Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)DOI: 10.25126/jtiik…
Vol. 1, No. 1, Juni 2025, hlm. 0-5 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

Klasifikasi Penyakit Jantung Menggunakan *Supervised* dan *Unsupervised Learning*

Nadhif Rif `at Rasendriya[1], Fikri Adyatma[2], Reyno Benedict[3]

[1,2,3]Universitas Brawijaya, Malang
Email: [1]nadhifrifat@student.ub.ac.id, [2]fikriadyatma@student.ub.ac.id, [3]benedictreyno@student.ub.ac.id (10pt)
[*]Penulis Korespondensi

(Naskah masuk: 02 Juni 2025, diterima untuk diterbitkan: 03 Juni 2025)

Abstrak

Penelitian ini bertujuan untuk membandingkan model pembelajaran terawasi *(supervised learning*) dan tidak terawasi (*unsupervised learning*) dalam masalah klasifikasi penyakit jantung. Model yang diuji termasuk Random Forest dan K-Nearest Neighbors untuk pembelajaran terawasi, serta K-Means Clustering dan DBSCAN untuk pembelajaran tidak terawasi. Hasil eksperimen menunjukkan bahwa model Random Forest mencapai akurasi tertinggi hingga 97%, sedangkan KNN mencapai 98.54% setelah dilakukan *hyperparameter tuning*. Proses penyetelan *hyperparameter* menggunakan RandomizedSearchCV menunjukkan hasil yang baik dalam menghindari *overfitting*, dengan evaluasi lebih lanjut yang diperlukan untuk validasi. AUC ROC model menunjukkan nilai 0.99, mengindikasikan performa yang sangat baik dalam membedakan antara kelas positif dan negatif. Penelitian ini memberikan wawasan penting tentang bagaimana teknik pembelajaran mesin dapat membantu dalam deteksi dini dan diagnosis penyakit jantung, meskipun validasi lebih lanjut diperlukan sebelum diterapkan di dunia nyata.

Kata Kunci: Klasifikasi Penyakit Jantung, Pembelajaran Mesin, Pembelajaran Terawasi, Pembelajaran Tidak Terawasi, Random Forest, K-Nearest Neighbors, Hyperparameter Tuning

Heart Disease Classification Using Supervised and Unsupervised Learning

Abstract

This study aims to compare the effectiveness of supervised learning and unsupervised learning models in predicting heart disease. The focus is on evaluating the performance of several machine learning algorithms, including Random Forest and K-Nearest Neighbors (KNN) for supervised learning, and K-Means Clustering and DBSCAN for unsupervised learning. A dataset from Kaggle, containing key features such as age, gender, cholesterol levels, and other health indicators, was used for model training and testing. Hyperparameter tuning was performed using RandomizedSearchCV to optimize the models' parameters and avoid overfitting. Among the models tested, Random Forest achieved the highest accuracy of 97%, while KNN demonstrated a notable accuracy of 98.54% (after hyperparameter tuning). The model evaluation was based on performance metrics such as accuracy, confusion matrix, and ROC curve. The results from the confusion matrix revealed that 15 data points were misclassified, while the ROC curve showed an Area Under the Curve (AUC) of 0.99, indicating excellent model performance. This study provides valuable insights into how machine learning techniques can assist in early detection and diagnosis of heart disease, although further validation is needed before real-world deployment.

Keywords: *Heart Disease Classification, Machine Learning, Supervised Learning, Unsupervised Learning, Random Forest, K-Nearest Neighbors, Hyperparameter Tuning*

1. PENDAHULUAN

Penyakit kardiovaskular (CVD) menjadi salah satu penyebab utama kematian di seluruh dunia dengan angka lebih dari 17 juta kematian setiap tahunnya. Angka ini setara dengan 32% dari keseluruhan angka kematian global, menjadikan penyakit kardiovaskular masalah kesehatan masyarakat yang serius dan membutuhkan penanganan yang efektif (Khan Minhas et al., 2024). Deteksi dini penyakit kardiovaskular menjadi sangat penting agar dapat segera menerima penanganan

medis untuk mencegah komplikasi yang lebih parah dari kondisi sebelumnya.

