ANALISIS PENGGUNAAN MACHINE LEARNING DALAM KLASIFIKASI PENENTUAN PENYAKIT JANTUNG

Anita Carolina Wibowo[1*], Sofiana Ardi Lestari[2], Nurchim[3]
[1,2,3]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Duta Bangsa Surakarta
*e-mail : [*1]1210101006@mhs.udb.ac.id, [2]2210101038@mhs.udb.ac.id, [3]nurchim@udb.ac.id*

Penyakit kardiovaskular menyebabkan sekitar 17,9 juta kematian setiap tahun, menjadikan penyakit tersebut sebagai penyebab utama kematian di seluruh dunia.. Penelitian ini bertujuan untuk menganalisis model klasifikasi penyakit jantung yang akurat menggunakan teknik machine learning. Algoritma machine learning yang dianalisis meliputi K-Nearest Neighbor (K-NN), Logistic Regression, dan Decision Tree. Penelitian ini menggunakan Heart Disease Dataset dari Kaggle terdiri dari 1025 record dengan 14 atribut. Tahapan penelitian dimulai dari pre-processing data hingga evaluasi performa dilakukan menggunakan bahasa pemrograman Python. Hasil penelitian menunjukkan bahwa algoritma K-NN mencapai akurasi tertinggi sebesar 94% diikuti algoritma Decision Tree memiliki akurasi sebesar 93% dan terakhir Logistic Regression dengan akurasi sebesar 86%. Dari evaluasi ini, dapat disimpulkan bahwa algoritma K-NN memiliki kinerja terbaik dalam klasifikasi data klinis pasien penyakit jantung. Penelitian ini diharapkan memberikan kontribusi signifikan dalam pemilihan algoritma machine learning untuk mendukung diagnosis medis yang lebih baik.

***Kata Kunci*: Penyakit Jantung, Klasifikasi, K-Nearest Neighbor, Logistic Regression, Decision Tree.**

I. PENDAHULUAN

Klasifikasi adalah teknik pembelajaran mesin yang digunakan untuk memilah berdasarkan pola data[1]. Klasifikasi menetapkan kelas atau grup dari setiap sampel data, menggunakan atribut sampel sebagai input dari model klasifikasi dan kelas sampel sebagai outputnya[2]. Dua jenis klasifikasi yang paling umum digunakan adalah klasifikasi biner yang hanya memiliki dua kelas dan klasifikasi multikelas yang memiliki lebih dari dua kelas. Klasifikasi menggunakan label atau kelas sebagai target pembelajaran mesin, dan fitur berfungsi sebagai data yang akan dicari polanya berdasarkan label. Dalam klasifikasi ini, data biasanya dibagi menjadi dua: data latih dan data uji. [4].

Algoritma K-Nearest Neighbor (K-NN) mengklasifikasikan objek berdasarkan data pembelajaran (neighbor) yang paling dekat dengannya. Algoritma ini merupakan salah satu algoritma yang sederhana dan efisien. Sedangkan Logistic regression adalah metode analisis data dalam statistika yang dimaksudkan untuk mengetahui hubungan antara setiap variable dengan menggunakan probabilitas untuk memprediksi data kategorikal, dan dengan menggunakan fungsi sigmoid untuk menggabungkan nilai input dan nilai koefisien secara linear untuk memprediksi hasilnya.[1]. Sementara Decision Tree merupakan algoritma pembelajaran mesin untuk regresi dan klasifikasi[5]. Decision tree terdiri dari akar (root), brach node, dan leaf node. Prosesnya dilakukan dari akar hingga leaf node secara rekursif, yang membuat proses pengambilan keputusan menjadi lebih mudah[6].

