Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 220-225
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

PERBANDINGAN METODE MACHINE LEARNING UNTUK MENDETEKSI PENYAKIT JANTUNG

Yutri Amelia[1*]
[1]Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia
Email: [1]14210237@nusamandiri.ac.id
(* : corresponding author)

Abstrak-Penyakit jantung termasuk ke dalam bagian penyakit kardiovaskular (CVD) atau sekelompok penyakit yang melibatkan pembuluh darah dan jantung yang merupakan salah satu penyakit serius yang diderita banyak orang secara global. Setiap tahunnya ada 17,9 juta jiwa yang meninggal akibat penyakit ini setiap tahunnya. Mendeteksi dini penyakit jantung sangat penting untuk perawatan dan pengobatan yang efektif. Penelitian ini memprediksi penyakit jantung dengan menggunakan metode *Machine Learning* (ML). ML memiliki efektifitas dan harga yang lebih murah untuk mendeteksi suatu penyakit. Penelitian ini bertujuan untuk memprediksi penyakit jantung dengan menggunakan perbandingan dari algoritma ML. Dalam penelitian ini menggunakan dataset dari *UCI Machine Learning Repository*. Pada penelitian ini, metode yang digunakan meliputi *Random Forest, Support Vector Machine* (SVM), *XGBoost, K-Nearest Neighbor* (KNN), *Decision Tree, Logistic Regression serta Multi-Layer Perceptron Classifier* (MLP). Dari penelitian ini didapatkan akurasi terbaik menggunakan algoritma XGBoost dengan akurasi mencapai 95,08%.

Kata Kunci: Penyakit Jantung, *Machine Learning*, SVM, XGBoost, *Random Forest*

Abstract-*Heart disease is included in the cardiovascular disease (CVD) section or a group of diseases involving the heart and blood vessels, which is a serious disease that affects many people globally. Every year there are 17.9 million people who die from this disease every year. Early detection of heart disease is essential for effective care and treatment. This study predicts heart disease using the Machine Learning (ML) method. ML has effectiveness and lower cost for detecting a disease. This study aims to predict heart disease using a comparison of the ML algorithm. In this study using datasets from the UCI Machine Learning Repository. In this study, the methods used include Random Forest, Support Vector Machine (SVM), XGBoost, K-Nearest Neighbor (KNN), Decision Tree, Logistic Regression and Multi-Layer Perceptron Classifier (MLP). From this study, the best accuracy was obtained using the XGBoost algorithm with an accuracy of 95.08%.*

Keywords*: Heart Disease, Machine Learning, SVM, XGBoost, Random Forest*

1. PENDAHULUAN

Penyakit yang paling mematikan di dunia menurut WHO[1] salah satunya adalah penyakit kardiovaskular (CVD). Kardiovaskular merupakan gangguan pembuluh darah dan jantung . Penyakit yang termasuk kedalam CVD diantaranya penyakit serebrovaskular, penyakit jantung koroner, penyakit jantung rematik, penyakit arteri perifer, penyakit jantung bawaan dan thrombosis vena dalam serta emboli paru. CVD disebabkan oleh pola makan yang kurang sehat, kurangnya aktivitas fisik, penggunaan tembakau (merokok) dan konsumi alkohol. Penyakit CVD disebut sebagai *"silent killer"* karena mengakibatkan kematian tanpa gejala yang spesifik [2]. Untuk itu perlu deteksi dini CVD untuk mengurangi resiko kematian mendadak karena serangan jantung.

Machine Learning (ML) merupakan salah satu ilmu yang paling banyak digunakan untuk peramalan penyakit. ML memiliki algoritma yang dapat membantu untuk mendeteksi suatu penyakit. Selain itu, dengan menggunakan ML, penyakit dapat dideteksi lebih dini [3] dan dapat menghemat biaya pengobatan. Efektivitas dari ML juga mampu membantu dokter untuk mendeteksi penyakit jantung dengan lebih baik dan tepat. ML menyediakan banyak metode pemrosesan yang dapat digunakan untuk meramalkan penyakit dengan efisien dan harga yang murah. Algoritma ini diklasifikasikan ke dalam *reinforcement learning*, *unsupervised*, dan *supervised*, masing-masing terdiri dari beberapa jenis algoritma.