Seiring dengan perkembangan teknologi komputasi, *machine learning* merupakan pendekatan yang menjanjikan dalam pengembangan sistem pendukung *decision making* untuk diagnosis penyakit. Berbagai studi menunjukkan bahwa algoritma *supervised learning* seperti Random Forest dan K-Nearest Neighbors (KNN) mampu menyediakan akurasi tinggi dalam mengklasifikasi kondisi pasien berdasarkan fitur klinis yang dimiliki seperti usia, tekanan darah, kadar kolesterol, detak jantung maksimum, dan riwayat kesehatan lainnya (Saxena & Sharma, 2021). Model tersebut telah membuktikan efektifitasnya dalam mengidentifikasi pola yang tidak mudah dikenali oleh diagnosis manual atau konvensional.

Di lain sisi, *unsupervised learning* juga mulai menjadi pilihan alternatif eksploratif untuk mendeteksi anomali atau *clustering* data pasien tanpa label. Teknik seperti K-Means Clustering dan DBSCAN digunakan untuk menemukan struktur tersembunyi dalam data medis yang dapat mendukung diagnosis atau pengembangan sistem *monitoring* pasien (Mansoori et al., 2020). Meskipun pendekatan ini tidak menggunakan label dalam proses *training*, hasil yang diberikan dapat memberikan gambaran awal dalam kondisi saat data berlabel terbatas atau tidak tersedia.

Berdasarkan latar belakang tersebut, penelitian ini ditujukan untuk membandingkan performa *supervised learning* dan *unsupervised learning* dalam klasifikasi penyakit kardiovaskular. Penelitian ini berfokus dalam mengevaluasi tingkat akurasi, keandalan, dan kemampuan generalisasi dari masing masing pendekatan, juga menilai *tuning* hiperparameter dalam mempengaruhi performa model secara keseluruhan.

2. Metode Penelitian

Penelitian ini dilakukan dengan pendekatan kuantitatif eksperimental untuk mengevaluasi dan membandingkan performa algoritma klasifikasi (*supervised learning*) dan algoritma klastering (*unsupervised learning*) dalam konteks diagnosis penyakit jantung menggunakan dataset terbuka (*Heart Disease Dataset*). Penelitian dibagi ke dalam dua pendekatan utama, yaitu *Supervised Learning* dan *Unsupervised Learning*, masing-masing melalui tahapan *Think*, *Design*, dan *Test* yang merepresentasikan iterasi pemodelan dan evaluasi model.

2.1 Think

a. Pernyataan Masalah dan Asumsi

Penelitian ini berangkat dari permasalahan medis terkait klasifikasi dan eksplorasi pola penyakit jantung. Diketahui bahwa *dataset* yang digunakan memiliki label target (1 = sakit, 0 = tidak sakit), sehingga memungkinkan pendekatan klasifikasi (*supervised learning*). Namun, untuk eksplorasi tambahan, pendekatan *unsupervised learning* juga digunakan.

Permasalahan yang diangkat adalah:

- Bagaimana membangun model klasifikasi terbaik berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score?
- Bagaimana efektivitas model *unsupervised learning* dalam mengidentifikasi pola laten jika label target tidak tersedia?
- Fitur mana yang paling berpengaruh dalam klasifikasi penyakit jantung?

Asumsi awal:

- Random Forest akan menunjukkan performa terbaik karena sifatnya yang *robust* terhadap *overfitting*.
- Fitur seperti “cp” (*chest pain type*), “thal”, dan “ca” memiliki kontribusi besar terhadap prediksi.
- Pada *unsupervised learning*, K-Means lebih stabil dibandingkan DBSCAN karena DBSCAN sensitif terhadap parameter eps dan skala data.

b. Hipotesis

1. Diyakini bahwa model Random Forest dapat mengklasifikasikan penyakit jantung dengan akurasi tinggi. Dianggap benar berdasarkan metrik akurasi dan AUC.
2. Diyakini bahwa K-Means akan menghasilkan klaster yang lebih terdefinisi dibanding DBSCAN. Dianggap benar berdasarkan *silhouette score* dan visualisasi PCA/t-SNE.

c. Proto-Persona

Persona dari penelitian ini adalah praktisi kesehatan atau analis data medis yang ingin:

- Mengidentifikasi pasien berisiko tinggi.
- Melakukan eksplorasi awal tanpa label untuk mendeteksi pola risiko.
- Menggunakan model klasifikasi sebagai alat bantu diagnosis klinis.