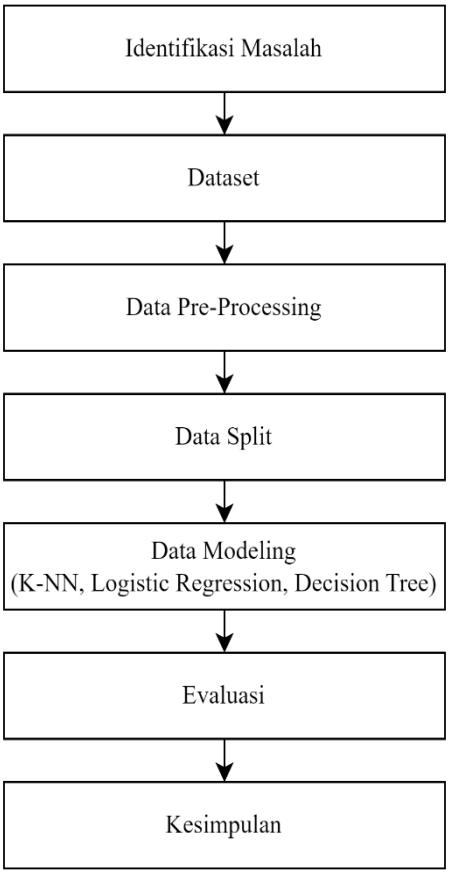
Sebelumnya pernah dilakukan penelitian klasifikasi penyakit jantung dengan algoritma K-Nearest Neighbor (KNN). Hasil penelitian menunjukkan bahwa metode K-NN efektif dalam klasifikasi penyakit jantung, dengan nilai K terbaik 6, akurasi 85%, presisi 78%, recall 93%, dan f-measure 85%[7]. Penelitian lain melakukan klasifikasi sama namun dengan algoritma Logistic Regression. Dengan sensitivitas tertinggi sebesar 88,54% dan spesifisitas tertinggi sebesar 87,50% pada data pelatihan, hasil ini menunjukkan bahwa regresi logistik adalah alat yang efektif untuk mendeteksi penyakit jantung dan memberikan kontribusi yang signifikan dalam sistem pendukung keputusan medis[1]. Selain dua penelitian tersebut, terdapat penelitian dengan kasus yang sama namun menggunakan algoritma Decision Tree. Dengan nilai recall terbaik sebesar 81,9% dan nilai ROC AUC sebesar 79,6%, model Decision Tree yang dioptimalkan dengan teknik tuning dan resampling menunjukkan performa yang stabil[1]. Dari ketiga penelitian yang pernah dilakukan dapat dilihat bahwa klasifikasi dengan performa terbaik adalah Logistic Regression.

Setelah melihat hasil dari penelitian-penelitian sebelumnya, pada penelitian ini akan melakukan perbandingan ketiga metode tersebut. Tujuan dari penelitian ini adalah untuk membuat model klasifikasi penyakit jantung yang lebih akurat yang menggunakan machine learning. Diharapkan bahwa model ini dapat memberikan prediksi yang lebih tepat mengenai kemungkinan terjadinya penyakit jantung pada individu dengan menggunakan dataset klinis yang besar dan beragam. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi seberapa baik berbagai algoritma machine learning bekerja dalam klasifikasi penyakit jantung. Hasil

penelitian ini diharapkan dapat memberikan kontribusi besar di masa mendatang dalam meningkatkan kualitas diagnosis dan perawatan penyakit jantung.

II. METODE PENELITIAN

Penelitian ini menggunakan metodologi yang terdiri dari beberapa tahapan. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

A. Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari website Kaggle yaitu Heart Disease Dataset. Dataset berisi data yang dikumpulkan dari empat database yang digabungkan. Dataset ini memiliki 14 atribut, dimana 13 atribut merupakan atribut biasa sedangkan 1 atribut sebagai class dan memiliki 1025 record.

B. Data Pre-processing

Penelitian ini menggukan bahasa pemrograman python untuk pre-processing data hingga evaluasi performa. Library yang digunakan diantaranya Pandas, NumPy, Scikit-Learn, Matplotlib, serta Seaborn.

C. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) termasuk dalam learning instance-based dan lazy learning. K-NN menggunakan k objek dalam data pembelajaran yang paling mirip dengan objek dalam data pengujian. Sistem klasifikasi ini mencari informasi dengan menggunakan solusi pasien lama untuk mencari solusi pasien baru. Perhitungan jarak ketetanggaan dilakukan dengan algoritma Euclidean, seperti yang ditunjukkan dalam persamaan 1.

$$euc = \sqrt{((a_1 - b_1)^2 + \cdots + (a\_n - b\_n)^2)} \quad (1)$$

Keterangan :
a = a1, a2, ..., an
b = b1, b2, ..., bn mewakili n atribut dari dua record. Untuk atribut dengan nilai kategori[7].