Pada penelitian sebelumnya untuk memprediksi penyakit jantung dengan menggunakan metode *Random Forest* yang dilakukan oleh [4] mendapatkan akurasi sebesar 93,3%. Pada tahun 2021 [5] melalukan penelitian dalam memprediksi penyakit jantung menggunakan teknik ML *Random Forest* dan KNN. Hasilnya, KNN mendapatkan akurasi 86,88% dan akurasi sebesar 81,96% untuk algoritma *Random Forest*. Penelitian yang dilakukan [6] menggunakan algoritma MLP *Classifier* dan KNN dalam memprediksi CVD. Hasilnya. MLP mengungguli KNN dengan perolehan akurasi sebesar 82,47%. Menggunakan algoritma Random Forest, penelitian yang dilakukan [7] mendapatkan akurasi sebesar 83% untuk mendeteksi penyakit jantung. Menurut [8] pada penelitian yang mereka lakukan di tahun 2022 untuk memprediksi resiko penyakit jantung dengan

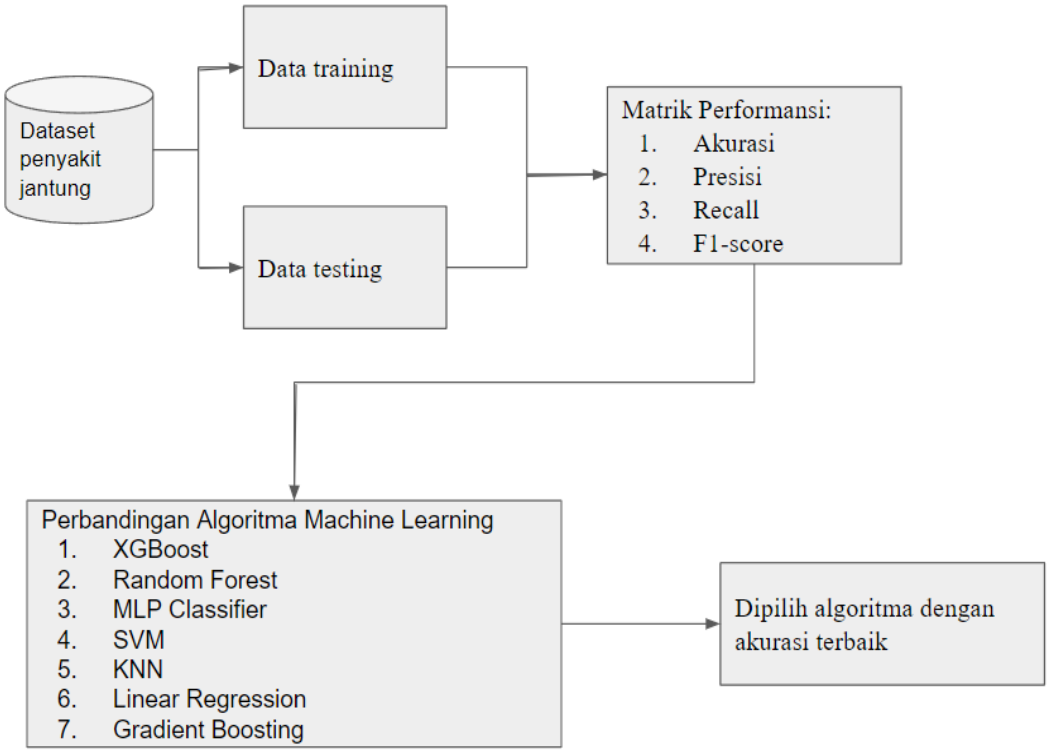
Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 220-225
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

menggunakan algoritma ML dan didapatkan hasil akurasi tertinggi sebesar 88,50% menggunakan metode *Random Forest*.

Dari beberapa penelitian tersebut, penelitian menggunakan berbagai algoritma ML telah dilakukan. Penelitian ini bertujuan untuk mendeteksi penyakit jantung dengan membandingkan beberapa algoritma ML, diantaranya *XGBoost*, *Random Forest*, MLP *Classifier*, SVM, KNN, *Linear Regression* serta *Gradient Boosting* untuk mendapatkan akurasi yang lebih baik dari penelitian sebelumnya. Hasil dari penelitian ini dapat digunakan oleh ahli medis dalam memprediksi pasien dengan penyakit jantung.

