
**PREDIKSI PENYAKIT JANTUNG CARDIOVASCULAR
MENGUNAKAN MODEL ALGORITMA KLASIFIKASI**

Wahyu Nugraha

PREDIKSI PENYAKIT JANTUNG CARDIOVASCULAR MENGGUNAKAN MODEL ALGORITMA KLASIFIKASI

Wahyu Nugraha¹

¹Program Studi Sistem Informatika, Universitas Bina Sarana Informatika

e-mail: *wahyu.whn@bsi.ac.id

Abstrak

Penderita penyakit cardiovascular (CVD) adalah gangguan pada jantung atau penyakit jantung koroner di seluruh dunia terus mengalami peningkatan dan menjadi penyakit yang paling mematikan. Sistem perawatan kesehatan di seluruh dunia mengalami kesulitan karena kurangnya keahlian staf medis dalam menentukan dan memprediksi penyakit ini. Salah satu cara efektif dalam mengidentifikasi dan memprediksi penyakit jantung adalah dengan memanfaatkan algoritma machine learning. Machine learning mampu mengatasi kerumitan dalam mendiagnosis penyakit jantung dengan model prediksi. Penelitian ini bertujuan untuk melakukan prediksi Penyakit Cardiovascular menggunakan beberapa algoritma klasifikasi pada machine learning seperti Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST, Light GBM. Hasil Penelitian memperlihatkan bahwa model prediksi menggunakan algoritma XGBOOST memperoleh rata-rata nilai tertinggi dalam melakukan prediksi.

Kata kunci— Cardiovascular, Machine Learning, Algoritma Klasifikasi.

Abstract

Cardiovascular disease (CVD) is a disorder of the heart or coronary heart disease worldwide continues to increase and become the most deadly disease. Health care systems around the world are experiencing difficulties due to the lack of expertise of medical staff in determining and predicting this disease. One effective way to identify and predict heart disease is to utilize machine learning algorithms. Machine learning can overcome the complexity of diagnosing heart disease with predictive models. This study aims to predict Cardiovascular Disease using several classification algorithms in machine learning such as Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST, Light GBM. The results showed that the prediction model using the XGBOOST algorithm obtained the highest average value in making predictions.

Keywords— Cardiovascular, Machine Learning, Classification Algorithm

I. PENDAHULUAN

Penyakit Cardiovascular merupakan salah satu penyakit paling mematikan pada manusia. Penyakit kardiovaskular membunuh sekitar 17 juta orang di seluruh dunia setiap tahunnya (Chicco & Jurman, 2020). Meningkatnya angka penyakit kardiovaskular dengan angka kematian yang tinggi menyebabkan risiko dan menjadi beban cukup berat bagi sistem perawatan kesehatan di seluruh dunia (Ghosh et al., 2021). Penyakit kardiovaskular lebih banyak diderita pada pria dibandingkan

dengan wanita terutama pada usia paruh baya hingga tua (Oh & Jeong, 2020). Menurut data WHO (World Health Organization), sepertiga dari kematian secara global disebabkan oleh penyakit jantung. Cardiovascular menyebabkan kematian sekitar 17,9 juta orang setiap tahun di seluruh dunia dan memiliki prevalensi yang lebih tinggi di Asia (Ghosh et al., 2021). Penyakit kardiovaskular (CVD) adalah gangguan pada jantung dan pembuluh darah termasuk, penyakit jantung koroner atau serangan jantung), penyakit

serebrovaskular atau stroke, gagal jantung, dan jenis patologi lainnya. Rata-rata setengah dari semua pasien yang didiagnosis dengan Penyakit Jantung meninggal hanya dalam waktu 1-2 tahun (Haq et al., 2018). Sebagian besar penyakit kardiovaskular dapat dicegah dengan mengatasi faktor risiko perilaku seperti penggunaan tembakau, diet tidak sehat dan obesitas, kurangnya aktivitas fisik, dan minuman alkohol.