D. Logistic Regression

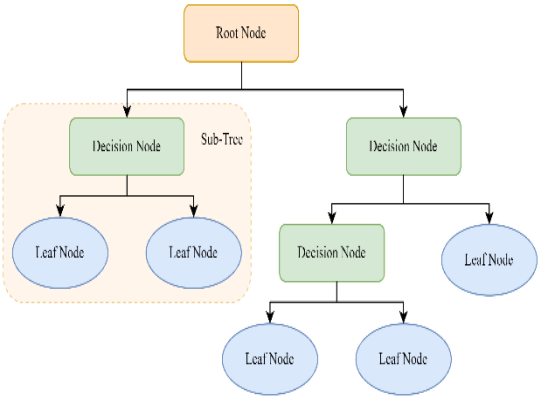
Logistic regression adalah metode analisis data statistika yang dimaksudkan untuk menentukan hubungan antara masing-masing variabel. Ini menggunakan probabilitas untuk memprediksi data kategorikal dan memprediksi hasilnya dengan menggabungkan nilai input dan koefisien secara linear dengan fungsi sigmoid.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_2 + \cdots + \beta_n X_n \quad (2)$$

Keterangan :
$\ln\left(\frac{p}{1-p}\right)$ = logit dari variabel dependen yaitu probabilitas sukses dibagi dengan probabilitas gagal
β0 = konstanta (intercept) dari model
X1, X2, ..., Xn = variabel independen yang digunakan dalam model
β1, β2, …, βn = koefisien regresi dari masing-masing variable independen[1].

E. Decision Tree

Salah satu algoritma prediktif yang digunakan untuk klasifikasi atau pengelompokan adalah Decision Tree[8]. Decision tree, juga disebut sebagai pohon keputusan, adalah algoritma yang biasanya digunakan untuk pengambilan keputusan dengan mencari solusi masalah berdasarkan kriteria dan membuat struktur yang menyerupai pohon. Setiap pohon ini memiliki cabang yang mewakili tugas yang harus dipenuhi agar dapat naik ke cabang berikutnya dan berakhir di daun[9]. Penggambaran algoritma decision tree dapat dilihat pada Gambar 2.



Gambar 1. Decision Tree[10]

F. Evaluasi

Pada tahap ini dilakukan evaluasi performa dari setiap model. Tahap ini membandingkan CA, Precision, Recall dan F1-Score dari perhitungan Confussion Matrix. Alat evaluasi yang dikenal sebagai confusion matrix digunakan untuk mengevaluasi tingkat akurasi model klasifikasi

dengan membandingkan data terklasifikasi dengan data latih[11]. Classification Accuracy (CA) merupakan pengukuran proporsi prediksi yang benar. Precision adalah akurasi data yang memungkinkan dua kejadian, yaitu 1 dan 0, dan recall mengukur rasio. F1-Score adalah perbandingan antara recall dan presisi[12]. Pada Tabel 1 dapat dilihat bagaimana plot confusion matrix.

Tabel 1. Confusion Matrix

Predicted Class	**Actual Class**	
	+	-
+	True Positives (TP)	False Positives (FP)
-	False Negatives (FN)	True Negatives (TN)

Ketika model memprediksi bahwa pasien memiliki penyakit jantung, True Positives (TP)adalah hasil yang benar, sedangkan True Negatives (TN)adalah hasil yang salah. False Negatives (FN) adalah hasil yang salah ketika model memprediksi bahwa pasien tidak memiliki penyakit jantung, tetapi pasien sebenarnya memilikinya. False Positives (FP) adalah hasil yang salah ketika model memprediksi bahwa pasien memiliki penyakit jantung, tetapi pasien sebenarnya tidak memilikinya.[1].

Selanjutnya dilakukan perhitungan Classification Accuracy, F1, Precision, dan Recall. Berikut adalah rumus perhitungan Classification Accuracy, F1, Ketepatan, dan Recall.

$$CA = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Keterangan :
TP = True Positives
TN = True Negatives
FN = False Negatives
FP = False Positives

III. HASIL DAN IMPLEMENTASI

Penelitian ini dimulai dengan menginput dataset yang akan digunakan. Heart Disease Dataset adalah dataset yang menjadi objek penelitian karena memuat informasi data klinis pasien termasuk klasifikasi apakah pasien tersebut memiliki penyakit jantung atau tidak. Pada Gambar 3 dapat dilihat dataset yang akan digunakan.

