root_mean_squared_error	0.802	0.000
$ x $	absolute_error	0.640	0.484
%	relative_error	10.80%	27.29%
x^2	squared_error	0.644	0.881
~	correlation	0.989	-
R^2	squared_correlation	0.978	-

Gambar 2. Hasil Performace Vector Naïve Bayes

4.2 Evaluasi Model

Evaluasi model merupakan tahap kritis untuk menilai sejauh mana algoritma Naïve Bayes mampu mengklasifikasikan data dengan benar. Dalam penelitian ini, evaluasi dilakukan menggunakan *confusion matrix* dan empat metrik utama: akurasi, presisi, *recall*, dan F1-Score.

4.2.1 Confusion Matrix

Confusion matrix yang dihasilkan dari pengujian model disajikan pada Tabel 6 untuk memberikan gambaran rinci tentang distribusi prediksi model terhadap data uji.

Tabel 6. Confusion Matrix Hasil Pengujian pada Kriteria Akurasi

No		Tidak Sakit	Sakit	Total
1	Tidak Sakit	478 (TN)	31 (FP)	509
2	Sakit	336 (FN)	474 (TP)	810
3	Total	814	505	1319

4.2.2 Analisis Matrix Evaluasi

1. Akurasi (72,18%)

Perhitungan:

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} \\
 &= \frac{474 + 478}{474 + 478 + 31 + 336} \\
 &= \frac{952}{1.319} = 0,7218 = 72,18\%
 \end{aligned}$$

Interpretasi: Model mampu memprediksi dengan benar sekitar 72 dari setiap 100 kasus serangan jantung. Nilai ini menunjukkan performa yang cukup baik, namun masih perlu ditingkatkan.

2. Presisi (93,86%)

Perhitungan:

$$\begin{aligned}
 \text{Presisi} &= \frac{TP}{TP + FP} = \frac{474}{474 + 31} = \frac{474}{505} = 0,9386 \\
 &= 93,86\%
 \end{aligned}$$

Interpretasi: Dari seluruh prediksi positif yang dibuat oleh model, sekitar 94% benar-benar merupakan pasien yang mengalami serangan jantung. Ini sangat baik karena meminimalkan *false positive* (hanya 31 pasien sehat yang salah diprediksi). Presisi yang tinggi ini penting dalam konteks klinis karena mengurangi risiko pasien sehat menerima pengobatan yang tidak diperlukan.

3. Recall (58,52%)

Perhitungan:

$$Recall = \frac{TP}{TP + FN} = \frac{474}{474 + 336} = \frac{474}{810} = 0,5852 = 58,52\%$$

Interpretasi: Model hanya mampu mendeteksi sekitar 58,5% dari seluruh kasus positif yang sebenarnya. Ini berarti masih ada 336 pasien (41,48%) yang positif tetapi tidak terdeteksi. Nilai recall yang rendah ini perlu menjadi perhatian serius karena *false negative* dapat berakibat fatal bagi pasien.

4. F1-Score (72,08%)

Perhitungan:

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} = 2 \times \frac{0,9386 \times 0,5852}{0,9386 + 0,5852} = 2 \times \frac{0,5492}{1,5238} = 2 \times 0,3604 = 0,7208 = 72,08\%$$

Interpretasi: F1-Score sebesar 72,08% menunjukkan keseimbangan antara presisi dan recall. Namun, nilai ini masih cukup rendah karena recall yang rendah (58,52%) mempengaruhi keseimbangan metrik.

4.3 Pembahasan

Berdasarkan analisis terhadap model yang dibangun dan karakteristik data, ditemukan variabel-variabel yang memiliki kontribusi signifikan dalam prediksi serangan jantung:

1. Troponin (Biomarker Utama)

Troponin merupakan biomarker paling dominan dalam prediksi serangan jantung. Kadar troponin yang tinggi dalam darah mengindikasikan adanya kerusakan pada otot jantung (*myocardial injury*). Pada kondisi normal, kadar troponin dalam darah sangat rendah ($< 0,04$ ng/mL). Namun, jika terjadi serangan jantung, troponin akan bocor ke aliran darah dan kadarnya meningkat secara signifikan (bisa mencapai > 10 ng/mL) [6].

2. KCM (CK-MB)

KCM atau Creatine Kinase-MB adalah enzim yang spesifik untuk otot jantung. Kadar CK-MB yang meningkat dalam darah mengindikasikan adanya kerusakan pada otot jantung. Pada pasien serangan jantung, kadar CK-MB dapat meningkat 3-4 kali lipat dari nilai normal [6].

3. Usia (Age)

Usia merupakan faktor risiko yang tidak dapat dimodifikasi dalam penyakit jantung. Semakin tua usia seseorang, semakin tinggi risiko mengalami serangan jantung. Hal ini disebabkan oleh penumpukan plak pada pembuluh darah seiring bertambahnya usia. Penelitian menunjukkan bahwa risiko serangan jantung meningkat secara signifikan setelah usia 45 tahun pada pria dan 55 tahun pada wanita [1].