2. METODE PENELITIAN

Berikut adalah alur metode penelitian yang dilakukan yang tergambar pada Gambar1.



Gambar 1. Metode Penelitian

2.1. Algoritma *Machine Learning*

a. XGBoost
XGBoost merupakan algoritma yang disempurnakan berdasarkan decision tree dalam peningkatan gradien yang mampu membangun pohon yang ditingkatan secara efisien dan berjalan secara paralel [9]. Pada algoritma XGBoost, peningkatan pohon terbagi menjadi pohon klasifikasi dan pohon regresi. Algoritma ini mengoptimalkan nilai fungsi tujuan.

b. *Random Forest* (RF)
RF adalah teknik klasifikasi yang menghasilkan banyak pohon keputusan berdasarkan titik-titik vektor acak yang dikumpulkan secara independen dan tidak berubah-ubah. RF juga merupakan bagian dari kelas pembelajaran terbimbing yang dapat dimanfaatkan untuk membuat prediksi. Ini adalah pendekatan klasifikasi yang memiliki tingkat akurasi tinggi dan dapat menangani beberapa parameter masukan tanpa overfitting [10].

c. MLP *Classifier*
MLP merupakan gabungan dari suatu unit syaraf yang disebut sebagai *perceptron*. Pada MLP terdapat lapisan yang berisi sejumlah nilai melalui perceptron yang terhubung satu sama lain. Untuk melatih *feed-forward neural network*, digunakan algoritma *backpropagation*. Bobot nilai disesuaikan guna meminimalisir kesalahan saat pelatihan *neural network* [6].

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 220-225
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

d. SVM
SVM merupakan suatu cara untuk percabangan data non-linier dan data linier [11]. SVM paling baik menggunakan hyperlane yang dapat memisahkan titik-titik didalam ruang variabel input yang terdapat kelasnya, antara 0 dan 1.

e. KNN
KNN adalah suatu algoritma data mining yang paling sering dipakai dalam klasifikasi. KNN disebut sebagai *k-Memory Based Classification* [12] dimana data *training* harus ada di memori ketika *run-time*.

f. Logistic Regression
Logistic Regression digunakan untuk memprediksi kemungkinan sebuah variabel target dimana algoritma variabel target bergantung kepada metode kategorikal yang digunakan. Selain itu variabel dapat diklasifikasikan kedalam dua kelompok [7].

g. Gradient Boosting
Algoritma *Gradient Boosting* meningkatkan dan memperluas algoritma *Decision Tree* yang sederhana dengan menggunakan *stochastic gradient boosting*. Algoritma ini memberikan hasil model prediksi dalam bentuk ansambel dari beberapa *decision tree* sederhana yang mana algoritma ini menurunkan keunggulan dari algoritma *decesion tree* dengan meningkatkan akurasi dan ketahanan [13].

2.2. Matrik Performansi

a. Akurasi
Akurasi mewakili jumlah instance data yang diklasifikasikan dengan benar di atas jumlah total instance data. Akurasi dapat dihitung dengan menggunakan persamaan (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

TP = *True Positive*
TN = *True Negative*
FP = *False Positive*
FN = *False Negative*

b. Presisi
Presisi adalah performa model yang memprediksi berapa banyak prediksi positif yang dibuat oleh model yang sebenarnya benar. Untuk menghitung presisi dapat digunakan rumus seperti pada persamaan (2) berikut ini.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

TP = *True Positive*
FP = *False Positive*

c. *Recall*
Recall adalah ukuran kinerja sistem klasifikasi biner. *Recall* dihitung sebagai rasio prediksi True Possitive (yaitu, berapa kali model memprediksi kelas positif dengan benar) terhadap jumlah total kasus positif aktual. Persamaan (3) menggambarkan rumus *recall.*

$$recall = \frac{TP}{TP + FN} \quad (3)$$

TP = *True Positive*
FN = *False Negative*

d. *F1-score*
F1-score merupakan metrik evaluasi pembelajaran mesin yang mengukur akurasi model. F1-score menggabungkan skor presisi dan perolehan model. Menghitung *f1-score* dapar menggunakan formula pada persamaan (4).

