JURNAL REKAYASA INFORMATIKA (JRIT),
Vol. 3, No.1 2026, Hal.65-72

ISSN : 3089-5375

Penerapan Data Mining untuk Prediksi Penyakit Jantung Menggunakan Metode Decision Tree

Risky Yanuar Firdaus[1], Heti Mulyani[2]
[1,3]Teknologi Rekayasa Perangkat Lunak, Politeknik Enjinering Indorama, Indonesia
Email Korespondensi: riskyyanuarfirdaus@gmail.com

Info Artikel

Histori Artikel:

Dikirim 20-05-2026
Revisi 29-05-2026
Diterima 01-06-2026

Kata Kunci:
Data Mining
Decision Tree
Penyakit Jantung
Streamlit

ABSTRAK

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia dan menjadi masalah kesehatan yang memerlukan penanganan serta deteksi dini yang efektif. Oleh karena itu, diperlukan sistem yang mampu membantu proses prediksi penyakit jantung secara cepat, akurat, dan efisien. Penelitian ini bertujuan untuk menerapkan teknik data mining menggunakan algoritma Decision Tree dalam membangun sistem prediksi penyakit jantung berbasis web. Metode yang digunakan adalah CRISP-DM (Cross Industry Standard Process for Data Mining) yang terdiri dari enam tahapan, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Dataset yang digunakan diperoleh dari Kaggle dan melalui proses preprocessing untuk meningkatkan kualitas data sebelum dilakukan pemodelan. Model dibangun menggunakan Decision Tree Classifier dan dievaluasi dengan accuracy score, confusion matrix, serta classification report. Hasil penelitian menunjukkan bahwa model mampu mencapai tingkat akurasi sebesar 73,77%. Selanjutnya, model diimplementasikan ke dalam aplikasi berbasis web menggunakan Streamlit sehingga pengguna dapat melakukan prediksi penyakit jantung dengan lebih mudah, cepat, dan interaktif

Article Info

Article history:
Received 20-05-2026
Revised 29-05-2026
Accepted 01-06-2026

Keywords:
Data Mining
Decision Tree
Heart Disease
Streamlit

ABSTRACT

Heart disease is one of the leading causes of death worldwide and is a health problem that requires effective treatment and early detection. Therefore, a system capable of assisting the process of predicting heart disease is needed quickly, accurately, and efficiently. This study aims to apply data mining techniques using the Decision Tree algorithm in building a web-based heart disease prediction system. The method used is CRISP-DM (Cross Industry Standard Process for Data Mining), which consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The dataset used was obtained from Kaggle and went through a preprocessing process to improve data quality before modeling. The model was built using the Decision Tree Classifier and evaluated with an accuracy score, confusion matrix, and classification report. The results showed that the model was able to achieve an accuracy level of 73.77%. Furthermore, the model was implemented into a web-based application using Streamlit so that users can predict heart disease more easily, quickly, and interactively.

1. PENDAHULUAN

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia dan menjadi masalah Kesehatan yang serius hingga saat ini[1]. Penyakit seperti ini dapat dipengaruhi oleh berbagai faktor seperti usia, tekanan darah tinggi, kadar kolestrol, gula darah, polah hidup, serta kondisi Kesehatan yang lain. Proses diagnosis penyakit jantung umumnya memerlukan pemeriksaan secara

medis yang cukup kompleks. Oleh karena itu, diperlukan suatu sistem yang mampu membantu dalam melakukan prediksi penyakit jantung secara lebih cepat dan efisien dengan memanfaatkan teknologi *data mining* dan *machine learning.*

Perkembangan teknologi data mining saat ini memungkinkan proses pengolahan data Kesehatan dilakukan secara otomatis untuk menghasilkan informasi yang bermanfaat dan membantu proses pengambilan keputusan. Salah satu metode klasifikasi yang sering digunakan dalam *data mining* adalah algoritma *decision tree.* Metode ini memiliki kemampuan untuk melakukan klasifikasi data dengan baik serta menghasilkan model keputusan yang lebih mudah dipahami[2]. Selain itu, algoritma *decision tree* juga sering digunakan dalam berbagai penelitian Kesehatan karena mampu menghasilkan tingkat akurasi yang lebih baik dalam proses prediksi penyakit[3].