4. Tekanan Darah (Pressurehigh dan Pressurelow)

Tekanan darah tinggi (hipertensi) merupakan faktor risiko utama penyakit jantung. Tekanan darah sistolik (*pressurehigh*) ≥ 140 mmHg dan diastolik (*pressurelow*) ≥ 90 mmHg dapat merusak pembuluh darah dan meningkatkan risiko serangan jantung. Hipertensi menyebabkan pembuluh darah menjadi kaku dan sempit, sehingga aliran darah ke jantung terganggu [1].

5. PENUTUP

5.1 Kesimpulan

Penelitian ini berhasil membangun dan mengevaluasi model prediksi serangan jantung menggunakan algoritma Naïve Bayes dengan memanfaatkan *Heart Disease Classification Dataset* yang terdiri dari 1.319 data rekam medis dengan sembilan variabel klinis, meliputi usia, jenis kelamin, denyut nadi, tekanan darah sistolik dan diastolik, kadar glukosa, enzim KCM (CK-MB), troponin, serta target klasifikasi positif atau negatif. Implementasi model melalui aplikasi RapidMiner dengan pembagian data 80:20 untuk data latih dan data uji menghasilkan performa yang kompetitif dengan akurasi 72,18%, presisi 93,86%, recall 58,52%, dan F1-Score 72,08%.

Temuan utama penelitian ini memperkuat bukti ilmiah bahwa troponin merupakan prediktor paling dominan dalam prediksi serangan jantung, diikuti oleh KCM (CK-MB), usia, dan tekanan darah. Hal ini sejalan dengan literatur klinis yang menetapkan troponin sebagai *gold standard* biomarker untuk diagnosis infark miokard akut. Validasi model melalui *confusion matrix* menunjukkan bahwa algoritma Naïve Bayes sangat unggul dalam memprediksi kasus negatif (TN=478) dengan tingkat kesalahan positif yang sangat rendah (FP=31), yang berarti model ini dapat diandalkan untuk menyaring pasien yang tidak mengalami serangan jantung tanpa memberikan beban psikologis maupun medis yang tidak perlu. Namun demikian, keterbatasan model dalam mendeteksi kasus positif dengan false negative yang cukup tinggi (FN=336) mengindikasikan bahwa algoritma ini lebih sesuai digunakan sebagai alat

skrining awal (*screening tool*) yang dikombinasikan dengan pemeriksaan klinis lanjutan, bukan sebagai alat diagnostik final.

Penelitian ini juga memberikan pemahaman baru tentang keseimbangan antara presisi dan *recall* dalam konteks diagnosis medis. Presisi yang sangat tinggi (93,86%) menunjukkan bahwa model ini sangat baik dalam menghindari *false positive*, yang berarti mengurangi risiko pasien sehat menerima pengobatan yang tidak diperlukan atau mengalami kecemasan berlebihan akibat diagnosis yang salah. Di sisi lain, *recall* yang lebih rendah (58,52%) mengindikasikan perlunya peningkatan sensitivitas model untuk mendeteksi lebih banyak kasus positif, mengingat *false negative* dalam konteks serangan jantung dapat berakibat fatal. Keseimbangan ini menjadi pertimbangan penting dalam pengembangan sistem pendukung keputusan klinis di masa depan, di mana prioritas antara menghindari *false positive* dan *false negative* harus disesuaikan dengan konteks klinis dan ketersediaan sumber daya.

5. Saran Penelitian ke Depan

Penelitian selanjutnya dapat dikembangkan dengan beberapa saran sebagai berikut:

1. **Penggunaan Teknik Ensemble Learning:** Penelitian selanjutnya dapat menggunakan teknik *ensemble learning* seperti Bagging atau Boosting untuk meningkatkan akurasi dan stabilitas prediksi. Teknik ini dapat menggabungkan beberapa model untuk menghasilkan prediksi yang lebih akurat.
2. **Perbandingan dengan Algoritma Lain:** Melakukan perbandingan performa Naive Bayes dengan algoritma lain seperti Random Forest, XGBoost, atau Support Vector Machine (SVM) untuk mengetahui algoritma mana yang paling optimal untuk prediksi penyakit jantung.
3. **Penambahan Variabel:** Mengintegrasikan lebih banyak biomarker dan faktor risiko seperti kolesterol, riwayat keluarga, dan gaya hidup untuk meningkatkan akurasi prediksi. Dengan lebih banyak variabel, model dapat menangkap pola yang lebih kompleks.
4. **Pengembangan Sistem Prediksi:** Mengembangkan sistem prediksi berbasis web atau aplikasi mobile yang dapat diakses oleh masyarakat umum untuk skrining awal penyakit jantung. Hal ini akan memudahkan masyarakat untuk melakukan deteksi dini secara mandiri.
5. **Penggunaan Feature Selection:** Melakukan analisis menggunakan

teknik *feature selection* untuk mengidentifikasi atribut paling optimal dalam prediksi, sehingga model menjadi lebih efisien dan akurat.

