Perbandingan Random Forest, Decision Tree, Gradient Boosting, Logistic Regression untuk Klasifikasi Penyakit Jantung

I Made Krisna Dwipa Jaya[a1], I Gusti Agung Gede Arya Kadyanan[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Udayana, Bali
Jln. Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, 08261, Bali, Indonesia
[1]dwipajaya19@gmail.com
[2]gungde@unud.ac.id

Abstract

Heart disease is a condition characterized by disorders affecting the heart. These heart disorders include infections, abnormalities in heart valves, blockages in the heart's blood vessels, irregular heartbeats, and so on. According to a report by the World Health Organization (WHO) in 2019, approximately 17.9 million people died from cardiovascular diseases, with 85% of them attributed to heart attacks and strokes. The shortage of doctors and specialists can lead to negligence and the overlooking of patients' symptoms, which can result in disabilities or even death for the patients. Therefore, the need for an expert system arises, which can be utilized as a tool to classify or detect heart diseases based on patients' medical records. Based on the results of the conducted research, random forest is a fairly effective algorithm for classifying heart diseases, with a recall value of 80.6% and ROC AUC of 76.3%.

Keywords: *Classification, Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, Heart Disease*

1. Pendahuluan

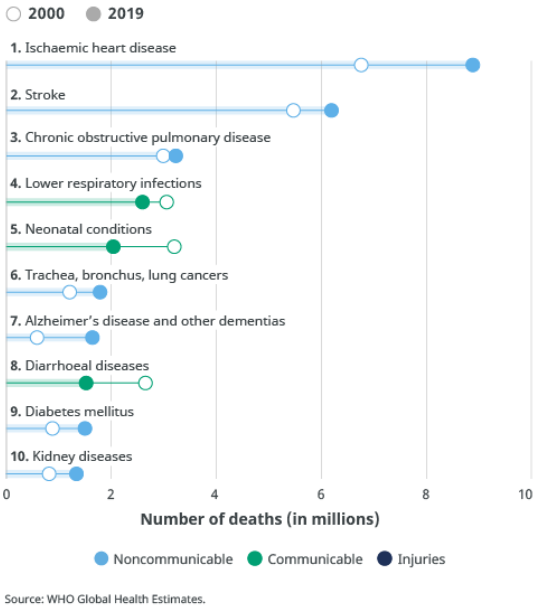
Penyakit jantung merupakan suatu kondisi dimana terdapat gangguan pada jantung. Gangguan pada jantung ini bermacam-macam seperti infeksi, kelainan pada katup jantung, penyumbatan pembuluh darah jantung, dan sebagainya. Berdasarkan laporan dari WHO (World Health Organization) pada tahun 2019, sekitar 17.9 juta orang meninggal karena penyakit kardiovaskuler yang mana 85% nya disebabkan oleh serangan jantung dan stroke.

Kurangnya dokter dan ahli yang menyebabkan kelalaian dan mengabaikan gelaja pasien dapat menyebabkan cacat hingga kematian bagi pasien. Oleh karena itu dibutuhkannya sistem pakar yang dapat digunakan sebagai alat untuk mengklasifikasi atau mendeteksi penyakit jantung berdasarkan catatan rekam medis pasien [1].

Klasifikasi merupakan metode yang terdapat pada *machine learning* yang digunakan untuk memilah berdasarkan pola dari data tersebut. Terdapat dua macam klasifikasi yang biasanya digunakan yaitu klasifikasi biner yang memiliki 2 kelas saja dan klasifikasi multikelas yang memiliki lebih dari dua kelas [2]. Dalam klasifikasi biasanya terdapat label atau kelas sebagai target dari *machine learning* dan fitur sebagai data yang akan dicari polanya berdasarkan label yang dimiliki. data yang digunakan dalam klasifikasi ini biasanya dibagi menjadi dua yaitu data latih dan data uji [3].