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right) \quad (4)$$

2.3. Dataset

Penelitian ini menggunakan dataset penyakit jantung yang bersumber dari UCI *Machine Learning Repository* [14]. Dataset terdiri dari 303 jumlah pasien dengan 14 atribut. Fitur dari dataset dapat dilihat dari Tabel 1.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 220-225
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Tabel 1. Deskripsi Dataset

No	Variabel	Deskripsi	Tipe
1	*Age*	Umur	numerik
2	Sex	Jenis Kelamin	numerik
3	cp	*Chest Pain* (Nyeri Dada)	numerik
4	trtbps	*Resting blood pressure results during hospitalised* (Hasil tekanan darah istirahat selama dirawat di rumah sakit)	numerik
5	chol	*Cholesterol* (Kolesterol)	numerik
6	fbs	*Fasting blood sugar* (gula darah puasa)	numerik
7	restecg	*Electrocardiographic results during resting* (hasil elektrokardiografi saat istirahat 1=benar)	numerik
8	thalachh	*Maximum heart rate achieved* (detak jantung maksimum yang dicapai)	numerik
9	exng	*Exercise induced angina* (Latihan induksi angina)	numerik
10	oldpeak	ST depression	desimal
11	slp	ST segment slope	numerik
12	caa	*Number of major vessels coloured by fluoroscopy* (Jumlah pembuluh darah besar yang diwarnai dengan fluoroskopi)	numerik
13	thall	*Thallium stress result* (hasil tegangan thallium)	numerik
14	output	1=pasien sakit jantung 0=pasien sehat	numerik

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan algoritma ML XGBoost, *Random Forest*, MLP, SVM, KNN, *Logistic Regression* dan *Gradient Boosting*. Data di analisis menggunakan *Python* versi 3. Pemisahan *dataset* menjadi subset *training* dan *testing* dengan menggunakan metode pemisahan acak. Perbandingan antara data *training* dan *testing* adalah 80% dari 303 data yang berjumlah 242 data berbanding 20% dari 303 data yang berjumlah 61 data. Parameter yang digunakan untuk setiap algoritma dijelaskan dalam Tabel 2.

Tabel 2. Parameter Algoritma

Algoritma	Parameter
XGBoost	*objective=binary:logistic* *learning_rate*=0.1 *max_depth*=1 *n_estimators* = 50 *colsample_bytree* = 0.5
Random Forest	*n_estimators*=300 *criterion*="gini" *random_state*=5 *max_depth*=100
MLP	*random_state*=48 *hidden_layer_sizes*=(150,100,50) *max_iter*=150 *activation* = 'relu' *solver*='adam'
SVM	*kernel*="rbf"
KNN	*n_neighbors*=15
Gradient Boosting	*random_state*=10 *n_estimators*=20 *learning_rate*=0.29 *loss*="deviance"

C*onfusion matrix* digunakan untuk melihat seberapa banyak kelas negatif dan kelas positif yang berhasil diramalkan oleh sistem. Hasil dari *confusion matrix* dapat dilihat dalam Tabel 3.

Idealis: Indonesia Journal Information System
Volume 6, Nomor 2, Juli 2023
ISSN 2684-7280 (*online*)
Halaman 220-225
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Tabel 3. Hasil *Confusion Matrix*

Algoritma	True Possitive	True Negative	False Positive	False Negative
XGBoost	30	28	1	2
Random Forest	29	27	2	3
MLP	30	27	2	2
SVM	30	25	4	2
KNN	29	25	4	3
Logistic Regression	30	25	4	2
Gradient Boosting	28	26	3	4

Dari *confusion matrix* diperoleh:

a. *True Positive* (TP): dimana pasien diprediksi memiliki penyakit jantung, dan benar bahwa pasien tersebut memiliki penyakit jantung.
b. *True Negative* (TN): dimana pasien diprediksi tidak memiliki penyakit jantung, dan benar bahwa pasien tidak memiliki penyakit jantung.
c. *False Positve* (FP): dimana pasien diprediksi memiliki penyakit jantung, ternyata pasien tidak memiliki penyakit jantung
d. *False Negative* (FN): dimana pasien diprediksi tidak memiliki penyakit jantung, ternyata pasien memiliki penyakit jantung.

Pada penelitian ini matrik performansi digunakan untuk mendapatkan akurasi, sensitifitas (*recall*), presisi dan *f1-score* untuk masing-masing algoritma yang ditampilkan dalam Tabel 4.