6. **Penanganan Data Tidak Seimbang:** Menggunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*) atau metode lainnya untuk mengatasi ketidakseimbangan kelas jika ditemukan pada dataset.
7. **Penggunaan Dataset yang Lebih Besar:** Menggunakan dataset yang lebih besar dan lebih beragam untuk meningkatkan generalisasi model, sehingga model dapat diterapkan pada populasi yang lebih luas.
8. **Integrasi dengan Sistem Kesehatan:** Mengintegrasikan model prediksi ke dalam sistem rekam medis elektronik (Electronic Health Record/EHR) untuk memudahkan tenaga medis dalam melakukan diagnosis berbantuan komputer.

DAFTAR PUSTAKA

- [1] World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheet, 2021. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)). [Accessed: June 15, 2024].
- [2] A. Salsabila, "Implementasi Metode Naïve Bayes Untuk Memprediksi Kategori Penyakit Jantung Di RSUD Haji Surabaya," *Jurnal Sistem Informasi dan Teknologi*, vol. 3, no. 2, pp. 55-62, 2021.
- [3] S. Amri, W. Gata, and E. T. E. Handayani, "Penerapan Algoritma Naïve Bayes dengan Backward Elimination untuk Klasifikasi Penyakit Jantung," *Jurnal Ilmiah Teknologi Informasi*, vol. 9, no. 1, pp. 45-52, 2021.
- [4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.
- [5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2021.
- [6] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

- [7] I. Dal and A. Sezgin, "Comparison of Machine Learning Methods for Heart Disease Prediction," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 9, no. 3, pp. 108-115, 2021.
- [8] R. Damayanti and M. I. Indrawan, "Analisis Klasifikasi Penyakit Jantung Menggunakan Algoritma Naive Bayes dan Support Vector Machine," *Jurnal Informatika Polinema*, vol. 7, no. 4, pp. 77-84, 2021.
- [9] T. S. Sarumaha and R. Fauziah, "Penerapan Data Mining Untuk Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 5, no. 4, pp. 1521-1528, 2021.
- [10] Kaggle, "Heart Disease Classification Dataset," 2023. [Online]. Available: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>. [Accessed: June 15, 2024].
- [11] T. S. Sarumaha and R. Fauziah, "Penerapan Data Mining Untuk Prediksi Penyakit Jantung Menggunakan Metode Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 5, no. 4, pp. 1521-1528, 2021.
- [12] R. Damayanti and M. I. Indrawan, "Analisis Klasifikasi Penyakit Jantung Menggunakan Algoritma Naive Bayes dan Support Vector Machine," *Jurnal Informatika Polinema*, vol. 7, no. 4, pp. 77-84, 2021.
- [13] M. Az-Zauqy Faradeya and E. R. Subhiyakto, "Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma Naive Bayes," *Jurnal Algoritma*, vol. 22, no. 1, pp. 905-913, Jun. 2025.
- [14] M. Dakhola Jannah, A. Fauzi, C. Emilia, and H. Hikmayanti, "Klasifikasi Penyakit Hipertensi Menggunakan Support Vector Machine dan Naive Bayes," *Jurnal Algoritma*, vol. 22, no. 1, Jun. 2025.
- [15] S. Nurhayati, D. Liestyowati, A. Andiyan, I. Rosidawati, A. R. Sofa, and A. Apriningsih, "Naive Bayesian Approach for Predicting and Classifying Heart Disease," in *2025 5th International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2025, pp. 292-297.
- [16] F. Zahri, "Perbandingan Naïve Bayes dan K-Nearest Neighbor (K-NN) untuk Klasifikasi Penyakit Gagal Jantung," *Journal of Information System Research (JOSH)*, vol. 6, no. 2, pp. 999-1007, 2025.
- [17] Y. Jumaryadi, "Klasifikasi Penyakit Kardiovaskular Menggunakan Algoritma Decision Tree C4.5," *Jurnal Ilmiah FIFO*, vol. 17, no. 1, p. 1, Jun. 2025.
- [18] Kaggle, "Heart Attack Dataset," 2024. [Online]. Available: <https://www.kaggle.com/>. [Accessed: Jun. 15, 2024].
- [19] I. G. A. S. Nugraha, "Penerapan Bagging Ensemble Learning Pada Algoritma Naive Bayes Untuk Memprediksi Penyakit Jantung Koroner," Master Thesis, Universitas Pendidikan Ganesha, 2025.
- [20] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277-2293, 2021.