SENAFTI
SEMINAR NASIONAL MAHASISWA
FAKULTAS TEKNOLOGI INFORMASI

6th Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
20 September 2025 – Jakarta, Indonesia
Volume 4, Nomor 2, September 2025 - ISSN 2962-8628 (*online*)

IMPLEMENTASI *NAIVE BAYES* DAN *LOGISTIC REGRESSION* UNTUK DIAGNOSIS DINI PENYAKIT JANTUNG

M. Ridhoni[1*], Gandung Triyono[2]

[1,2] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1*]ridhonimuhammad@gmail.com, [2]gandung.triyono@budiluhur.ac.id
(* : corresponding author)

Abstrak- Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga deteksi dini menjadi langkah penting dalam pencegahan dan penanganannya. Penelitian ini bertujuan untuk membandingkan performa dua algoritma *machine learning*, yaitu *Naive Baye*s dan *Logistic Regression*, dalam memprediksi risiko penyakit jantung berdasarkan data klinis pasien. *Dataset* yang digunakan berasal dari platform *Kaggle* dengan judul *Heart Failure Prediction Dataset*, yang berisi 918 data pasien dengan 11 atribut klinis, seperti usia, tekanan darah, kolesterol, serta riwayat angina, dan satu label target (*HeartDisease*). Penelitian ini mengikuti tahapan CRISP-DM, yang mencakup eksplorasi data, *preprocessing*, seleksi fitur, pelatihan model, evaluasi, serta simulasi prediksi data baru. Evaluasi dilakukan menggunakan tiga rasio pembagian data, yaitu 70:30, 80:20, dan 90:10, serta metode 5-fold *cross-validation* untuk memastikan keandalan dan kestabilan hasil. Hasil penelitian menunjukkan bahwa *Naive Bayes* menghasilkan akurasi tertinggi sebesar 87,31% dengan *precision* 91% dan *F1-score* 89%, sedangkan *Logistic Regression* memiliki akurasi 86,59%, dengan lebih tinggi yaitu 92% dan *F1-score* 88% yang lebih stabil. Analisis fitur menunjukkan bahwa *ST_Slope, Exprecision erciseAngina*, dan *Cholesterol* merupakan faktor klinis yang paling berpengaruh terhadap prediksi risiko penyakit jantung. Temuan ini mengindikasikan bahwa *Naive Bayes* lebih unggul dalam hal akurasi keseluruhan, sementara *Logistic Regression* lebih sesuai untuk meminimalkan kesalahan deteksi kasus positif. Secara praktis, hasil penelitian ini menekankan pentingnya menjaga pola hidup sehat, mengontrol kadar kolesterol, dan mewaspadai gejala angina sebagai langkah pencegahan dini terhadap penyakit jantung.

Kata Kunci: Penyakit Jantung, Naive Bayes, Logistic Regression, Klasifikasi, Data Klinis.

IMPLEMENTATION OF NAIVE BAYES AND LOGISTIC REGRESSION FOR EARLY DIAGNOSIS OF HEART DISEASE

Abstract-*Heart disease is one of the leading causes of death in the world, making early detection an important step in its prevention and management. This study aims to compare the performance of two machine learning algorithms, namely Naive Bayes and Logistic Regression, in predicting the risk of heart disease based on clinical data from patients. The dataset used comes from the Kaggle platform titled Heart Failure Prediction Dataset, which contains 918 patient data entries with 11 clinical attributes, such as age, blood pressure, cholesterol, as well as angina history, and one target label (HeartDisease). This research follows the CRISP-DM stages, which include data exploration, preprocessing, feature selection, model training, evaluation, and simulation of predictions on new data. Evaluation is conducted using three data splitting ratios, namely 70:30, 80:20, and 90:10, as well as the 5-fold cross-validation method to ensure the reliability and stability of the results. The research results show that Naive Bayes yields the highest accuracy of 87.31% with a precision of 91% and an F1-score of 89%, while Logistic Regression has an accuracy of 86.59%, with a higher precision of 92% and a more stable F1-score of 88%. Feature analysis indicates that ST_Slope, ExerciseAngina, and Cholesterol are the most influential clinical factors in predicting the risk of heart disease. These findings suggest that Naive Bayes is superior in terms of overall accuracy, while Logistic Regression is more suitable for minimizing false positives. Practically, this research emphasizes the importance of maintaining a healthy lifestyle, controlling cholesterol levels, and being aware of angina symptoms as early preventive measures against heart disease.*

Keywords*: heart disease, naïve bayes, Logistic Regression, classification, clinical data.*

1. PENDAHULUAN

Penyakit jantung merupakan salah satu jenis penyakit tidak menular dengan tingkat mortalitas yang tinggi di dunia. Berdasarkan data *World Health Organization* (WHO), setiap tahunnya lebih dari 17 juta kematian disebabkan oleh penyakit jantung dan pembuluh darah. Di Indonesia, penyakit ini menempati urutan pertama sebagai penyebab kematian dengan jumlah kasus kematian mencapai lebih dari 651.000 jiwa pada tahun 2023 [1].

Situasi ini menjadikan penyakit jantung sebagai permasalahan kesehatan global yang mendesak dan perlu mendapatkan perhatian khusus.

Fenomena ini diperparah oleh pola hidup masyarakat modern yang cenderung tidak sehat, seperti konsumsi makanan tinggi lemak, kurangnya aktivitas fisik, serta tingginya tingkat stres. Kombinasi faktor tersebut menyebabkan risiko penyakit jantung meningkat bahkan pada kelompok usia produktif [2]. Di sisi lain, keterbatasan fasilitas dan biaya dalam diagnosis konvensional menyebabkan banyak kasus penyakit jantung tidak terdeteksi sejak dini, sehingga meningkatkan kemungkinan komplikasi fatal [3].