Dalam penelitian ini, peneliti bertujuan untuk membandingkan performa empat algoritma machine learning yaitu Random Forest, Decision Tree, Gradient Boosting, dan Logistic Regression untuk klasifikasi penyakit jantung. Peneliti akan menggunakan dataset yang mencakup data pasien dengan berbagai gejala, riwayat medis, dan hasil tes diagnostik terkait penyakit jantung. Model nantinya akan dievaluasi menggunakan metrik performa seperti recall, dan ROC-AUC untuk

membandingkan efektivitas masing-masing algoritma dalam mengklasifikasikan penyakit jantung. Hasil dari penelitian ini diharapkan dapat memberikan wawasan dalam pemilihan algoritma machine learning yang paling sesuai untuk klasifikasi penyakit jantung.



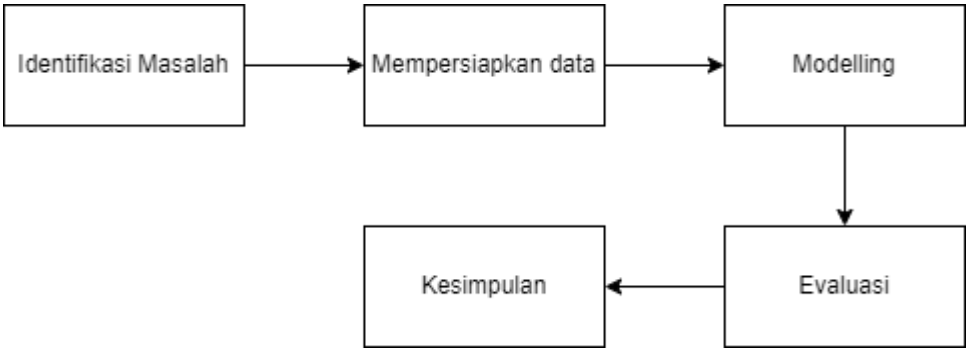
Gambar 1. Penyebab Kematian Terbanyak di Dunia

2. Metode Penelitian

Penelitian terkait perbandingan performa dari algoritma *Machine Learning* ini sudah banyak dilakukan dan dipublikasikan. Penelitian yang sudah dipublikasi akan dikaji agar dapat lebih memahami penelitian yang akan dilakukan. Penelitian perbandingan kinerja algoritma untuk prediksi penyakit jantung dengan teknik *data mining oleh* Derisma ini menghasilkan model dengan tepat dan akurat sebesar 83% menggunakan algoritma Naive Bayes [1]. Adapun penelitian dari Muhammad Aminullah terkait klasifikasi *machine learning* dengan teknik resampling pada dataset tidak seimbang [2]. Penelitian ini membuktikan dengan melakukannya teknik resampling dapat meningkatkan performa klasifikasi *machine learning* pada dataset tidak seimbang dengan akurasi rata-rata sebesar 83% menggunakan algoritma *neural networks* dan teknik *resampling* SMOTEENN. Pada penelitian selanjutnya dilakukan oleh Mufti, Kusrini, dan Sudarmawan dengan judul perancangan sistem klasifikasi penyakit jantung menggunakan Naive Bayes menghasilkan akurasi sebesar 90.61%, nilai presisi 87.44%, dan nilai *recall* 87.95% [3]. Selain itu, penelitian yang dilakukan Riski Annisa yang menganalisa komparasi diantara algoritma decision tree, K-Nearest Neighbour (KNN), Naive Bayes, Random Forest, dan Decision Stump untuk prediksi penderita penyakit jantung [4]. Hasil dari penelitian ini yaitu algoritma random forest mendapatkan performa terbaik sebesar 80.38% berdasarkan akurasinya. Penelitian terakhir yang dikaji yaitu algoritme *stacking* untuk klasifikasi penyakit jantung pada dataset *imbalanced class* dari atik dan yoga yang mendapatkan hasil bahwa penggunaan algoritme *stacking* mampu menjadi solusi untuk permasalahan *imbalanced class* dengan akurasi sebesar 81% dan AUC sebesar 87% [5].