Untuk memprediksi penyakit jantung beberapa tes diperlukan. Kurangnya keahlian staf medis dapat menghasilkan prediksi yang salah (Pouriye et al., 2017). Diagnosa secara dini terhadap penyakit ini menjadi sangat sulit. Tantangan terhadap penanganan bedah penyakit jantung semakin sulit, terutama di negara berkembang yang kekurangan ahli tenaga medis terlatih serta sumber daya lain yang diperlukan untuk diagnosis dan perawatan yang tepat bagi pasien dengan masalah jantung (Ghwanmeh et al., 2013). Evaluasi berdasarkan prediksi yang akurat dari risiko gagal jantung akan sangat membantu pasien mencegah serangan jantung yang parah dan meningkatkan persentase angka keselamatan pasien (Al-Shayea, 2011). Salah satu cara efektif dalam mengidentifikasi dan memprediksi penyakit jantung adalah dengan memanfaatkan algoritma machine learning (Ghosh et al., 2021). Machine learning mampu mengatasi kerumitan dalam mendiagnosis penyakit jantung dengan model prediksi pembelajaran mesin seperti mesin vektor dukungan (SVM), k-nearest tetangga (K-NN), jaringan saraf tiruan (JST), pohon keputusan (DT), regresi logistik (LR), AdaBoost (AB), Naive Bayes (NB), logika fuzzy (FL) (Haq et al., 2018).

Algoritma machine learning sangat efektif ketika dilatih pada kumpulan data yang sesuai (Mienye et al., 2020). Para peneliti telah menerapkan teknik yang berbeda untuk mengembangkan algoritma.

klasifikasi. Masih ada beberapa masalah yang dapat mengurangi akurasi dari prediksi penyakit jantung, seperti terbatasnya dataset medis, pemilihan fitur, aplikasi algoritma ML, dan kurangnya analisis yang mendalam (Ghosh et al., 2021).

Penelitian kami bertujuan untuk membuat beberapa model terbaik untuk prediksi penyakit cardiovascular menggunakan beberapa algoritma klasifikasi yang ada pada machine learning.

Dalam penelitian ini, kami menggunakan dataset yang berasal dari Kaggle Datasets yang terdiri dari 299 pasien dengan gagal jantung. Kumpulan data ini berisi 12 fitur yang dapat digunakan untuk memprediksi kematian akibat gagal jantung. Dataset dapat diunduh pada <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>. Penelitian ini bertujuan untuk melakukan prediksi Penyakit Cardiovascular menggunakan beberapa algoritma klasifikasi pada machine learning seperti Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST, Light GBM. Dari model prediksi yang telah dibangun selanjutnya akan dilakukan perbandingan hasil menggunakan Confusion Matrix. Dari matrik konfusi akan diperoleh nilai accuracy, F1 Score dan AUC.

Dari hasil eksperimen semua model prediksi akan diperoleh model terbaik untuk menghasilkan prediksi yang lebih akurat. Hasil Eksperimen memperlihatkan bahwa model prediksi menggunakan algoritma klasifikasi Xgboost memperlihatkan rata-rata nilai tertinggi dibandingkan algoritma yang lain dengan nilai AUC sebesar 0.75, F1 Score sebesar 0.86 dan Accuracy sebesar 0.80. Model prediksi ini diharapkan dapat berdampak pada praktik klinis agar menjadi alat pendukung untuk memprediksi apakah pasien gagal jantung akan bertahan hidup atau tidak.

II. TINJAUAN PUSTAKA

2.1 Data Preparation

Data preparation pada proses data mining selalu dilakukan yang bertujuan untuk mendapatkan data yang siap diolah dan berkualitas baik (Vercellis, 2009). Beberapa teknik data preparation diantaranya: Data validation digunakan untuk mengatasi masalah Incompleteness atau data yang mengalami missing value. Data integration dan transformation, Data yang bernilai

kategorikal ditransformasikan ke dalam angka misalkan seperti atribut Outcome (positive, negative) diubah kedalam bentuk angka (0,1). Selain itu konversi tipe data ke bentuk format numerik juga diperlukan agar dataset dapat diuji coba menggunakan model klasifikasi.

2.2 Confusion Matrix

Hasil pengujian akan diukur menggunakan confusion matrix atau matrik konfusi yang dapat dilihat pada Tabel 1 untuk mendapatkan nilai akurasi, sensitivity, specificity dan Area Under Curve (AUC) yang dinyatakan dalam pernyataan 1, 2, 3, dan 4. Confusion matrix memberikan penilaian kinerja model klasifikasi berdasarkan jumlah objek yang diprediksi dengan benar dan salah (Gorunescu, 2011).