	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	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	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Gambar 2. Heart Disease Dataset

A. Hasil Data Pre-processing

Dataset yang digunakan dalam penelitian ini adalah database pasien penyakit jantung dengan kondisi medisnya yang didapat dari website Kaggle yaitu Heart Disease Dataset. Jumlah data adalah 1025 record hasil praproses data dari database ini memiliki empat belas atribut. Kondisi klinis pasien digunakan sebagai atribut prediksi pada 13 atribut, dan kelas target adalah atribut ke-14. Kelas target memiliki dua nilai. Pasien dengan penyakit jantung memiliki nilai 1 dan pasien tanpa penyakit jantung memiliki nilai 0. Rincian atribut prediksi dan kelas target dapat dilihat pada Tabel 2.

Tabel 2. Atribut Prediksi dan Kelas Target

No	Atribut	Tipe Data	Keterangan
1	Age	Integer	Usia dalam tahun
2	Sex	Integer	Jenis kelamin
3	ChestPainType	Integer	Jenis nyeri dada (1 = Angina typica, 2 = Angina atipika, 3 = Non-anginal, 4 = Asimptomatik)
4	RestingBP	Integer	Tekanan darah pasien saat istirahat (dalam mmHg)
5	Cholesterol	Integer	Kolesterol serum pasien dalam mg/dl
6	FastingBS	Integer	Apakah gula darah puasa pasien lebih dari 120 mg/dl (True/False)
7	RestingECG	Integer	Hasil elektrokardiografi pasien saat istirahat (0 = Normal, 1 = ST-T abnormalitas, 2 = Hipertrofi ventrikel kiri)
8	MaxHR	Integer	Denyut jantung maksimum yang tercapai selama tes (bpm)
9	ExcerciseAngina	Integer	Apakah pasien mengalami angina yang diinduksi oleh olahraga (True/False)
10	Oldpeak	Float	Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat
11	ST_Slope	Integer	Kemiringan segmen ST puncak selama olahraga (0 = Naik, 1 = Datar, 2 = Turun)
12	Ca	Integer	Jumlah pembuluh besar (0-3) yang terlihat berwarna saat fluoroskopi
13	Thalasemia	Integer	Jenis thalassemia yang dideteksi (0 = Normal, 1 = Defek tetap, 2 = Defek reversibel)
14	HeartDisease	Integer	1 = positif heart disease, 0 = negatif heart disease

Setelah diketahui atribut yang akan digunakan, selanjutnya dilakukan split data. Split data bertujuan untuk membagi dataset menjadi data latih dan data uji. Pada penelitian ini, dataset dibagi menjadi 80% data latih dan 20% data uji.

B. Data Latih

Dari hasil split data, data latih yang akan digunakan pada penelitian ini berjumlah 820 record. Pada Gambar 4

dapat dilihat beberapa data latih yang akan digunakan.

	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	2	2	3	0
137	64	0	0	180	325	0	1	154	1	0	2	0	2	1
797	65	0	0	150	225	0	0	114	0	1	1	3	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
322	45	1	0	142	309	0	0	147	1	0	1	3	3	0
13	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
631	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1