2.2 Design

a. Pemodelan *Supervised Learning*

Pada pendekatan ini, dilakukan eksplorasi terhadap tiga algoritma klasifikasi yaitu Logistic Regression, K-Nearest Neighbors (KNN), dan Random Forest. Masing-masing model diuji performanya dengan menggunakan metrik:

- Akurasi
- Presisi
- *Recall*
- F1-Score
- *Confusion Matrix*
- AUC (*Area Under Curve*)

Model dengan akurasi awal tertinggi (Random Forest) kemudian dioptimalkan menggunakan *hyperparameter tuning* dengan teknik RandomizedSearchCV. Hasil tuning menunjukkan bahwa Random Forest menghasilkan akurasi sebesar 0.9268 dan AUC sebesar 0.99.

b. Pemodelan *Unsupervised Learning*
Pada pendekatan ini, digunakan dua algoritma:

- K-Means Clustering dengan 2 klaster
- DBSCAN dengan parameter eps = 1.2 dan min_samples = 5

Model K-Means menghasilkan *Silhouette Score* sebesar 0.1687 dan ARI sebesar 0.376, sedangkan DBSCAN menunjukkan skor negatif (-0.211) dan ARI sebesar 0.0068. Visualisasi dilakukan menggunakan PCA dan t-SNE untuk menilai seberapa baik klaster terbentuk secara spasial.

2.3 Test

a. Evaluasi *Supervised Learning*

Evaluasi model dilakukan dengan membandingkan metrik performa antar model. Random Forest unggul pada semua metrik (akurat dan stabil). Evaluasi stabilitas juga dilakukan dengan menghitung standar deviasi metrik. Random Forest memiliki deviasi terkecil (paling stabil), sedangkan KNN menunjukkan performa tidak stabil dan overfitting.

Confusion matrix digunakan untuk mengidentifikasi jumlah salah klasifikasi (*false positive* dan *false negative*), dan ROC *curve* digunakan untuk melihat kurva performa klasifikasi.

b. Evaluasi *Unsupervised Learning*
Evaluasi dilakukan menggunakan:

- *Silhouette Score*: menilai kekompakan dan pemisahan klaster
- *Adjusted Rand Index* (ARI): mengukur kemiripan hasil klastering terhadap label asli

Visualisasi PCA dan t-SNE digunakan untuk memperkuat interpretasi spasial dari hasil klastering.

2.4 Refleksi dan Kesimpulan

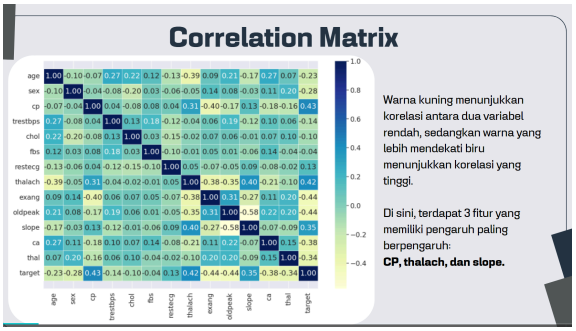
Refleksi terhadap dataset menunjukkan bahwa:

- *Supervised learning* jauh lebih tepat digunakan untuk *dataset* ini karena tersedia label target.
- Random Forest adalah model paling akurat dan stabil, cocok untuk konteks medis karena minim *false positive*.
- KNN memberikan *recall* yang baik, berguna dalam konteks deteksi dini.
- DBSCAN kurang cocok digunakan pada *dataset* ini tanpa tuning parameter lebih lanjut atau teknik reduksi dimensi.