Seiring berkembangnya teknologi informasi, pendekatan berbasis *machine learning* (ML) mulai dimanfaatkan untuk membantu proses diagnosis penyakit jantung secara lebih cepat dan efisien. Teknik *machine learning* memungkinkan komputer mempelajari pola dari data medis untuk menghasilkan prediksi risiko penyakit yang akurat. Di antara algoritma yang sering digunakan adalah *Naive Bayes* dan *Logistic Regression*. *Naive Bayes* dikenal sebagai algoritma klasifikasi berbasis probabilitas yang sederhana namun efektif untuk *dataset* berskala besar dengan fitur-fitur independen [3]. Sementara itu, *Logistic Regression* merupakan algoritma klasifikasi biner yang banyak digunakan dalam prediksi medis karena modelnya mudah diinterpretasi dan memiliki performa yang baik [4]

Data mining merupakan proses untuk mengekstraksi pengetahuan dari data dalam jumlah besar dengan memanfaatkan teknik statistik, kecerdasan buatan, dan pembelajaran mesin. Salah satu cabang penting dari data mining adalah *machine learning* yang memungkinkan komputer belajar dari data historis untuk melakukan prediksi atau klasifikasi terhadap data baru. Dalam konteks penelitian medis, *machine learning* dapat membantu mengidentifikasi pola risiko penyakit berdasarkan data klinis pasien secara lebih cepat dan akurat dibandingkan metode konvensional [3].

Algoritma *Naïve Bayes* merupakan metode klasifikasi probabilistik yang sederhana namun efektif, terutama untuk dataset dengan jumlah fitur besar dan asumsi independensi antar variabel. Sedangkan *Logistic Regression* adalah model klasifikasi *biner* yang banyak digunakan di bidang medis karena mudah diinterpretasikan, mampu mengukur kontribusi tiap variabel, serta memiliki performa prediksi yang stabil [4].

Berdasarkan fenomena yang ada, permasalahan utama dalam penelitian ini adalah keterbatasan deteksi dini penyakit jantung dengan metode konvensional yang sering kali membutuhkan biaya tinggi, waktu lama, dan belum tentu akurat. Padahal, kebutuhan diagnosis dini sangat mendesak untuk menekan angka kematian akibat penyakit jantung.

Penelitian yang dilakukan menunjukkan bahwa algoritma *Naive Bayes* memiliki akurasi tertinggi yaitu 84,67% dalam klasifikasi penyakit jantung, diikuti oleh *Logistic Regression* dengan akurasi 84,30% [5]. Hasil serupa, yang membandingkan beberapa model dan menyimpulkan bahwa *Logistic Regression* merupakan salah satu metode dengan tingkat akurasi yang stabil dan tinggi [4]. Bahkan, penggabungan *Logistic Regression* dengan *Random Forest* dalam model *hybrid* berhasil meningkatkan akurasi prediksi menjadi 84,48%, menunjukkan bahwa metode ini sangat kompetitif dalam konteks klasifikasi penyakit jantung.

Dari sisi penelitian terdahulu, mayoritas studi membandingkan performa algoritma *machine learning* dalam klasifikasi penyakit jantung, Namun, sebagian besar penelitian sebelumnya hanya menyoroti perbandingan akurasi model tanpa menganalisis kestabilan performa pada berbagai skenario pembagian data serta perbedaan kontribusi fitur klinis antar algoritma. Kesenjangan ini menunjukkan perlunya penelitian yang secara khusus membandingkan *Naive Bayes* dan *Logistic Regression* tidak hanya dari sisi akurasi, tetapi juga konsistensi metrik evaluasi serta faktor klinis yang paling berpengaruh. Oleh karena itu, penelitian ini dilakukan untuk mengisi gap tersebut dan memberikan rekomendasi praktis bagi diagnosis dini penyakit jantung.

Rumusan masalah dalam penelitian ini adalah: (1) Bagaimana performa algoritma *Naïve Bayes* dalam klasifikasi risiko penyakit jantung? (2) Bagaimana performa algoritma *Logistic Regression* dalam klasifikasi risiko penyakit jantung? (3) Algoritma mana yang lebih optimal digunakan sebagai sistem pendukung keputusan klinis? Tujuan penelitian adalah untuk menganalisis dan membandingkan efektivitas kedua algoritma tersebut, sekaligus mengidentifikasi faktor klinis yang paling berpengaruh dalam diagnosis dini penyakit jantung.

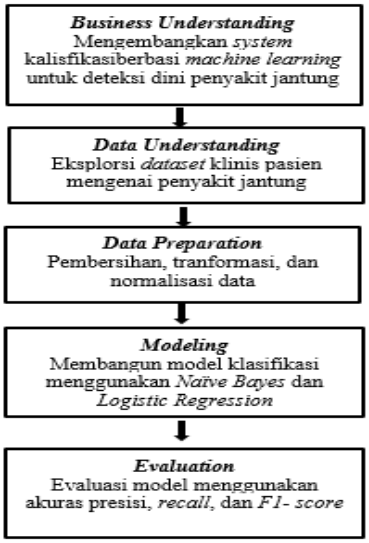
Melihat dari kompleksitas dan dampak besar yang ditimbulkan oleh penyakit jantung, serta potensi algoritma *machine learning* dalam membantu proses diagnosis dini, maka penelitian ini dilakukan untuk mengimplementasikan dan membandingkan dua algoritma klasifikasi yaitu *Naive Bayes* dan *Logistic Regression* dalam memprediksi risiko penyakit jantung berdasarkan data klinis [6]. Diharapkan penelitian ini dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan medis yang lebih cepat, efisien, dan akurat.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksperimen dengan algoritma *machine learning Naïve Bayes* dan *Logistic Regression*. Tahapan penelitian diawali dengan identifikasi masalah, yaitu tingginya

angka kematian akibat penyakit jantung yang membutuhkan solusi diagnosis dini berbasis data klinis. Selanjutnya dilakukan studi literatur untuk meninjau penelitian terdahulu terkait penerapan algoritma klasifikasi pada kasus penyakit jantung. Berdasarkan hasil kajian tersebut, dirumuskan rumusan masalah penelitian, yakni membandingkan performa algoritma *Naïve Bayes* dan *Logistic Regression* dalam diagnosis dini penyakit jantung. Setelah itu dilakukan pengumpulan data, yaitu dengan memanfaatkan *Heart Failure Prediction Dataset* yang bersumber dari platform *Kaggle*, berisi 918 data pasien dengan 11 atribut klinis dan 1 target klasifikasi.

Setelah tahapan awal tersebut, penelitian ini mengacu pada metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri atas pemahaman data, eksplorasi data, persiapan data (*preprocessing*), pemodelan, evaluasi, dan *deployment* awal. Diagram alur tahapan penelitian ditunjukkan pada Gambar 1. [7].