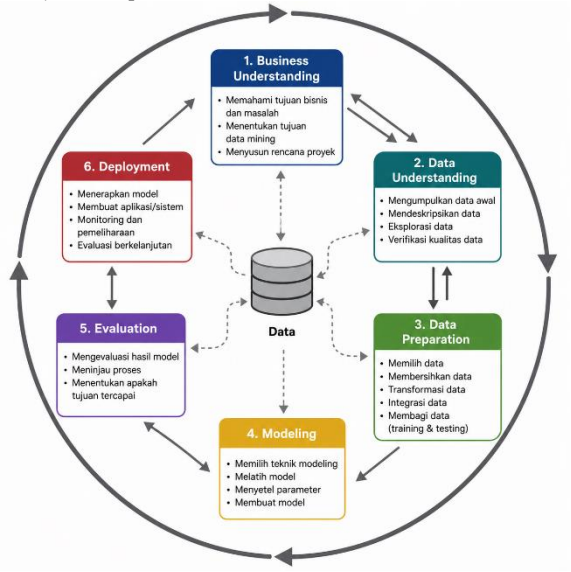
Beberapa penelitian sebelumnya yang telah menerapkan metode *decision tree* dalam prediksi jantung dan menujukkan hasil yang cukup efektif. Seperti contohnya penelitian pada Nafi'ah dan fatah menujukkan bahwa penerapan algoritma *decision tree* pada prediksi penyakit jantung mampu menghasilkan tingkat akurasi yang baik dalam proses klasifikasi data penyakit jantung[4]. Penelitian tersebut menjelaskan bahwa metode *decision tree* efektif digunakan untuk membantu proses analisis data Kesehatan berdasarkan beberapa atribut seperti tekanan darah

Namun, penilitian sebelumnya juga sebagian besar hanya berfokus pada proses klasifikasi data dan evaluasi model tanpa mengimplementasi hasil dari model ke dalam suatu sistem yang dapat digunakan secara langsung oleh pengguna. Oleh karena itu, penelitian memiliki nilai tambah berupa implementasi model prediksi penyakit jantung ke dalam aplikasi berbasis web sehingga lebih mudah digunakan dan diakses.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menerapkan Teknik *data mining* menggunakan algoritman *decision tree* dalam memprediksi penyakit jantung. Keunggulan utama dalam algoritma ini Adalah kemampuannya untuk memvisualisasikan keterkaitan antara variable, sehingga memperudah keputusan dalam situasi yang kompleks[5]. Dataset yang digunakan berasal dari Kaggle dan dilakukan proses preprocessing sebelum model dibangun. Sistem yang dihasilkan diharapkan mampu membantu pengguna dalam melakukan prediksi awal terhadap kemungkinan penyakit jantung berdasarkan data kesehatan pasien secara cepat dan mudah.

2. METODE

Penelitian ini menggunakan metodologi Cross-Industry Standard Process for Data Mining (CRISP-DM) dengan pendekatan analisis prediktif kuantitatif[6]. Metode CRISP-DM dipilih karena memiliki tahapan yang sistematis dalam proses pengolahan data mining, mulai dari pengumpulan data hingga implementasi model ke dalam sistem berbasis web. Tahapan utama dalam metode *CRISP-DM* terdiri dari *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* seperti yang ditunjukkan pada Gambar 1.



Gambar 1 Metode CRISP-DM

2.1. Pengumpulan Data

Tahap awal penelitian dilakukan dengan pengumpulan dataset penyakit jantung yang diperoleh dari platform Kaggle. Dataset tersebut terdiri dari beberapa atribut kesehatan seperti age, sex, chest pain type (cp), trestbps, chol, fbs, thalach, exang, dan atribut lainnya yang digunakan untuk proses klasifikasi penyakit jantung. Variabel target pada dataset terdiri dari dua kelas, yaitu pasien yang memiliki penyakit jantung dan pasien yang tidak memiliki penyakit jantung[7].

2.2. *Business Understanding*

Tahap *Business Understanding* dilakukan untuk memahami permasalahan penelitian terkait prediksi penyakit jantung menggunakan teknik *data mining*. Pada tahap ini dilakukan identifikasi masalah mengenai pentingnya sistem prediksi penyakit jantung yang mampu membantu proses analisis data kesehatan secara lebih cepat dan efisien. Selain itu dilakukan studi literatur dari berbagai penelitian sebelumnya terkait penggunaan algoritma *Decision Tree* dalam klasifikasi data kesehatan. Tujuan penelitian ini adalah membangun model prediksi penyakit jantung menggunakan algoritma *Decision Tree* dan mengimplementasikannya ke dalam aplikasi berbasis web.