3. Hasil dan Pembahasan

3.1 Iterasi Fase Pertama

Penelitian dimulai pada fase think, yang berfokus pada eksplorasi masalah dan asumsi awal berdasarkan data penyakit jantung. Pada fase ini dilakukan analisis korelasi fitur terhadap target (penyakit jantung) untuk menyusun asumsi dan hipotesis awal. Berdasarkan hasil eksplorasi, ditemukan bahwa fitur-fitur seperti *chest pain type* (cp), *maximum heart rate* (thalach), dan *slope* memiliki korelasi tinggi terhadap target.



Gambar 1. Korelasi antar fitur (*Correlation Matrix*)

Hipotesis awal dalam penelitian ini adalah:

- "Diyakini bahwa model Random Forest dapat mengklasifikasikan penyakit jantung dengan performa tinggi. Dianggap benar berdasarkan metrik akurasi dan AUC."
- "Diyakini bahwa model K-Means menghasilkan klaster yang lebih stabil dan terpisah dibandingkan DBSCAN, berdasarkan visualisasi PCA dan *silhouette score*."

Proses selanjutnya adalah identifikasi karakteristik pengguna atau proto-persona, yaitu:

1. Analis Data Kesehatan, yang bertugas mengevaluasi pasien dengan data klinis historis.

2. Sistem prediksi mandiri, yang dirancang untuk skenario awal ketika label belum tersedia (menggunakan *unsupervised learning*).

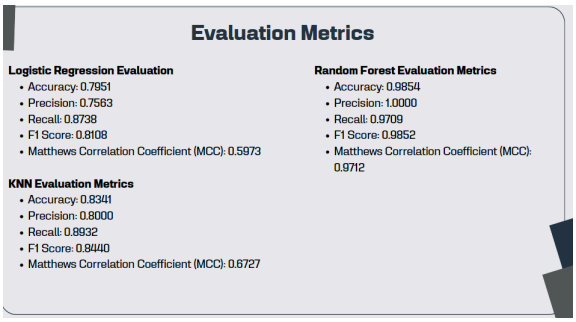
3.2 Iterasi Fase Kedua – Design

a. *Supervised Learning*

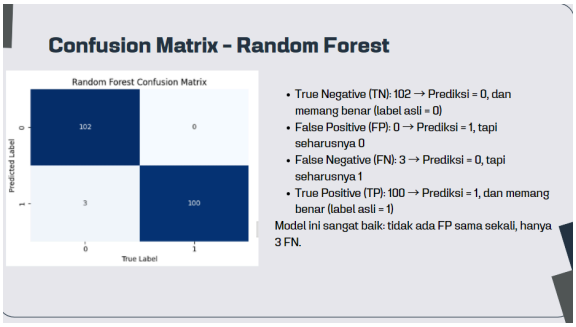
Pada fase desain, dilakukan implementasi 3 model klasifikasi:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Random Forest

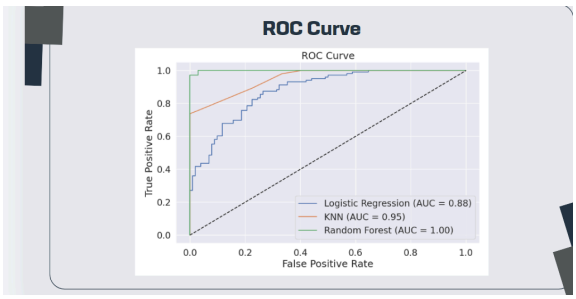
Model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, F1-*score*, dan AUC. Model Random Forest menunjukkan hasil terbaik pada semua metrik:



Gambar 2. Tabel Evaluasi Model *Supervised Learning*



Gambar 3. *Confusion Matrix Random Forest*



Gambar 4. ROC *Curve Random Forest* (AUC = 0.88)

Model Random Forest kemudian dituning menggunakan RandomizedSearchCV untuk menghindari *overfitting*.