Tabel 1 Model Confusion Matrix

Class f i c a t i o n	Predicted Class	
	Class = YES	Class = NO
Observed Class	Class = YES	A (true positive-TP)
	Class = NO	C (false positive-FP)
Observed Class	Class = YES	B (false negative-FN)
	Class = NO	D (true negative-TN)

2.3 Validation dan evaluasi model

Model 10 fold cross-validation akan digunakan dalam proses learning dan testing karena metode ini telah menjadi metode validasi standar dan state-of-the-art (Witten et al., 2011). Model 10 fold cross-validation bekerja dengan cara membagi data training menjadi 10 bagian yang sama kemudian proses pembelajaran (learning) dilakukan sebanyak jumlah k yaitu 10 kali. Dari Gambar 1 dapat dilihat bahwa setiap kali dipilih satu bagian sebagai data testing. Maka, data yang lain sebanyak 9 bagian digunakan sebagai data pembelajaran (learning). Setelah itu, dihitung nilai rata-rata dan nilai penyimpangan (deviation value) dari 10 kali jumlah pengujian yang berbeda (Wahono et al., 2014).

n-validation	Dataset's Partition									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Gambar 1 Stratified 10 Fold Cross Validation

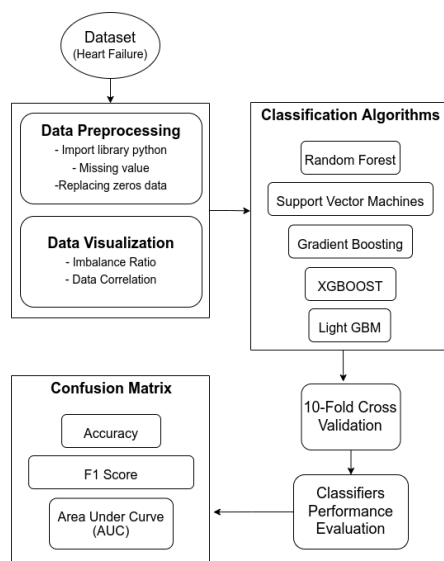
III. METODE PENELITIAN

3.1 Pengumpulan Data

Dataset yang digunakan pada penelitian ini merupakan dataset public yang merupakan contoh nyata dari penyakit Cardiovascular. Dataset yang digunakan berasal dari Kaggle dataset repository (Heart Failure Prediction) yang dapat diunduh dari <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>. Dataset ini memiliki 12 atribut dengan total sample valid sebanyak 299 sample. Dataset ini telah banyak digunakan oleh peneliti. Seorang peneliti bernama Davide Chicco merupakan peneliti awal serta merupakan kontributor dari dataset ini (Chicco & Jurman, 2020). Dataset terdiri dari beberapa variabel prediktor medis dan satu variabel target (DEATH_EVENT). Variabel prediktor meliputi umur, anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, dan lain sebagainya.

3.2 Metode Penelitian

Pada penelitian ini beberapa model yang akan dibuat menggunakan algoritma klasifikasi Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST, dan Light GBM. Gambar 2 Menunjukkan tahapan eksperimen yang dilakukan yaitu dimulai dari import dataset, data preprocessing, data visualization, model fitting dan evaluasi hasil.



Gambar 2 Model Tahapan Penelitian

IV. HASIL DAN PEMBAHASAN

Tahapan Eksperimen yang telah dilakukan untuk mencapai hasil dari penelitian ini terdiri dari beberapa tahap

diantaranya data preprocessing, data visualization, model fitting (pengujian beberapa model) dan evaluasi hasil menggunakan confusion matrix untuk menentukan model terbaik dari seluruh model yang diuji.

4.1 Data preprocessing

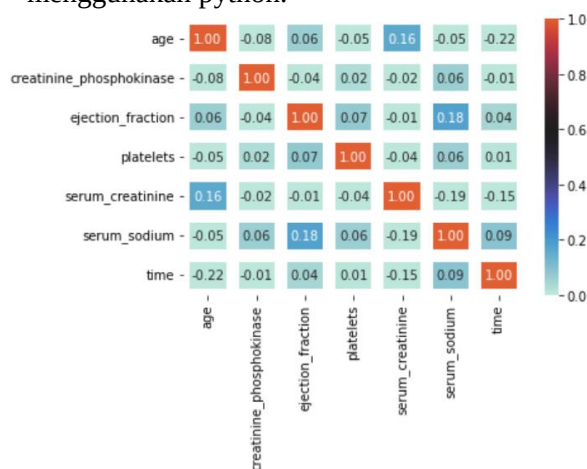
Langkah pertama eksperimen adalah mengimpor beberapa library yang diperlukan sebelum data diproses. Eksperimen ini menggunakan bahasa pemrograman python3 sehingga library yang digunakan diantaranya sklearn, numpy, pandas, matplotlib, seaborn dan plotly. Langkah selanjutnya adalah memuat dataset CSV format menggunakan library pandas. Setelah dataset dimuat, kita harus membuat feature matrix X dan vector observasi Y terhadap X. Langkah selanjutnya menemukan dan mengatasi nilai yang hilang (missing value) agar tidak menyebabkan masalah pada saat pelatihan. Gambar 3 menunjukkan detail statistik data seperti percentile, mean, std dan lainnya dari data frame.