2.3. *Data Understanding*

Tahap Data Understanding dilakukan untuk memahami karakteristik dataset sebelum proses pemodelan dilakukan. Analisis data dilakukan menggunakan pendekatan Exploratory Data Analysis (EDA) dengan melakukan analisis statistik deskriptif, visualisasi distribusi data target, dan eksplorasi hubungan antar atribut. Selain itu, dilakukan juga pembuatan pivot table untuk melihat distribusi data berdasarkan beberapa atribut seperti jenis kelamin dan kondisi penyakit jantung. Tahap ini bertujuan untuk memahami pola data yang akan digunakan dalam proses klasifikasi[8].

2.4. *Data Preparation*

Tahap Data Preparation dilakukan untuk mempersiapkan data sebelum proses pelatihan model. Proses ini meliputi pengecekan missing value, pengecekan data duplikat, dan penghapusan data duplikat agar kualitas data menjadi lebih baik. Setelah proses pembersihan data selesai dilakukan, dataset dipisahkan menjadi fitur (X) dan target (y). Selanjutnya data dibagi menjadi data training dan data testing menggunakan metode train_test_split dengan rasio 80% data training dan 20% data testing[9].

2.5. *Modeling*

Tahap *Modeling* dilakukan dengan menerapkan algoritma *Decision Tree Classifier* untuk membangun model prediksi penyakit jantung. Model dilatih menggunakan data training yang telah dipersiapkan sebelumnya. Algoritma *Decision Tree* dipilih karena mampu melakukan proses klasifikasi dengan baik serta menghasilkan struktur keputusan yang mudah dipahami[10]. Setelah proses pelatihan selesai, model digunakan untuk melakukan prediksi terhadap data testing.

2.6. *Evaluation*

Tahap Evaluation dilakukan untuk mengetahui performa model dalam melakukan klasifikasi penyakit jantung. Evaluasi model dilakukan menggunakan beberapa metrik pengujian seperti accuracy score, confusion matrix, dan classification report. Pengujian ini bertujuan untuk mengetahui tingkat akurasi model serta melihat performa prediksi pada masing-masing kelas data.

2.7. *Deployment*

Tahap terakhir adalah Deployment, yaitu implementasi model ke dalam aplikasi berbasis web menggunakan framework Streamlit. Model yang telah selesai dibuat disimpan dalam format .pkl menggunakan library pickle agar dapat digunakan kembali tanpa perlu melakukan proses pelatihan ulang. Pada aplikasi web, pengguna dapat memasukkan data kesehatan pasien melalui form input yang tersedia, kemudian sistem akan menampilkan hasil prediksi penyakit jantung secara otomatis. Implementasi berbasis web ini bertujuan untuk mempermudah penggunaan sistem oleh pengguna secara lebih interaktif dan mudah diakses.

3. HASIL DAN PEMBAHASAN

Pada tahap hasil dan pembahasan dilakukan analisis terhadap dataset penyakit jantung untuk memahami karakteristik data sebelum proses pemodelan dilakukan. Dataset yang digunakan diperoleh dari Kaggle dan terdiri dari beberapa atribut kesehatan seperti age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, dan target. Dataset ini digunakan untuk melakukan klasifikasi kemungkinan pasien memiliki penyakit jantung berdasarkan kondisi kesehatan pasien.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Gambar 2 Dataset Penyakit Jantung

3.1. Hasil Analisis Deskriptif dan EDA

Tahap awal dilakukan analisis statistik deskriptif menggunakan fungsi df.describe() untuk mengetahui karakteristik data numerik. Hasil analisis menunjukkan bahwa dataset memiliki 302 data setelah proses penghapusan data duplikat. Variabel usia memiliki rata-rata sekitar 54 tahun dengan rentang usia antara 29 hingga 77 tahun. Selain itu, nilai kolesterol memiliki rata-rata sebesar 246,5 dan detak jantung maksimum (thalach) memiliki rata-rata sebesar 149,5. Hasil statistik deskriptif menunjukkan bahwa data memiliki variasi nilai yang cukup beragam pada setiap atribut kesehatan.

```
df.describe()
```