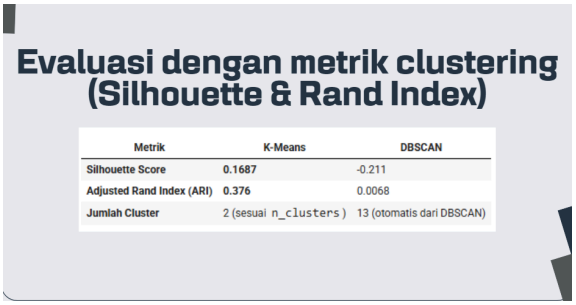
Gambar 5. Hasil *Hyperparameter Tuning Random Forest*

b. *Unsupervised Learning*

Model *unsupervised* yang digunakan adalah:

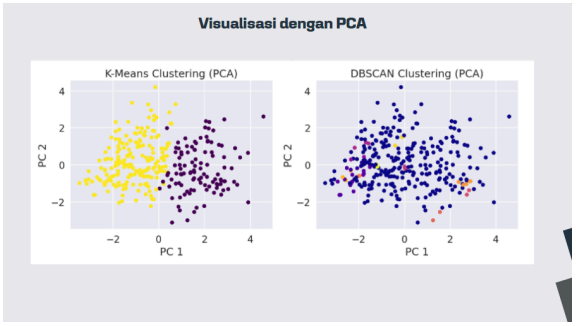
- K-Means Clustering
- DBSCAN

Evaluasi dilakukan menggunakan *Silhouette Score* dan *Adjusted Rand Index* (ARI). Hasil menunjukkan bahwa K-Means lebih baik dalam membentuk klaster yang padat dan terdefinisi.

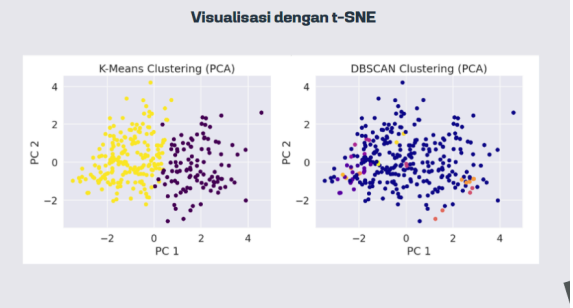


Gambar 6. Hasil *Clustering* KMeans dan DBSCAN (tabel skor)

Visualisasi dilakukan menggunakan PCA dan t-SNE.



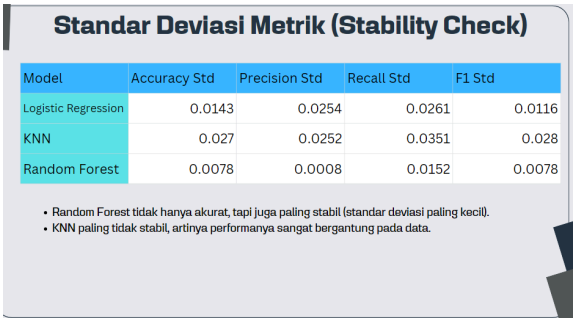
Gambar 7. Visualisasi PCA KMeans vs DBSCAN



Gambar 8. Visualisasi t-SNE *KMeans* vs DBSCAN

3.3 Iterasi Evaluasi dan Refleksi

Evaluasi hasil iterasi pertama dan kedua dilakukan dengan membandingkan stabilitas dan konsistensi antar model. Random Forest menunjukkan deviasi terkecil dan performa paling konsisten:



Standar Deviasi Metrik (Stability Check)

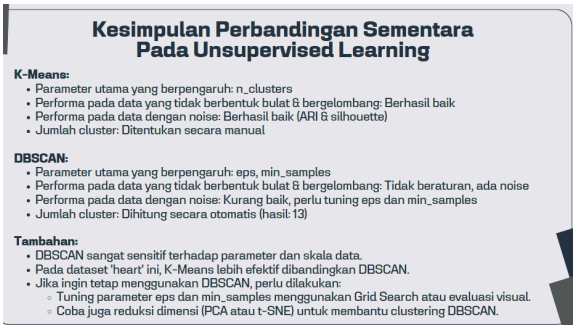
Model	Accuracy Std	Precision Std	Recall Std	F1 Std
Logistic Regression	0.0143	0.0254	0.0261	0.0116
KNN	0.027	0.0252	0.0351	0.028
Random Forest	0.0078	0.0008	0.0152	0.0078

- Random Forest tidak hanya akurat, tapi juga paling stabil (standar deviasi paling kecil).
- KNN paling tidak stabil, artinya performanya sangat bergantung pada data.