	count	mean	std	min	25%	50%	75%	max
age	299.0	60.833893	11.894809	40.0	51.0	60.0	70.0	95.0
anaemia	299.0	0.431438	0.496107	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	299.0	581.839465	970.287881	23.0	116.5	250.0	582.0	7861.0
diabetes	299.0	0.418060	0.494067	0.0	0.0	0.0	1.0	1.0
ejection_fraction	299.0	38.083612	11.834841	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	299.0	0.351171	0.478136	0.0	0.0	0.0	1.0	1.0
platelets	299.0	263358.029264	97804.236869	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	299.0	1.393880	1.034510	0.5	0.9	1.1	1.4	9.4
serum_sodium	299.0	136.625418	4.412477	113.0	134.0	137.0	140.0	148.0
sex	299.0	0.648829	0.478136	0.0	0.0	1.0	1.0	1.0
smoking	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0
time	299.0	130.260870	77.614208	4.0	73.0	115.0	203.0	285.0
DEATH_EVENT	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0

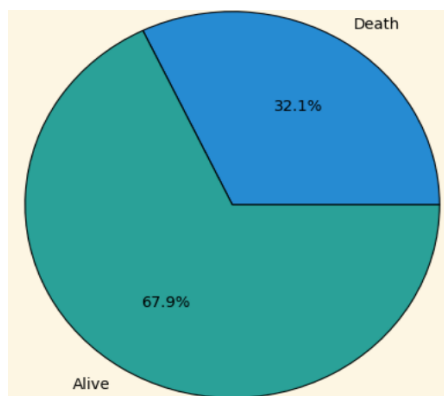
Gambar 3 Detail Statistik Data

1.1. Data Visualization

Analisis data visualisasi pada tahap ini dilakukan untuk menemukan jawaban atas pertanyaan seperti rasio dari imbalance class dan melihat hubungan antar atribut data dan feature yang berkaitan erat dengan penyakit jantung. Gambar 5 Menggambarkan persentase imbalance data 67.9% Alive dan 32.1% mengalami kematian. Sedangkan Gambar 4 menggambarkan hubungan korelasi antar variabel yang digambarkan dengan seaborn correlation heatmap menggunakan python.