Gambar 3. Data Latih

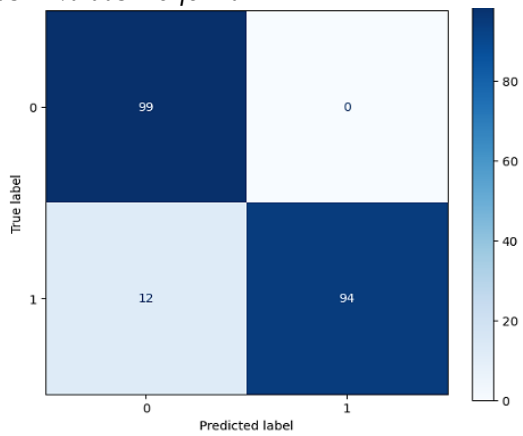
C. Data Uji

Dari hasil split data, data uji yang akan digunakan pada penelitian ini berjumlah 205 record. Data uji yang akan digunakan dapat dilihat pada Gambar 5.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
346	50	1	0	150	243	0	0	128	0	2.6	1	0	3	0
475	57	1	2	150	126	1	1	173	0	0.2	2	1	3	1
627	38	1	3	120	231	0	1	182	1	3.8	1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
553	53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
694	39	1	0	118	219	0	1	140	0	1.2	1	0	3	0
642	64	1	0	128	263	0	1	105	1	0.2	1	1	3	1

Gambar 4. Data Uji

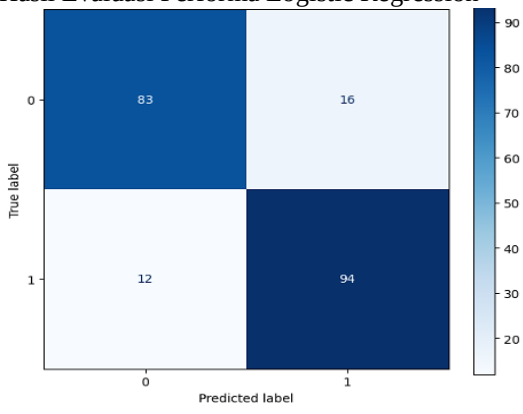
D. Hasil Evaluasi Performa K-NN



Gambar 5. Confusion Matrix

Pada gambar 6 menunjukkan bahwa dari 205 data uji, model yang menggunakan algoritma K-Nearest Neighbor (K-NN) dapat melakukan 193 prediksi dengan benar dimana 99 prediksi pada kelas negatif (TN) dan 94 prediksi pada kelas positif(TP). Namun model ini masih melakukan kesalahan pada kelas negatif(FN) sebanyak 12 prediksi dari total 205 data uji. Dari confusion matrix pada gambar 6 didapatkan hasil perhitungan classification accuracy (CA) dari model K-NN sebesar 0.94, kemudian nilai precision model ini sebesar 1, nilai Recall sebesar 0.84 dan nilai F1-Score sebesar 0.94.

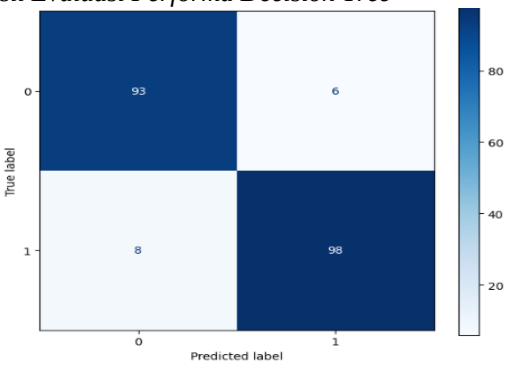
E. Hasil Evaluasi Performa Logistic Regression



Gambar 6. Confusion Matrix Logistic Regression

Pada gambar 7 menunjukkan bahwa model yang menggunakan algoritma Logistic Regression dapat melakukan 177 prediksi dengan benar dimana 83 prediksi pada kelas negatif (TN) dan 94 prediksi pada kelas positif (TP). Namun model ini masih melakukan kesalahan sebanyak 28 prediksi dimana 16 prediksi pada kelas positif (FP) dan 12 prediksi pada kelas negatif (FN). Dari confusion matrix pada gambar 7 didapatkan hasil perhitungan classification accuracy (CA) dari model Logistic Regression sebesar 0.86, kemudian nilai precision model ini sebesar 0.85, nilai Recall sebesar 0.89 dan nilai F1-Score sebesar 0.87.

F. Hasil Evaluasi Performa Decision Tree



Gambar 7. Confusion Matrix Decision Tree

Pada gambar 8 menunjukkan bahwa model yang menggunakan algoritma Decission Tree dapat melakukan 191 prediksi dengan benar dimana 93 prediksi pada kelas negatif (TN) dan 98 prediksi pada kelas positif (TP). Namun model ini masih melakukan kesalahan sebanyak 14 prediksi dimana 6 prediksi pada kelas positif (FP) dan 8 prediksi pada kelas negatif (FN). Dari confusion matrix pada gambar 8 didapatkan hasil bahwa classification accuracy (CA) dari model Decission Tree sebesar 0.93, kemudian nilai precision model ini sebesar 0.94, nilai Recall sebesar 0.92 dan nilai F1-Score sebesar 0.93.