Gambar 9. Tabel Standar Deviasi Model

Beberapa catatan evaluasi:

- Logistic Regression unggul pada recall (berguna untuk deteksi awal).
- KNN *overfitting* pada nilai k kecil.
- DBSCAN terlalu sensitif terhadap parameter eps.



Kesimpulan Perbandingan Sementara Pada Unsupervised Learning

K-Means:
- Parameter utama yang berpengaruh: n_clusters
- Performa pada data yang tidak berbentuk bulat & bergelombang: Berhasil baik
- Performa pada data dengan noise: Berhasil baik (ARI & silhouette)
- Jumlah cluster: Ditentukan secara manual

DBSCAN:
- Parameter utama yang berpengaruh: eps, min_samples
- Performa pada data yang tidak berbentuk bulat & bergelombang: Tidak beraturan, ada noise
- Performa pada data dengan noise: Kurang baik, perlu tuning eps dan min_samples
- Jumlah cluster: Dihitung secara otomatis (hasil: 13)

Tambahan:
- DBSCAN sangat sensitif terhadap parameter dan skala data.
- Pada dataset 'heart' ini, K-Means lebih efektif dibandingkan DBSCAN.
- Jika ingin tetap menggunakan DBSCAN, perlu dilakukan:
  - Tuning parameter eps dan min_samples menggunakan Grid Search atau evaluasi visual.
  - Coba juga reduksi dimensi (PCA atau t-SNE) untuk membantu clustering DBSCAN.

Gambar 10. Penjelasan Parameter DBSCAN dan Kesimpulan t-SNE

4. Kesimpulan

Penelitian ini membandingkan efektivitas beberapa model *machine learning* dengan pendekatan *supervised learning* dan *unsupervised learning* dalam klasifikasi penyakit kardiovaskular dengan menggunakan *dataset* berdasarkan fitur klinis. Dalam penelitian ini didapatkan bahwa pendekatan *supervised learning* lebih unggul dibandingkan *unsupervised learning*.

Model Random Forest menunjukkan hasil terbaik dengan tingkat akurasi 98.54% dan nilai AUC mencapai 1.00 yang menunjukkan kapabilitas yang sangat baik dalam membedakan kelas positif dan negatif. Model K-Nearest Neighbor memperoleh tingkat akurasi sebesar 83.41%, sedangkan Logistic Regression memperoleh tingkat akurasi sebesar 79.51%. Hasil ini sejalan dengan literatur Chandrasekhar dan Peddakrishna (2023) yang menyatakan bahwa Random Forest merupakan salah satu model terbaik dalam klasifikasi penyakit kardiovaskular. Selain itu, efektifitas KNN sebagai model yang kompetitif juga didukung oleh penelitian Assegie (2021) yang menunjukkan performa baik dalam memprediksi penyakit kardiovaskular.

Sebaliknya, model *unsupervised learning* seperti K-Means CLustering dan DBSCAN menunjukkan performa yang jauh lebih rendah. Nilai ARI untuk K-Means sebesar 0.376 dan *silhouette* sebesar 0.1687, sedangkan DBSCAN menunjukkan ARI sebesar 0.0068 dengan *silhouette* negatif. Hal ini menunjukkan bahwa model *clustering* kurang tepat untuk penggunaan klasifikasi jika tidak terdapat label pada data. Temuan ini sesuai dengan studi oleh Jetty et al. (2025), yang mengindikasikan bahwa K-Means dapat mengelompokkan data dengan tingkat akurasi tertentu namun efektivitasnya terbatas untuk klasifikasi medis.

Dari hasil tersebut dapat disimpulkan bahwa pendekatan *supervised learning* lebih direkomendasikan untuk pengembangan sistem klasifikasi penyakit kardiovaskuler yang membutuhkan akurasi dan reliabilitas tinggi. Meskipun demikian, *unsupervised learning* tetap memiliki potensi dalam tahap eksplorasi awal, khususnya dalam segmentasi data pasien berdasarkan pola laten yang belum dikenali.