Gambar 1. Tahapan Penelitian

Pada tahap *data preprocessing*, dilakukan *exploratory data analysis* (EDA) untuk melihat distribusi data, mendeteksi anomali, serta memahami korelasi antar variabel. Data kategorikal diubah menjadi numerik melalui *encoding*, sementara data numerik dinormalisasi untuk menjaga skala yang seimbang. Tahap ini dilanjutkan dengan seleksi fitur menggunakan analisis korelasi *Pearson*, serta penilaian *feature importance* berdasarkan koefisien *Logistic Regression* dan probabilitas *Naïve Bayes*. *Dataset kemudian* dibagi menjadi data latih dan uji menggunakan tiga skenario (70:30, 80:20, 90:10) dengan teknik *stratified split*.

Model Naïve Bayes dibangun menggunakan algoritma *GaussianNB*, sementara *Logistic Regression* diimplementasikan dengan fungsi *LogisticRegression* dari pustaka *scikit-learn*. Seluruh pengolahan data, pelatihan model, serta evaluasi dilakukan menggunakan bahasa pemrograman *Python* dengan lingkungan kerja *Jupyter Notebook*. Pustaka pendukung yang digunakan antara lain *pandas* dan numpy untuk pengolahan data, *matplotlib* dan *seaborn* untuk visualisasi, serta *scikit-learn* untuk pemodelan dan evaluasi performa algoritma.

3. HASIL DAN PEMBAHASAN

3.1 *Business Understanding*

Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Deteksi dini risiko penyakit ini menjadi tantangan bagi dunia medis, terutama dalam memastikan akurasi dan efisiensi diagnosis. Dalam konteks ini, penelitian ini dihadapkan pada beberapa permasalahan utama yang perlu diatasi, yaitu:

a. Kebutuhan Deteksi Risiko yang Cepat dan Akurat, sistem kesehatan membutuhkan metode yang dapat mendeteksi risiko penyakit jantung dengan cepat dan akurat berdasarkan data klinis pasien, seperti tekanan darah, kadar kolesterol, dan usia [8].
b. Pemilihan Metrik Evaluasi yang Tepat, untuk memastikan keandalan model prediksi, diperlukan metrik evaluasi yang dapat mengukur performa algoritma, seperti akurasi, *Area Under Curve* (AUC), *precision*, dan *recall* [9].
c. Efisiensi dalam Penggunaan Data, model prediksi harus mampu bekerja secara efektif dengan data medis yang sering kali tidak lengkap atau tidak seimbang, tanpa mengorbankan keakuratan hasilnya. Permasalahan ini mendorong penelitian untuk memanfaatkan algoritma *Machine learning*, khususnya *Naive Bayes* dan *Logistic*

Regression, guna memberikan solusi berbasis data yang dapat mendukung sistem pendukung keputusan klinis (*Clinical Decision Support System*) [10].

d. Kebutuhan *Skateholder*, *stakeholder* utama dalam penelitian ini adalah tenaga medis dan institusi kesehatan. Institusi kesehatan membutuhkan sistem berbasis data yang mampu meningkatkan efisiensi pelayanan, khususnya dalam deteksi dini risiko penyakit jantung. Institusi juga menginginkan solusi yang efisien secara biaya tanpa mengurangi kualitas hasil yang dihasilkan oleh sistem. Selain itu, penerapan teknologi prediktif yang canggih dan berbasis bukti diharapkan dapat meningkatkan kepuasan pasien, sehingga institusi kesehatan dapat memberikan layanan yang lebih optimal dan modern [11].

3.2 Data *Understanding*

a. *Dataset*

Dataset yang digunakan dalam penelitian ini adalah *Heart Failure Prediction Dataset* yang diunduh dari platform *Kaggle* [6]. Dataset ini berisi 918 data pasien dengan 11 atribut klinis dan 1 target klasifikasi (*HeartDisease*). Atribut klinis mencakup usia, jenis kelamin, tipe nyeri dada, tekanan darah istirahat, kadar kolesterol, gula darah puasa, hasil elektrokardiogram, detak jantung maksimum, angina akibat olahraga, nilai *oldpeak*, dan kemiringan segmen ST (*ST_Slope*). Target *HeartDisease* bernilai 1 jika pasien terindikasi penyakit jantung, dan 0 jika tidak.

Tabel 1. *Dataset* Dataset yang digunakan dalam Penelitian (diadaptasi dari *Kaggle Heart Failure Prediction Dataset* [6])

1	Age	40	49	37	48	54
2	*Sex*	M	F	M	F	M
3	*ChestPainType*	ATA	NAP	ATA	ASY	NAP
4	*RestingBP*	140	160	130	138	150
5	*Cholesterol*	289	180	283	214	195
6	*FastingBS*	0	0	0	0	0
7	*RestingECG*	*Normal*	*Normal*	ST	*Normal*	*Normal*
8	*MaxHR*	172	156	98	108	122
9	*ExerciseAngina*	N	N	N	Y	N
10	*Oldpeak*	0	1	0	1.5	0
11	*ST_Slope*	*Up*	*Flat*	*Up*	*Flat*	*Up*
12	*HeartDisease*	0	1	0	1	0

Seperti ditunjukkan pada Tabel 1, fitur *Age* merepresentasikan usia pasien dalam tahun, sementara *Sex* menunjukkan jenis kelamin pasien dengan nilai 0 untuk perempuan dan 1 untuk laki-laki. *ChestPainType* mencatat jenis nyeri dada dengan empat kategori, yaitu ATA, NAP, ASY, dan TA. *RestingBP* menggambarkan tekanan darah pasien saat istirahat dalam satuan mmHg, sedangkan *Cholesterol* mencatat kadar kolesterol dalam darah dalam satuan mg/dL. Selanjutnya, fitur *FastingBS* mengindikasikan kadar gula darah puasa (nilai 1 menunjukkan lebih dari 120 mg/dL, dan 0 sebaliknya). *RestingECG* adalah hasil elektrokardiogram saat istirahat dengan tiga kategori nilai (0–2). *MaxHR* menunjukkan detak jantung maksimum saat aktivitas, sedangkan *ExerciseAngina* mencatat ada atau tidaknya angina dengan nilai 0 untuk tidak dan 1 untuk ya. *Oldpeak* merepresentasikan depresi segmen ST sebagai perubahan dari kondisi istirahat, dan *ST_Slope* mencatat kemiringan segmen ST selama latihan (Up, Flat, atau Down). Akhirnya, *HeartDisease* merupakan label klasifikasi yang menunjukkan apakah pasien terdiagnosis penyakit jantung atau tidak. Informasi dalam Tabel 1 ini menjadi dasar penting dalam membangun model prediksi risiko penyakit jantung berbasis machine learning

b. Statistik Deskriptif

Dalam penelitian ini, dilakukan analisis deskriptif terhadap beberapa fitur penting seperti usia, tekanan darah istirahat, kadar kolesterol, detak jantung maksimum, nilai *Oldpeak*, kadar gula darah puasa (*FastingBS*), serta status penyakit jantung (*HeartDisease*).