1 to 8 of 8 entries Filter

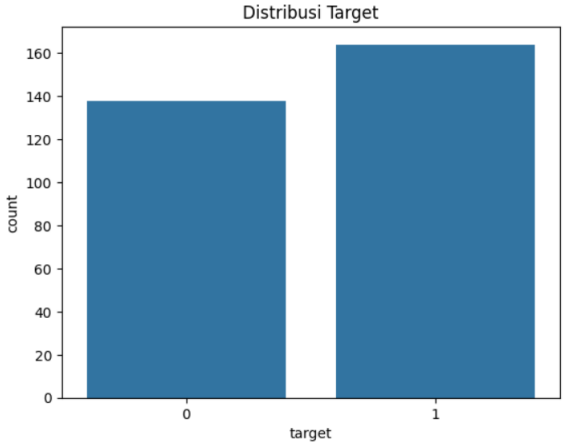
index	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
count	302.0	302.0	302.0	302.0	302.0	302.0	302.0	302.0	302.0
mean	54.420529801324506	0.6821192052980133	0.963576158940397	131.60264900662253	246.5	0.1490066225165563	0.5264900662251656	149.56953642384107	0.32781456953642385
std	9.047969746247 46	0.4664257380672639	1.0320436419542325	17.563394230037563	51.753488655740 56	0.3566860293648132	0.5260271694099755	22.903527251969837	0.4701959640097691 5
min	29.0	0.0	0.0	94.0	126.0	0.0	0.0	71.0	0.0
25%	48.0	0.0	0.0	120.0	211.0	0.0	0.0	133.25	0.0
50%	55.5	1.0	1.0	130.0	240.5	0.0	1.0	152.5	0.0
75%	61.0	1.0	2.0	140.0	274.75	0.0	1.0	166.0	1.0
max	77.0	1.0	3.0	200.0	564.0	1.0	2.0	202.0	1.0

Show 25 per page
Like what you see? Visit the data table notebook to learn more about interactive tables.

Gambar 3 Statistik Deskriptif Dataset

Selanjutnya dilakukan visualisasi distribusi target menggunakan *countplot*. Hasil visualisasi menunjukkan bahwa jumlah pasien yang memiliki penyakit jantung sedikit lebih banyak dibandingkan pasien yang tidak memiliki penyakit jantung. Distribusi data target yang cukup seimbang membantu model dalam melakukan proses klasifikasi secara lebih optimal.

 ISSN : 3089-5375

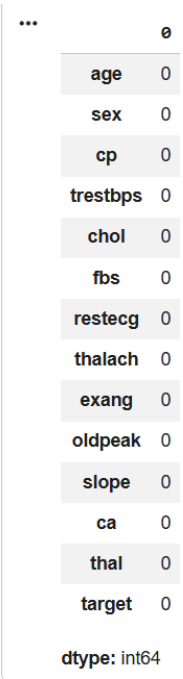


Gambar 4 Distribusi Target Penyakit Jantung

Selain itu, dilakukan analisis menggunakan pivot table untuk melihat distribusi pasien berdasarkan jenis kelamin dan kondisi penyakit jantung. Hasil analisis menunjukkan bahwa jumlah pasien laki-laki lebih dominan dibandingkan perempuan pada dataset yang digunakan.

3.2. Hasil Preprocessing Data

Pada tahap preprocessing dilakukan pengecekan kualitas data menggunakan fungsi isnull().sum() dan duplicated().sum(). Hasil pengecekan menunjukkan bahwa dataset tidak memiliki missing value pada seluruh atribut sehingga data dapat langsung digunakan dalam proses analisis.



	0
age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
target	0

dtype: int64

Gambar 5. Menampilkan duplikasi data

Namun, hasil pengecekan data duplikat menunjukkan terdapat 723 data duplikat pada dataset. Oleh karena itu dilakukan proses penghapusan data duplikat menggunakan fungsi drop_duplicates() agar kualitas data menjadi lebih baik sebelum proses pelatihan model dilakukan. Setelah proses

preprocessing selesai, dataset dipisahkan menjadi fitur (*X*) dan target (*y*). Selanjutnya data dibagi menjadi data training dan data testing menggunakan metode train_test_split dengan rasio 80% data training dan 20% data testing.

3.3. Hasil Pemodelan

Tahap pemodelan dilakukan menggunakan algoritma Decision Tree dengan library DecisionTreeClassifier. Model dilatih menggunakan data training yang telah dipersiapkan sebelumnya. Proses pelatihan model dilakukan menggunakan fungsi fit() untuk mempelajari pola data berdasarkan atribut kesehatan pasien.

```
model = DecisionTreeClassifier()

model.fit(X_train, y_train)
```

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

Gambar 6 Pemodelan decision tree

Algoritma Decision Tree dipilih karena mampu melakukan proses klasifikasi dengan baik dan menghasilkan struktur keputusan yang mudah dipahami. Pada proses pemodelan, atribut seperti chest pain type (*cp*), tekanan darah (*trestbps*), kolesterol (*chol*), dan detak jantung maksimum (*thalach*) menjadi faktor penting dalam proses klasifikasi penyakit jantung.