5. Sumber Pustaka/Rujukan

Sumber pustaka dalam penelitian ini sebagian besar berasal dari artikel-artikel ilmiah terbitan lima tahun terakhir, dengan fokus utama pada penggunaan algoritma pembelajaran mesin untuk klasifikasi penyakit jantung. Minhas et al. (2024) dalam peninjauan komprehensifnya mengulas berbagai pendekatan *machine learning* dan efektivitasnya dalam prediksi penyakit jantung, memberikan dasar teori yang kuat untuk pemilihan model dalam penelitian ini. Saxena dan Sharma (2021) secara khusus membandingkan performa beberapa algoritma seperti KNN dan Logistic Regression dalam konteks diagnosis klinis, yang menjadi acuan dalam analisis model supervised learning. Di sisi lain, pendekatan *unsupervised learning* dikaji secara mendalam oleh Mansoori et

al. (2020), yang menunjukkan potensi *clustering* dalam mengelompokkan data medis tanpa label.

Lebih lanjut, studi oleh Assegie (2021) membuktikan efektivitas algoritma KNN dalam prediksi penyakit jantung dengan akurasi yang kompetitif, mendukung hasil empiris dalam penelitian ini. Penelitian oleh Jetty et al. (2025) secara spesifik mengevaluasi metode *unsupervised learning* seperti K-Means dan *DBSCAN* dalam konteks diagnosis jantung, dan menjadi referensi utama dalam bagian evaluasi clustering. Selain itu, penelitian oleh Chandrasekhar dan Peddakrishna (2023) menjadi pelengkap dalam mengeksplorasi peningkatan akurasi melalui teknik optimasi model, yang menginspirasi penggunaan *hyperparameter tuning* menggunakan RandomizedSearchCV pada penelitian ini.

Dengan demikian, keseluruhan pustaka yang digunakan tidak hanya sesuai dengan konteks topik penelitian, namun juga mencerminkan perkembangan terkini dalam penerapan *machine learning* untuk sistem pendukung keputusan medis.

6. Aturan Lain

Penulisan artikel ini mengikuti seluruh pedoman penulisan yang ditetapkan oleh Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), termasuk penggunaan format dua kolom, penggunaan bahasa Indonesia yang sesuai dengan Ejaan Yang Disempurnakan (EYD), serta penyesuaian ukuran huruf, format kutipan, dan tata letak tabel maupun gambar.

Seluruh konten dalam artikel ini, termasuk rancangan prototipe, visualisasi antarmuka pengguna, dan hasil evaluasi, merupakan hasil karya orisinal penulis berdasarkan proses perancangan dan pengujian mandiri. Setiap referensi, kutipan, maupun sumber pendukung telah dicantumkan secara eksplisit sesuai kaidah penulisan ilmiah.

Penulis menyatakan bahwa naskah ini adalah karya asli yang belum pernah diterbitkan dalam jurnal atau prosiding manapun, dan tidak sedang diajukan di tempat lain. Segala bentuk tanggung jawab terkait orisinalitas, pemanfaatan perangkat lunak desain, serta kepatuhan terhadap hak kekayaan intelektual berada sepenuhnya di tangan penulis artikel ini.

7. Daftar Pustaka

KHAN MINHAS, M., et al., 2024. A comprehensive review of machine learning for heart disease prediction. Frontiers in Artificial Intelligence.

SAXENA, A. and SHARMA, A., 2021. Machine learning algorithms for heart disease diagnosis. Computer Methods and Programs in Biomedicine.

MANSOORI, M., et al., 2020. Detecting cardiovascular diseases using unsupervised machine learning. Journal of Biomedical Informatics.

ASSEGIE, T.A., 2021. Heart disease prediction model with k-nearest neighbor algorithm. International Journal of Informatics and Communication Technology.

JETTY, J., SULTANA, S.S., POLEPALLE, R.B. and PARUSU, V., 2025. Unsupervised learning for heart disease prediction: Clustering-based approach. ITM Web of Conferences, 74, p.01005.

CHANDRASEKHAR, N. and PEDDAKRISHNA, S., 2023. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. Processes, 11(4), p.1210.