Tabel 2. Statistik Deskriptif *Dataset* (Hasil Olahan Menggunakan *Python* [6])

Statistik	*Age*	*Resting BP*	*Cholesterol*	*FastingBS*	*MaxHR*	*Oldpeak*	*HeartDisease*
Count	918	918	918	918	918	918	918
Mean	53.51	132.40	198.80	0.23	136.81	0.89	0.55
Std	9.43	18.51	109.38	0.42	25.46	1.07	0.50
Min	28.00	0.00	0.00	0.00	60.00	-2.60	0.00

25%	47.00	120.00	173.25	0.00	120.00	0.00	0.00
50%	54.00	130.00	223.00	0.00	138.00	0.60	1.00
75%	60.00	140.00	267.00	0.00	156.00	1.50	1.00
Max	77.00	200.00	603.00	1.00	202.00	6.20	1.00

Tabel 2 menampilkan hasil analisis statistik deskriptif terhadap fitur-fitur penting pada dataset, meliputi *Age*, *RestingBP*, *Cholesterol*, *FastingBS*, *MaxHR*, *Oldpeak*, dan *HeartDisease*. Nilai mean dan standar deviasi menggambarkan rata-rata serta sebaran data dari setiap fitur. Terlihat bahwa fitur *Age* memiliki distribusi yang cukup normal dengan nilai mean sekitar 53,5 tahun. Fitur *RestingBP* dan *Cholesterol* masing-masing memiliki nilai minimum 0, yang mengindikasikan adanya potensi anomali atau kesalahan input data. Fitur *Oldpeak* juga memiliki nilai minimum negatif, yaitu -2,6, yang perlu diperhatikan lebih lanjut. Nilai-nilai kuartil (25%, 50%, dan 75%) memberikan gambaran distribusi data dan mendukung pemahaman terhadap sebaran nilai dalam setiap fitur. Informasi dalam Tabel 2 ini sangat berguna sebagai dasar untuk tahap preprocessing dan normalisasi sebelum proses pelatihan model dilakukan.
Analisis statistik deskriptif dan korelasi antar fitur dilakukan menggunakan bahasa pemrograman *Python* dengan bantuan pustaka *pandas* untuk perhitungan nilai statistik, serta *seaborn* dan *matplotlib* untuk visualisasi distribusi data dan *heatmap* korelasi. Hasil analisis menunjukkan bahwa atribut seperti *ST_Slope, ExerciseAngina, Oldpeak, MaxHR,* dan *ChestPainType* memiliki korelasi yang paling signifikan terhadap target *HeartDisease*, sehingga menjadi faktor penting dalam pemodelan.

c. Korelasi antar Fitur

Tabel 3. Korelasi Antar Fitur Numerik

Atribut 1	Atribut 2 (Target)	Nilai Korelasi	Korelasi
ST_Slope	*HeartDisease*	-0.56	Negatif Kuat
MaxHR	*HeartDisease*	-0.40	Negatif Sedang
ChestPainType	*HeartDisease*	-0.39	Negatif Sedang
ExerciseAngina	*HeartDisease*	0.49	Positif Sedang
Oldpeak	*HeartDisease*	0.40	Positif Sedang
Sex	*HeartDisease*	0.31	Positif Rendah
Age	*HeartDisease*	0.28	Positif Rendah
RestingBP	*HeartDisease*	~0.00	Lemah / Tidak Ada
Cholesterol	*HeartDisease*	~0.00	Lemah / Tidak Ada
RestingECG	*HeartDisease*	~0.00	Lemah / Tidak Ada

Tabel 3 menyajikan hasil korelasi antar fitur numerik dengan target *HeartDisease*. Fitur *ST_Slope* memiliki korelasi negatif paling tinggi terhadap penyakit jantung sebesar -0,56, diikuti oleh *MaxHR* sebesar -0,40 dan *ChestPainType* sebesar -0,39. Hal ini mengindikasikan bahwa penurunan nilai pada ketiga fitur tersebut cenderung berhubungan dengan peningkatan risiko penyakit jantung. Di sisi lain, fitur dengan korelasi positif tertinggi adalah *ExerciseAngina* sebesar 0,49 dan *Oldpeak* sebesar 0,40, yang menunjukkan bahwa peningkatan nilai pada kedua fitur ini berkaitan erat dengan risiko jantung lebih tinggi. Fitur *Sex* (0,31) dan *Age* (0,28) juga menunjukkan korelasi positif sedang, menandakan bahwa jenis kelamin dan usia turut memengaruhi kemungkinan seseorang terkena penyakit jantung. Sementara itu, fitur-fitur seperti *RestingBP*, *Cholesterol*, dan *RestingECG* memiliki korelasi mendekati nol, sehingga pengaruh liniernya terhadap *HeartDisease* relatif lemah. Secara keseluruhan, informasi pada Tabel 3 menegaskan bahwa fitur *ST_Slope*, *ExerciseAngina*, *Oldpeak*, *MaxHR*, dan *ChestPainType* merupakan variabel yang paling berpengaruh dalam proses klasifikasi penyakit jantung, sehingga menjadi pertimbangan utama dalam pengembangan model prediksi berbasis data klinis.

3.3 Data *Preparation (Preprocessing)*

Tabel 4. Data Hasil *Preprocessing*

Age	-1.43314	-0.47848	-1.75136	-0.58456	0.051881
Sex	1	0	1	0	1
ChestPainType	1	2	1	0	2
RestingBP	0.410909	1.491752	-0.12951	0.302825	0.951331
Cholesterol	0.82507	-0.17196	0.770188	0.13904	-0.03475
FastingBS	0	0	0	0	0
RestingECG	1	1	2	1	1
MaxHR	1.382928	0.754157	-1.52514	-1.13216	-0.58198

ExerciseAngina	0	0	0	1	0
Oldpeak	-0.83243	0.105664	-0.83243	0.574711	-0.83243
ST_Slope	2	1	2	1	2
HeartDisease	0	1	0	1	0

Tabel 4 menunjukkan sebagian dari isi *dataset heart_preprocessed.csv* yang telah melalui serangkaian transformasi. Beberapa fitur numerik seperti *Age*, *RestingBP*, *Cholesterol*, *MaxHR*, dan *Oldpeak* telah dinormalisasi dengan teknik standardisasi, sehingga nilai-nilainya berada dalam rentang distribusi normal, ditandai dengan angka desimal yang mendekati nol sebagai pusat.