Tabel 4. Hasil *Confusion Matrix*

Algoritma	Akurasi	Sensitifitas (*Recall*)	Presisi	F1 Score
XGBoost	95,08%	93,80%	96,80%	95,20%
Random Forest	91,80%	90,60%	93,50%	92,10%
MLP	93,44%	93,80%	93,80%	93,80%
SVM	90,16%	93,80%	88,20%	90,90%
KNN	88,52%	90,60%	87,90%	89,20%
Logistic Regression	90,16%	93,80%	88,20%	90,90%
Gradient Boosting	88,52%	87,50%	90,30%	88,90%

Dari Tabel 4 diperoleh nilai sensitifitas untuk masing-masing algoritma, yang mana menunjukkan bahwa algoritma XGBoost, MLP, SVM dan *Logistic Regression* memprediksi penyakit jantung 93,80% dengan benar. XGBoost memperoleh hasil presisi terbaik yang menunjukkan 96,80% pasien yang menderita penyakit jantung diprediksi menderita penyakit jantung. Algoritma XGBoost mendapatkan nilai akurasi tertinggi dalam memprediksi penyakit jantung yang memperoleh nilai sebesar 95,08%.

4. KESIMPULAN

Pada penelitian ini, algoritma ML diimplementasikan dalam memprediksi penyakit jantung. Dari analisis yang telah dikerjakan, algoritma XGBoost memperoleh nilai akurasi,sensitifitas, presisi dan *f1-score* tertinggi untuk memprediksi penyakit jantung, masing-masing nilainya 95,08%, 93,80%, 96,80% dan 95,20%. Model yang diusulkan mendapatkan perolehan akurasi yang lebih baik dari penelitian sebelumnya. Untuk penelitian mendatang, dapat diterapkan algoritma *deep learning* untuk mengelola *dataset* pasien dengan jumlah yang lebih besar.

DAFTAR PUSTAKA

[1] WHO, "Cardiovascular diseases (CVDs)," *World Health Organization*, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] K. A. N. G. Harinadha Babu, Gunda Jayasree, Chattu Ashika, Vajja Ahalya, "Heart Disease Prediction System Using Random Forest Technique G.," *Int. J. Res.*, vol. 4, no. 1, pp. 88–100, 2023.
[3] K. Kanagarathinam and K. Sekar, "Estimation of the reproduction number and early prediction of the COVID-19 outbreak in India using a statistical computing approach," pp. 1–5, 2020.

[4] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, 2021, doi: 10.1088/1742-6596/1817/1/012009.

[5] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012046.

[6] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Med.*, vol. 17, no. 1, pp. 1100–1113, 2022, doi: 10.1515/med-2022-0508.

[7] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc. Anal.*, vol. 2, pp. 100016, 2022, doi: 10.1016/j.health.2022.100016.

[8] K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," *Comput. Math. Methods Med.*, pp. 14, 2022, doi: 10.1155/2022/6517716.

[9] J. Liu, J. Wu, S. Liu, M. Li, K. Hu, and K. Li, "Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model," *PLoS One*, vol. 16, no. 2 February, pp. 1–11, 2021, doi: 10.1371/journal.pone.0246306.

[10] Mila Desi Anasanti, Khairunisa Hilyati, and Annisa Novtariany, "The Exploring feature selection techniques on Classification Algorithms for Predicting Type 2 Diabetes at Early Stage," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 5, pp. 832–839, 2022, doi: 10.29207/resti.v6i5.4419.

[11] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," *World J. Eng. Technol.*, vol. 06, no. 04, pp. 854–873, 2018, doi: 10.4236/wjet.2018.64057.

[12] L. Syafa'ah, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," *J. Online Inform.*, vol. 6, no. 1, pp. 73-78, 2021, doi: 10.15575/join.v6i1.719.

[13] P. Lu *et al.*, "A Gradient Boosting Crash Prediction Approach for Highway-Rail Grade Crossing Crash Analysis," *J. Adv. Transp.*, vol. 1, pp. 1-10, 2020, doi: 10.1155/2020/6751728.

[14] Janosi Andras, Steinbrunn William, Pfisterer Matthias, and Detrano Robert. (1988). Heart Disease. UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C52P4X.