3.4. Hasil Evaluasi Model

Tahap evaluasi dilakukan menggunakan beberapa metode pengujian seperti *accuracy score*, *confusion matrix*, dan *classification report*. Berdasarkan hasil pengujian, model Decision Tree menghasilkan nilai akurasi sebesar 73,77% pada data testing. Hasil ini menunjukkan bahwa model mampu melakukan klasifikasi penyakit jantung dengan cukup baik berdasarkan atribut kesehatan pasien yang digunakan dalam penelitian. Hasil confusion matrix menunjukkan bahwa model berhasil melakukan prediksi dengan cukup baik pada kedua kelas target. Nilai confusion matrix yang diperoleh adalah:

Aktual	Prediksi Tidak Sakit	Prediksi Sakit
Tidak Sakit	25	7
Sakit	9	20

Berdasarkan hasil classification report, model menghasilkan nilai precision, recall, dan f1-score yang cukup baik pada masing-masing kelas target. Hal ini menunjukkan bahwa algoritma Decision Tree mampu melakukan klasifikasi penyakit jantung secara cukup stabil meskipun masih terdapat beberapa kesalahan prediksi pada data testing.

3.5. Deployment Streamlit

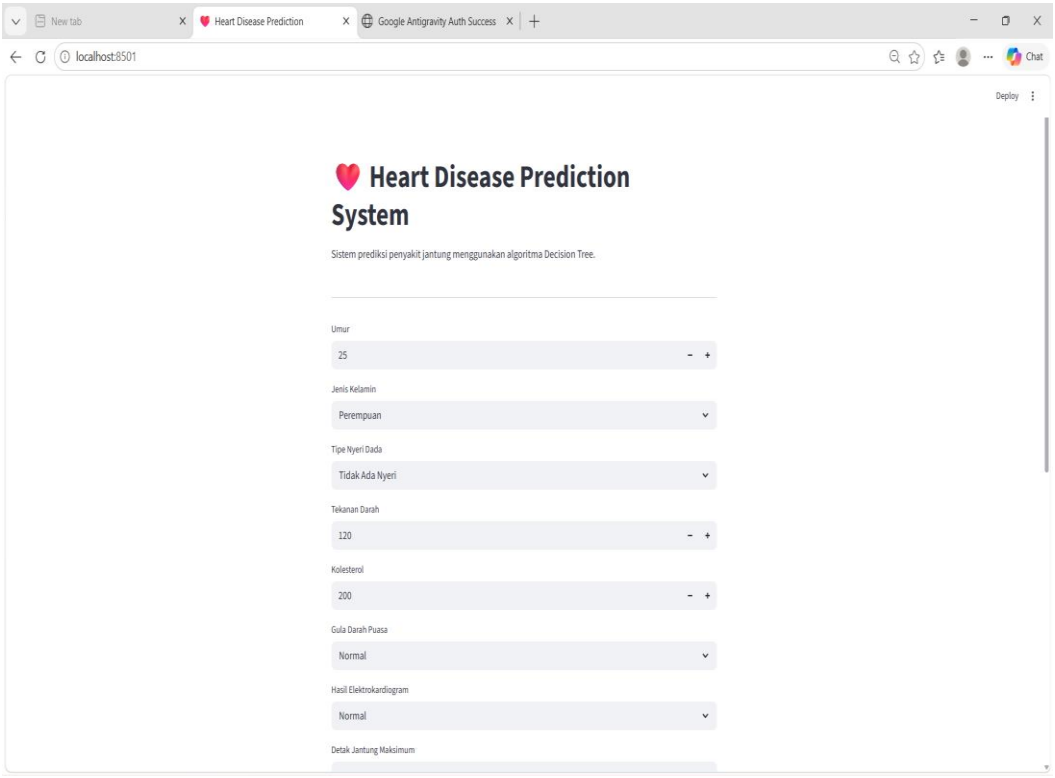
Tahap akhir penelitian adalah implementasi model ke dalam aplikasi berbasis web menggunakan framework Streamlit. Model yang telah selesai dibuat disimpan dalam format .pkl menggunakan library pickle agar dapat digunakan kembali tanpa perlu melakukan proses pelatihan ulang.