Fitur kategorikal seperti *Sex, ChestPainType, RestingECG, ExerciseAngina*, dan *ST_Slope* telah dikonversi ke dalam bentuk numerik melalui proses *encoding*, agar dapat diproses oleh algoritma pembelajaran mesin seperti *Naive Bayes* dan *Logistic Regression*. Misalnya, pada fitur *Sex*, data dikodekan menjadi 0 untuk perempuan dan 1 untuk laki-laki, sehingga model dapat memahami dan mengolah perbedaan jenis kelamin sebagai variabel numerik.

Sementara itu, fitur C*hestPainType* yang semula berupa data kategorikal dengan label teks seperti 'TA' (*typical angina*), 'ATA' (*asymptomatic*), 'NAP' (*non-anginal pain*), dan 'ASY' (*atypical angina*), diubah menjadi nilai numerik 0, 1, 2, dan 3, masing-masing mewakili kategori nyeri dada tersebut. Transformasi ini penting karena sebagian besar algoritma klasifikasi tidak dapat bekerja secara langsung dengan data teks. Dengan konversi ke bentuk numerik, seluruh atribut dalam *dataset* menjadi homogen secara format, yang memungkinkan proses pelatihan model berlangsung secara optimal.

Selain itu, fitur target yaitu *HeartDisease* tetap dipertahankan dalam bentuk biner, di mana nilai 1 menunjukkan bahwa pasien terindikasi memiliki penyakit jantung, dan 0 menunjukkan sebaliknya. Struktur data yang telah dibersihkan dan dinormalisasi ini memudahkan proses pelatihan model serta meningkatkan konsistensi hasil prediksi. Dengan bentuk data yang seperti ini, model *machine learning* dapat mengenali pola-pola klinis dengan lebih baik dan menghasilkan prediksi yang lebih akurat dalam mendeteksi risiko penyakit jantung.

Seluruh proses preprocessing dilakukan menggunakan bahasa pemrograman *Python* pada lingkungan kerja *Jupyter Notebook*. Proses transformasi data memanfaatkan pustaka *pandas* dan *numpy* untuk manipulasi data, *scikit-learn* untuk *encoding* (*One-Hot Encoding* dan *Label Encoding*), normalisasi (*Min-Max Normalization*), serta pembagian data (*train-test split*). Untuk eksplorasi awal data digunakan *seaborn* dan *matplotlib* yang memvisualisasikan distribusi fitur, *outlier*, serta hubungan antar variabel. Dengan *tools* tersebut, *dataset* menjadi siap untuk digunakan dalam tahap pemodelan.

3.4 Modeling

Dalam penelitian ini digunakan dua algoritma *supervised learning*, yaitu *Naive Bayes* dan *Logistic Regression*, untuk membangun model prediksi penyakit jantung berdasarkan atribut klinis pasien.

a. Evaluasi Performa Model *Naive Bayes* Pada Berbagai Rasio Data

Tabel 5. Performa Model *Naive Bayes* pada berbagai Rasio Data

Rasio Latih:Uji	Akurasi	*Precision* (Positif)	*Recall* (Positif)	*F1-score* (Positif)
70:30	87.31%	91%	87%	89%
80:20	84.24%	88%	84%	86%
90:10	84.78%	87%	87%	87%

Tabel 5 menunjukkan hasil performa model *Naive Bayes* pada tiga rasio pembagian data, yaitu 70:30, 80:20, dan 90:10. Dari hasil evaluasi, model menunjukkan akurasi tertinggi sebesar 87,31% pada rasio 70:30, disertai dengan *precision* sebesar 91% dan *F1-score* sebesar 89% pada kelas positif (penderita penyakit jantung). Hal ini mengindikasikan bahwa pada rasio tersebut, model memiliki keseimbangan yang optimal antara jumlah data latih dan data uji, sehingga mampu belajar pola dengan baik dan menghasilkan prediksi yang stabil.

Sementara itu, pada rasio 80:20, akurasi model menurun sedikit menjadi 84,24%, dengan nilai *F1-score* pada kelas positif sebesar 86%. Penurunan ini masih dalam batas wajar dan menunjukkan bahwa model tetap cukup andal meskipun dengan jumlah data uji yang lebih sedikit. Rasio ini umum digunakan karena dianggap seimbang antara pelatihan dan validasi, dan pada banyak kasus memberikan hasil yang *generalizable*.

Untuk rasio 90:10, model menghasilkan akurasi sebesar 84,78% dan *F1-score* sebesar 87%. Meskipun nilai *precision* dan *recall* cukup tinggi, jumlah data uji yang sangat terbatas pada rasio ini dapat membuat evaluasi menjadi kurang representatif terhadap performa model secara keseluruhan. Hal ini disebabkan oleh potensi bias akibat ukuran sampel uji yang terlalu kecil, yang dapat menyebabkan fluktuasi metrik kinerja yang tidak stabil.

Secara keseluruhan, model *Naive Bayes* menunjukkan performa terbaik pada rasio 70:30, diikuti oleh 90:10 dan 80:20. Namun, dalam konteks praktis dan generalisasi model, rasio 80:20 masih dapat dipertimbangkan sebagai pilihan yang seimbang antara efisiensi pelatihan dan validitas pengujian.

b. Evaluasi Performa Model *Logistic Regression* pada berbagai Rasio Data

Tabel 6. Hasil Evaluasi *Logistic Regression* pada berbagai Rasio Data

Rasio Latih:Uji	Akurasi	*Precision* (Positif)	*Recall* (Positif)	*F1-score* (Positif)
70:30	86,59%	0.92	0.85	0.88
80:20	84,24%	0.91	0.81	0.86
90:10	85,87%	0.89	0.87	0.88