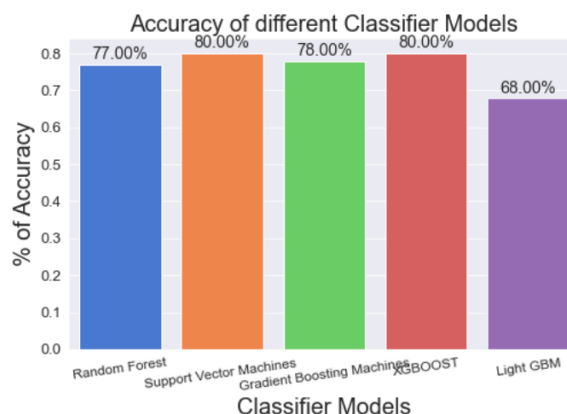
Gambar 4 Correlation Plot



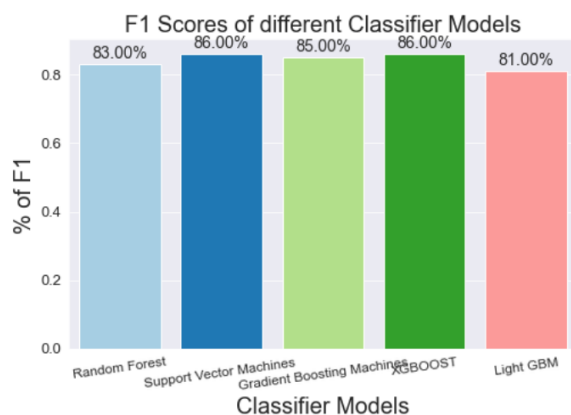
Gambar 5 Imbalance Class Distribution

4.2 Model Fitting

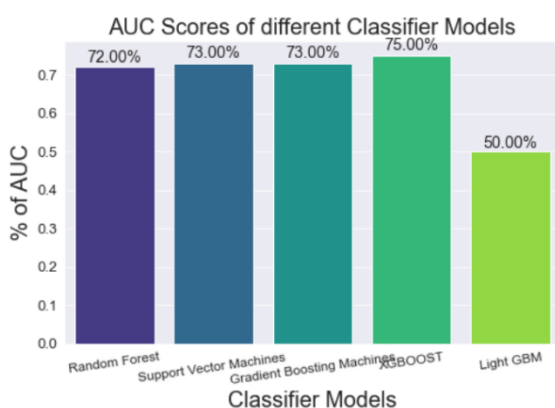
Pengujian terhadap model klasifikasi pada eksperimen ini akan menggunakan 5 algoritma klasifikasi yaitu Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST dan Light GBM. Dari seluruh model yang dibangun akan dicari model terbaik berdasarkan hasil eksekusi. Untuk menentukan model terbaik maka pengukuran terhadap model algoritma klasifikasi menggunakan metrik pengukuran confusion matrix. Dari hasil eksperimen dapat dilihat pada grafik gambar 6 menunjukkan bahwa dari model yang telah diujikan model SVM dan XGBoost memperoleh nilai accuracy sebesar 0.80, Gradient Boosting sebesar 0.78, Random forest 0.77 dan Light GBM sebesar 0.68. Pada Gambar 7 dapat dilihat bahwa nilai F1 Score masing-masing algoritma memperoleh nilai sebesar 0.86 untuk model SVM dan XGBoost, 0.85 untuk model Gradient Boosting, 0.83 untuk model Random forest dan 0.81 untuk model prediksi Light GMB. Sedangkan untuk pengukuran model menggunakan Area Under Curve (AUC) menunjukkan bahwa XGBOOST memperoleh nilai sebesar 0.75, SVM dan Gradient Boosting memperoleh nilai AUC 0.73, model random forest memperoleh nilai sebesar 0.72 dan Light GBM memperoleh nilai AUC sebesar 0.50.



Gambar 6 Nilai Accuracy Semua Model



Gambar 7 Nilai F1 Score Semua Model



Gambar 8 Nilai AUC Semua Model

V. KESIMPULAN

Berdasarkan penelitian yang dilakukan menggunakan dataset Kaggle dataset repository (Heart Failure Prediction) dapat disimpulkan bahwa model klasifikasi menggunakan XGBOOST rata-rata memperoleh nilai tertinggi baik menggunakan pengukuran accuracy, F1 score maupun AUC. Model klasifikasi SVM memperoleh nilai rata-rata hampir mirip dengan model Gradient Boosting. Sedangkan untuk model pengukuran dengan nilai terendah diperoleh dengan model klasifikasi LightGBM.

VI. DAFTAR PUSTAKA

Al-Shayea, Q. K. (2011). Artificial Neural Networks in Medical Diagnosis. *IJCSI International Journal of Computer Science*, 8(2), 150–154.

<https://doi.org/10.2478/v10136-012-0031-x>

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1), 1–16. <https://doi.org/10.1186/s12911-020-1023-5>

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, 9, 19304–19326. <https://doi.org/10.1109/ACCESS.2021.3053759>

Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis. *Journal of Intelligent Learning Systems and Applications*, 05(03), 176–183. <https://doi.org/10.4236/jilsa.2013.53019>

Gorunescu, F. (2011). Data mining: concepts and techniques. In *Springer-Verlag*. <https://doi.org/10.1007/978-3-642-19721-5>

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., Sun, R., & García-Magarinõ, I. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018. <https://doi.org/10.1155/2018/3860146>

Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, 20, 100402. <https://doi.org/10.1016/j.imu.2020.100402>

Oh, M. S., & Jeong, M. H. (2020). Sex Differences in Cardiovascular Disease Risk Factors among Korean Adults. *The Korean Journal of Medicine*,

- 95(4), 266–275.
<https://doi.org/10.3904/kjm.2020.95.4.266>
- Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *Proceedings - IEEE Symposium on Computers and Communications, Iscc*, 204–207. <https://doi.org/10.1109/ISCC.2017.8024530>
- Vercellis, C. (2009). *Business Intelligence : Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd.
- Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). *A Comparison Framework of Classification Models for Software Defect Prediction*. 20(10), 1945–1950. <https://doi.org/10.1166/asl.2014.5640>
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Third Edition*. In [Morgan Kaufmann series in data management systems]. Elsevier Inc. [https://doi.org/10.1002/1521-3773\(20010316\)40:6<9823::AID-ANIE9823>3.3.CO;2-C](https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C)