G. Perbandingan Hasil Evaluasi Performa

Setelah dilakukan evaluasi performa pada setiap model, langkah selanjutnya adalah membandingkan hasil performanya. Tabel perbandingan model dapat dilihat pada Tabel 3.

Tabel 3.Perbandingan Performa Model

Model	CA	F1	Precision	Recall
K-NN	0.94	0.94	1	0.89
Logistic Regression	0.86	0.87	0.85	0.89
Decision Tree	0.93	0.93	0.94	0.92

Algoritma K-Nearest Neighbor (K-NN) memiliki akurasi tertinggi untuk data klinis jantung, seperti yang ditunjukkan pada tabel 3. Besarnya nilai untuk CA adalah 0.94, nilai F1 0.94, Precision 1, dan Recall 0.89.

IV. KESIMPULAN DAN SARAN

Dari penelitian ini didapatkan hasil ahwa algoritma K-Nearest Neighbor (K-NN) memiliki akurasi sebesar 94%

dalam masalah klasifikasi penyakit jantung, algoritma Logistic Regression memiliki akurasi sebesar 86%, sedangkan untuk algoritma Decision Tree memiliki akurasi sebesar 93%. Dari hasil perbandingan evaluasi performa setiap model yang telah dilakukan didapatkan kesimpulan bahwa algoritma K-NN adalah algoritma dengan kinerja paling baik untuk klasifikasi data klinis pasien penyakit jantung.

DAFTAR PUSTAKA

[1] M. K. Dwipa Jaya, "Perbandingan Random Forest, Decision Tree, Gradient Boosting, Logistic Regression untuk Klasifikasi Penyakit Jantung," *Jnatia*, vol. 2, no. November, pp. 1–5, 2023.

[2] L. Susanti, P. Studi, T. Informatika, R. Miner, and N. Bayes, "KLASIFIKASI TINGKAT STRESS PADA MAHASISWA PERKULIAHAN METODE HYBRID MENGGUNAKAN," vol. 8, no. 3, pp. 243–248, 2024.

[3] M. Aminullah *et al.*, *PERBANDINGAN PERFORMA KLASIFIKASI MACHINE LEARNING DENGAN TEKNIK RESAMPLING PERBANDINGAN PERFORMA KLASIFIKASI MACHINE LEARNING Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer (S . Kom)*. 2021.

[4] C. Journal, M. A. Bianto, M. T. Informatika, and U. A. Yogayakarta, "Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes," vol. 6, no. 1, 2019.

[5] M. Serangan, C. Fraud, R. Firdaus, A. I. Hadiana, and F. Kasyidi, "Model Deteksi Botnet Menggunakan Algoritma Decision Tree Dengan Untuk," vol. 1089, 2022.

[6] H. A. Thooriqoh, Y. A. Tofan, and A. M. Shiddiqi, "MALICIOUS TRAFFIC DETECTION IN DNS INFRASTRUCTURE," pp. 45–52.

[7] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 6–10, 2020, [Online]. Available: http://bit.ly/datasetcardio.

[8] M. K. Kosentrasi, "KLASIFIKASI KOMPTENSI MAHASISWA DENGAN ALGORITMA DECISION TREE DALAM MENETUKAN KELAYAKAN," vol. 5, no. 2, pp. 3–6, 2020.

[9] F. Y. Pamuji and V. P. Ramadhan, "Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy," vol. 7, no. 1, pp. 46–50, 2021.

[10] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," vol. 02, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[11] N. N. K. Dalam, "IMPLEMENTASI ALGORITMA K-PERANCANGAN ALAT PENDETEKSI TINGKAT KESEGARAN DAGING," vol. 9, no. 1, 2024.

[12] H. Hasanah and Nurmalitasari, "Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C45 dalam Prediksi Penyakit Jantung," *Pros. Semin. Nas. Teknol. dan Sains*, vol. 2, pp. 13–18, 2023.