Tabel 5 menyajikan performa algoritma *Logistic Regression* pada tiga skenario rasio data yang berbeda. Model menunjukkan akurasi tertinggi sebesar 86,59% pada rasio 70:30, dengan nilai *F1-score* sebesar 88% pada kelas positif. Ini menunjukkan bahwa model mampu melakukan prediksi dengan presisi dan konsistensi yang baik ketika memperoleh proporsi data latih yang cukup besar, sekaligus data uji yang cukup representatif.
Pada rasio 80:20, akurasi model berada di angka 84,24%, sedikit lebih rendah dibandingkan dengan skenario sebelumnya. Meski begitu, nilai *precision* tetap tinggi sebesar 91%, namun *recall* menurun menjadi 81%, yang berdampak pada penurunan *F1-score*. Penurunan *recall* ini menunjukkan bahwa model cenderung lebih selektif dalam mengklasifikasikan kasus positif, sehingga beberapa kasus penyakit jantung mungkin tidak terdeteksi.
Pada rasio 90:10, akurasi meningkat menjadi 85,87%, dengan nilai *F1-score* sebesar 88%, sama seperti rasio 70:30. Namun, penting untuk dicatat bahwa karena ukuran data uji lebih kecil, maka keakuratan hasil evaluasi pada rasio ini bisa menjadi kurang stabil. Meski *precision* dan *recall* tinggi, jumlah sampel yang terbatas dapat mempengaruhi keandalan evaluasi model secara *general*.
Secara umum, *Logistic Regression* menunjukkan performa yang stabil pada semua rasio data, dengan nilai metrik yang konsisten tinggi. Namun, seperti pada model *Naive Bayes*, rasio 70:30 dapat dianggap paling optimal, karena memberikan keseimbangan yang baik antara pelatihan dan evaluasi, serta didukung oleh ukuran data uji yang cukup untuk validasi performa model.

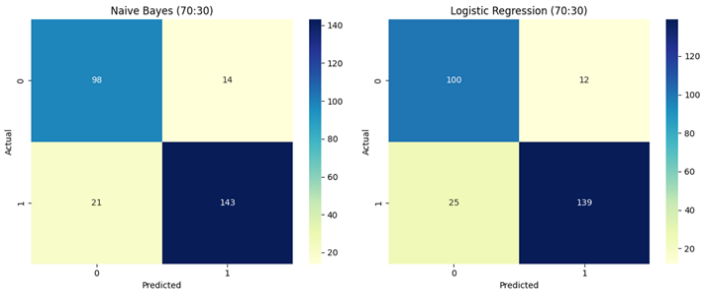
c. Evaluasi Akurasi Menggunakan *Cross-Validation*

Tabel 7. Evaluasi Akurasi *Cross-Validation*

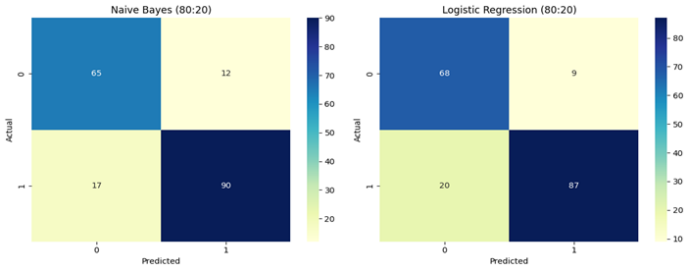
Model	*Fold* 1	*Fold* 2	*Fold* 3	*Fold* 4	*Fold* 5	Rata-rata Akurasi	Standar Deviasi
Naive Bayes	0.842	0.891	0.859	0.809	0.874	0.86	±0.03
Logistic Regression	0.842	0.875	0.864	0.809	0.858	0.85	±0.02

Berdasarkan Tabel 7, model Naive Bayes memperoleh rata-rata akurasi sebesar 86% dengan standar deviasi ±3%, sedangkan model Logistic Regression menunjukkan rata-rata akurasi sebesar 85% dengan standar deviasi ±2%. Nilai standar deviasi yang relatif kecil tersebut mengindikasikan bahwa performa kedua model stabil dan konsisten pada berbagai subset data..
Hasil ini memperkuat kesimpulan sebelumnya bahwa kedua model mampu melakukan klasifikasi penyakit jantung dengan cukup baik, namun *Naive Bayes* cenderung sedikit lebih unggul dalam hal konsistensi dan akurasi rata-rata. *Cross-validation* juga membantu dalam mengidentifikasi potensi *overfitting* yang mungkin tidak terlihat saat menggunakan satu kali pembagian data saja. Dengan demikian, teknik ini sangat berguna untuk memastikan keandalan model sebelum digunakan lebih lanjut dalam implementasi nyata.
Pada tahap modeling, algoritma *Gaussian Naïve Bayes* dan *Logistic Regression* diimplementasikan menggunakan pustaka *scikit-learn*. Parameter *max_iter* dan *random_state* disesuaikan untuk mengoptimalkan performa *Logistic Regression*, sementara *Naïve Bayes* dijalankan dengan asumsi distribusi *Gaussian*. Evaluasi performa model dilakukan dengan menghitung metrik akurasi, *precision, recall*, dan *F1-score* yang tersedia di modul *metrics scikit-learn*. Selain itu, *confusion matrix* divisualisasikan menggunakan *seaborn* untuk menunjukkan distribusi prediksi benar dan salah pada masing-masing kelas.

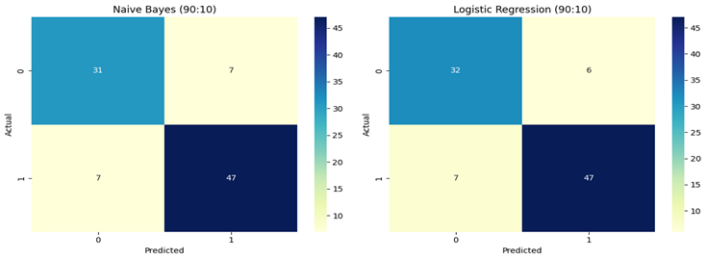
d. *Visualisasi Confusion Matrix*



Gambar 1. Visualisasi *Confusion Matrix Naive Bayes* dan *Logistik Regression* Pada Rasio 70:30



Gambar 2. Visualisasi *Confusion Matrix Naive Bayes* dan *Logistik Regression* Pada Rasio 80:20



Gambar 3. Visualisasi *Confusion Matrix Naive Bayes* dan *Logistik Regression* Pada Rasio 90:10