Pada aplikasi web, pengguna dapat memasukkan data kesehatan seperti usia, jenis kelamin, tekanan darah, kolesterol, gula darah, detak jantung maksimum, dan atribut kesehatan lainnya melalui form input yang tersedia. Sistem kemudian akan melakukan prediksi secara otomatis berdasarkan

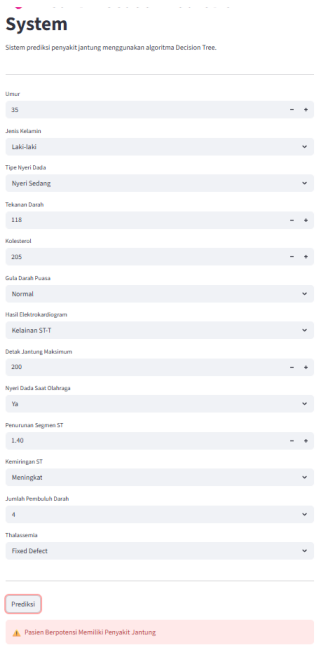
 ISSN : 3089-5375

model Decision Tree yang telah dibuat dan menampilkan hasil prediksi apakah pasien berpotensi memiliki penyakit jantung atau tidak.

Implementasi berbasis web ini mempermudah pengguna dalam melakukan prediksi penyakit jantung secara lebih cepat, mudah, dan interaktif tanpa perlu memahami proses machine learning secara teknis.



Gambar 7 Hasil aplikasi menggunakan streamlit



Gambar 8 Hasil Implementasi Sistem

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, penerapan teknik data mining menggunakan algoritma Decision Tree berhasil digunakan untuk melakukan prediksi penyakit jantung berdasarkan data kesehatan pasien. Proses penelitian dilakukan melalui beberapa tahapan mulai dari pengumpulan data, preprocessing, pemodelan, evaluasi, hingga implementasi sistem berbasis web menggunakan Streamlit. Hasil evaluasi model menunjukkan bahwa algoritma Decision Tree mampu menghasilkan tingkat akurasi sebesar 73,77% dalam melakukan klasifikasi penyakit jantung.

Implementasi sistem berbasis web mempermudah pengguna dalam melakukan prediksi penyakit jantung secara lebih cepat dan interaktif tanpa perlu memahami proses machine learning secara teknis. Dengan adanya sistem ini, pengguna dapat melakukan prediksi awal berdasarkan data kesehatan pasien yang dimasukkan melalui aplikasi web. Penelitian selanjutnya diharapkan dapat menggunakan dataset yang lebih besar serta mencoba algoritma machine learning lainnya seperti Random Forest, Support Vector Machine, atau Neural Network untuk meningkatkan performa model prediksi penyakit jantung.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Politeknik Enjinering Indorama khususnya Program Studi Teknologi Rekayasa Perangkat Lunak yang telah memberikan dukungan dalam proses penyusunan penelitian ini. Penulis juga mengucapkan terima kasih kepada dosen pembimbing, teman-teman, serta semua pihak yang telah membantu memberikan dukungan, masukan, dan motivasi selama proses penelitian dan pengembangan sistem prediksi penyakit jantung ini dilakukan.

DAFTAR PUSTAKA

[1] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques,” *IEEE Access*, vol. PP, p. 1, 2019, doi: 10.1109/ACCESS.2019.2923707.

[2] F. Azuaje, “Review of " Data Mining : Practical Machine Learning Tools and Techniques " by Witten and Frank,” vol. 2, pp. 1–2, 2006, doi: 10.1186/1475-925X-5-51.

[3] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” vol. 2018, 2018.

[4] L. Nafi and Z. Fatah, “Implementasi Algoritma Decision Tree Untuk Pendeteksian Penyakit Jantung,” vol. 3, no. 2, pp. 160–165, 2024.

[5] P. Algoritma, D. Tree, P. Penyakit, P. Valentino, and S. Narulita, “Jurnal Cakrawala Informasi,” vol. 3, no. 2, pp. 18–24, 2023.

[6] P. C. Ncr *et al.*, “Step-by-step data mining guide”.

[7] T. Application *et al.*, “Instal : Jurnal Komputer,” vol. 16, no. June, pp. 120–130, 2024.

[8] T. Edition, *Concepts and Techniques. Morgan Kaufmann, 2011.*

[9] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms,” vol. 2018, 2018, doi: 10.1155/2018/3860146.

[10] V. No *et al.*, “Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4 . 5 Dengan Rapidminer,” vol. 5, no. 2, pp. 73–83, 2023.