Gambar 1. menunjukkan bahwa pada rasio pembagian 70:30, model *Naive Bayes* berhasil mengklasifikasikan 98 data negatif dan 143 data positif dengan benar. Terdapat 14 data *false positive* dan 21 data *false negative.* Sementara itu, model *Logistic Regression* menghasilkan 100 prediksi benar untuk kelas negatif dan 139 untuk kelas positif, dengan jumlah *false positive* sebanyak 12 dan false negative sebanyak 25. Meskipun kedua model memberikan hasil yang baik, *Naive Bayes* memiliki sedikit keunggulan dalam mengenali kelas positif, yang menjadi fokus penting dalam prediksi penyakit jantung

Gambar 2. memperlihatkan bahwa pada rasio 80:20, model *Naive Bayes* berhasil mengklasifikasikan 65 data negatif dan 90 data positif secara tepat. Jumlah *false positive* tercatat sebanyak 12 dan *false negative* sebanyak 17. Sementara itu, model *Logistic Regression* mengklasifikasikan 68 data negatif dan 87 data positif dengan benar, disertai 9 false positive dan 20 *false negative*. Secara umum, performa kedua model cenderung seimbang, namun *Logistic Regression* menunjukkan keunggulan dari sisi presisi dalam mengklasifikasikan data negatif, sedangkan *Naive Bayes* memiliki *recall* yang sedikit lebih tinggi untuk kelas positif. Dengan demikian, pemilihan model dapat disesuaikan dengan kebutuhan sistem: apakah meminimalkan kesalahan deteksi positif palsu (*false positive*) atau memaksimalkan deteksi dini terhadap penyakit jantung (*true positive*).

Gambar 3. menunjukkan bahwa pada rasio 90:10, model *Naive Bayes* mengklasifikasikan 31 data negatif dan 47 data positif secara benar, dengan 7 *false positive* dan 7 *false negative*. Sementara itu, model *Logistic Regression* menghasilkan prediksi yang sedikit lebih baik, yaitu 32 benar untuk kelas negatif dan 47 benar untuk kelas positif, dengan hanya 6 *false positive* dan 7 *false negative*. Meskipun kedua model menunjukkan performa yang tinggi pada rasio ini, penting dicatat bahwa ukuran data uji yang kecil (hanya 10% dari total data) membuat evaluasi menjadi kurang representatif. Dengan data uji yang terbatas, model tampak memiliki performa tinggi, namun belum tentu konsisten jika diterapkan pada data yang lebih beragam. Oleh karena itu, meskipun akurasi terlihat tinggi, hasil pada rasio 90:10 perlu diinterpretasikan dengan lebih hati-hati.

e. Analisis Faktor Klinis Paling Berpengaruh

Tabel 8. Faktor Klinis yang Paling Berpengaruh menurut model *Logistic Regression*

Nama Fitur	Koefisien	Pengaruh
ST_Slope	-1.587206	Menurunkan Risiko
ExerciseAngina	1.140776	Meningkatkan Risiko
Sex	1.036052	Meningkatkan Risiko
FastingBS	0.934094	Meningkatkan Risiko
ChestPainType	-0.591502	Menurunkan Risiko
Oldpeak	0.459345	Meningkatkan Risiko
RestingECG	-0.222198	Menurunkan Risiko
Age	0.014821	Meningkatkan Risiko
MaxHR	-0.006034	Menurunkan Risiko
Cholesterol	-0.003665	Menurunkan Risiko
RestingBP	0.002439	Meningkatkan Risiko

Tabel 8 menunjukkan pengaruh masing-masing fitur klinis terhadap probabilitas prediksi pasien menderita penyakit jantung menurut model *Logistic Regression*. Koefisien positif menunjukkan bahwa peningkatan nilai fitur tersebut akan meningkatkan risiko penyakit jantung, sedangkan koefisien negatif menunjukkan efek sebaliknya. Fitur seperti *ST_Slope*, *ExerciseAngina*, dan *Sex* memiliki pengaruh yang paling signifikan.

Tabel 9. Faktor Klinis yang Paling Berpengaruh menurut model *Naïve Bayes*

Nama Fitur	Selisih Rata-rata
Cholesterol	49.578352
MaxHR	19.565027
Age	5.778172
RestingBP	5.047955
Oldpeak	0.937923
ST_Slope	0.682125
ChestPainType	0.680506
ExerciseAngina	0.504858
Sex	0.249044
FastingBS	0.209927
RestingECG	0.066880

Tabel 9. ini menampilkan selisih rata-rata nilai setiap fitur antara kelas positif (menderita penyakit jantung) dan kelas negatif (tidak menderita) menurut model *Naive Bayes*. Semakin besar selisih rata-rata, semakin besar pula kontribusi fitur tersebut dalam membedakan kedua kelas. Fitur *Cholesterol* dan *MaxHR* merupakan dua fitur yang memiliki selisih paling besar dan dengan demikian paling berpengaruh dalam klasifikasi menurut *Naive Bayes*.

3.5 *Evaluation*

Berdasarkan hasil evaluasi pada ketiga rasio pembagian data (70:30, 80:20, dan 90:10), rasio 70:30 terbukti memberikan performa terbaik bagi kedua model. Model *Naive Bayes* pada rasio ini mencatat akurasi 87,31% dengan *F1-score* 0,89, serta keseimbangan antara *precision* dan *recall* yang cukup baik. Model *Logistic Regression* juga menunjukkan performa yang kompetitif, dengan akurasi 86,59% dan *F1-score* 0,88, serta nilai *precision* tertinggi sebesar 92%. Meskipun akurasi terlihat tinggi pada rasio 90:10, ukuran data uji yang kecil membuat hasilnya kurang representatif secara umum.

Hasil evaluasi menunjukkan bahwa model *Naïve Bayes* cenderung lebih unggul pada metrik akurasi, sedangkan *Logistic Regression* memberikan nilai precision yang lebih konsisten. Evaluasi tambahan dengan 5-*fold cross-validation* juga memperlihatkan stabilitas performa kedua model, yang memperkuat kesimpulan penelitian ini.

Pengujian tambahan menggunakan teknik 5-*fold cross-validation* memperkuat hasil tersebut, dengan model *Naive Bayes* mencatat rata-rata akurasi 86% (±3%), dan *Logistic Regression* 85% (±2%). Ini menunjukkan bahwa kedua model bekerja cukup stabil di berbagai kombinasi data, meskipun *Naive Bayes* sedikit lebih unggul dalam hal konsistensi performa dan efisiensi pemrosesan.

Selain evaluasi performa, dilakukan pula analisis terhadap atribut klinis yang paling berpengaruh dalam proses klasifikasi. Berdasarkan model *Logistic Regression, fitur ExerciseAngina, Sex*, dan *FastingBS* memiliki pengaruh terbesar dalam meningkatkan risiko penyakit jantung, sedangkan fitur ST_*Slope* memiliki kontribusi

paling besar dalam menurunkan risiko. Sementara itu, berdasarkan hasil analisis *Naive Bayes* menggunakan selisih rata-rata antar kelas, fitur *Cholesterol, MaxHR*, dan *Age* menunjukkan perbedaan nilai yang paling signifikan, yang berperan penting dalam membedakan pasien yang terindikasi dan tidak terindikasi penyakit jantung.

Dengan demikian, kombinasi penggunaan algoritma *Naive Bayes* pada rasio 70:30 dapat direkomendasikan sebagai pendekatan yang efisien, akurat, dan konsisten dalam membangun sistem prediksi awal penyakit jantung berdasarkan data klinis, serta memberikan pemahaman yang berguna bagi analisis atribut klinis yang berisiko.

4. KESIMPULAN

Penelitian ini mengimplementasikan dan membandingkan dua algoritma klasifikasi, yaitu *Naive Bayes* dan *Logistic Regression*, dalam diagnosis dini penyakit jantung berdasarkan data klinis pasien dengan menggunakan *Heart Failure Prediction Dataset* dari Kaggle yang berisi 918 data pasien. Hasil penelitian menunjukkan bahwa faktor klinis yang paling berpengaruh berbeda pada tiap algoritma, di mana *Logistic Regression* menekankan *ExerciseAngina*, *Sex*, dan *FastingBS* sebagai faktor risiko utama, sementara *Naive Bayes* lebih menonjolkan perbedaan pada *Cholesterol*, *MaxHR*, dan *Age*. Proses pemodelan dilakukan melalui tahapan preprocessing (normalisasi, encoding, analisis korelasi), pengujian dengan rasio pembagian data 70:30, 80:20, dan 90:10, serta evaluasi tambahan menggunakan 5-fold cross-validation. Hasil evaluasi memperlihatkan bahwa *Naive Bayes* unggul dengan akurasi tertinggi 89,1% dan rata-rata 86%, sedangkan *Logistic Regression* mencatat akurasi tertinggi 87,5% dan rata-rata 85%, sehingga keduanya layak dijadikan pendekatan diagnosis dini penyakit jantung dengan keunggulan masing-masing. Penelitian ini menyarankan agar penelitian selanjutnya menguji algoritma lain seperti *Random Forest* atau *Gradient Boosting*, memperbesar ukuran dataset dengan data klinis dari rumah sakit lokal, serta mengembangkan integrasi model ke dalam sistem pendukung keputusan medis berbasis aplikasi yang mudah digunakan oleh tenaga kesehatan maupun masyarakat awam untuk deteksi dini risiko penyakit jantung.

Penelitian selanjutnya disarankan untuk memperluas cakupan dengan menggunakan dataset klinis yang lebih besar dan beragam, termasuk data real-time dari rumah sakit lokal agar model lebih representatif terhadap kondisi populasi yang berbeda. Selain itu, dapat dilakukan pengujian terhadap algoritma lain seperti *Random Forest*, *Support Vector Machine*, atau *Gradient Boosting* yang berpotensi memberikan performa lebih tinggi dalam klasifikasi penyakit jantung. Integrasi model ke dalam sistem pendukung keputusan medis berbasis aplikasi juga perlu dikembangkan, sehingga hasil prediksi dapat dimanfaatkan secara langsung oleh tenaga kesehatan maupun masyarakat awam sebagai alat bantu deteksi dini dan pencegahan risiko penyakit jantung.

DAFTAR PUSTAKA

[1] R. Hidayat, Y. S. Sy, T. Sujana, M. Husnah, H. T. Saputra, And F. Okmayura, "Implementasi *Machine Learning* Untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine," *Bios : Jurnal Teknologi Informasi Dan Rekayasa Komputer*, Vol. 5, No. 2, Pp. 161–168, Sep. 2024, Doi: 10.37148/Bios.V5i2.152.

[2] Y. Amelia, "Perbandingan Metode *Machine Learning* Untuk Mendeteksi Penyakit Jantung," *Idealis : Indonesia Journal Information System*, Vol. 6, No. 2, Pp. 220–225, Jul. 2023, Doi: 10.36080/Idealis.V6i2.3043.

[3] D. Lestari, "Metode Naïve Bayes Dalam *Machine Learning* Untuk Memprediksi Penyakit Jantung Dalam Tubuh," *Jurnal Teknologi Komputer Dan Sistem Informasi) Februari*, Vol. 2, No. 1, Pp. 23–28, 2022, [Online]. Available: Http://Jurnal.Goretanpena.Com/Index.Php/Teknisi

[4] S. A. T. 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, Jul. 2022, Doi: 10.54914/Jtt.V8i1.539.

[5] Ratnasari, A. J. Wahidin, A. E. Setiawan, And P. Bintoro, "*Machine Learning* Untuk Klasifikasi Penyakit Jantung," *Aisyah Journal Of Informatics And Electrical Engineering (A.J.I.E.E)*, Vol. 6, No. 1, Pp. 145–150, Feb. 2024, Doi: 10.30604/Jti.V6i1.272.

[6] A. Jayadie *Et Al.*, *Pembiayaan Kesehatan*, Vol. 11. Media Sains Indonesia, 2023.

[7] C. P. López, *Data Mining. The Crisp-Dm Methodology. The Clem Language And Ibm Spss Modeler*. 2020.

[8] Y. Pintaningrum, B. Rahmat, And R. Ermawan, *Buku Ajar Ilmu Penyakit Jantung*. Mataram: Pt. Percetakan Bali, 2019.

[9] G. Rani And P. K. Tiwari, *Handbook Of Research On Disease Prediction Through Data Analytics And Machine Learning*. Igi Global, 2021. Doi: 10.4018/978-1-7998-2742-9.

[10] I. D. Id, *Machine Learning : Teori, Studi Kasus Dan Implementasi Menggunakan Python*. Pekanbaru: Ur Press, 2021.

[11] G. N. Elwirehardja, T. Suparyanto, And B. Pardamean, "Pengenalan Konsep *Machine Learning* Untuk Pemula," *Publisher: Instiper Press. Isbn: 978-623-5979-10*, Vol. 6, 2023.