2.1. Tahapan Penelitian

Adapun tahapan penelitian pada kali ini yaitu



Gambar 2. Alur Penelitian

a. Identifikasi masalah
Pada awal penelitian, penulis mengidentifikasi masalah yang ada di dunia nyata dan melakukan kajian dengan hasil penelitian yang sudah didapatkan sebelumnya. Masalah yang ditemukan yaitu banyaknya kematian yang disebabkan oleh penyakit jantung. Selain itu, dokter spesialis jantung dan pembuluh darah (SpJP) di Indonesia yang masih sedikit (sekitar 1.821 dokter) [6].

b. Persiapan Data
Data yang akan digunakan pada penelitian ini adalah data sekunder yang berasal dari *kaggle*. Dataset awal yang digunakan berjumlah 319.795 ribu dengan 18 atribut. Data ini akan dilakukan *preprocessing* seperti *encode* yaitu mengubah data yang awalnya bertipe string menjadi numerikal. Selain itu, data juga akan diperiksa apakah terdapat nilai yang kosong ataupun duplikat.

Tabel 1. Contoh Data Mentah

Heart Disease	BMI	Smoking	Alcohol Drinking	Stroke	Physical Health	Mental Health	Diff Walking	Sex
No	16.60	Yes	No	No	3	30	No	Female
No	20.34	No	No	Yes	0	0	No	Female
No	26.58	Yes	No	No	20	30	No	Male
No	24.21	No	No	No	0	0	No	Female
No	23.71	No	No	No	28	0	Yes	Female
Age Category	Race	Diabetic	Physical Activity	Gen Health	Sleep Time	Asthma	Kidney Disease	Skin Cancer
55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
80 or older	White	No	Yes	Very good	7	No	No	No
65-69	White	Yes	Yes	Fair	8	Yes	No	No
75-79	White	No	No.	Good	6	No	No	Yes
40-44	White	No	Yes	Very good	8	No	No	No

Tabel 2. Penjelasan Dataset

No.	Atribut	Tipe Data	Keterangan
1	Heart Disease	Kategorikal	Pasien yang memiliki penyakit jantung atau tidak
2	BMI	Numerik	Perkiraan lemak tubuh yang didasarkan pada tinggi dan berat badan

No.	Atribut	Tipe Data	Keterangan
3	Smoking	Kategorikal	Pasien yang memiliki riwayat merokok paling sedikit 100 puntung rokok atau 5 bungkus
4	Alcohol Drinking	Kategorikal	Pasien yang memiliki riwayat minum alkohol lebih dari 14 kali selama satu minggu untuk laki-laki dewasa dan lebih dari 7 kali selama satu minggu untuk wanita dewasa
5	Stroke	Kategorikal	Pasien pernah mengalami struk atau tidak
6	Physical Health	Numeric	Berapa lama pasien merasa kesehatan fisik pasien dalam keadaan tidak baik atau terluka dalam rentang 1 bulan terakhir
7	Mental Health	Numeric	Berapa lama pasien merasa kesehatan mental pasien dalam keadaan tidak baik dalam rentang 1 bulan terakhir
8	Diff Walking	Kategorikal	Pasien yang mengalami kesusahan berjalan atau saat menaiki tangga
9	Sex	Kategorikal	Gender pasien
10	Age Category	Kategorikal	Kategori usia (14 kategori)
11	Race	Kategorikal	Ras yang dimiliki pasien
12	Diabetic	Kategorikal	Pasien memiliki riwayat penyakit diabetes atau tidak
13	Physical Activity	Kategorikal	Pasien yang melakukan aktivitas fisik atau olahraga dalam 1 bulan terakhir selain pekerjaan utama pasien
14	Gen Health	Kategorikal	Kesehatan pasien secara umum
15	Sleep Time	Numerik	Rata-rata jumlah waktu tidur dalam 1 hari
16	Asthma	Kategorikal	Pasien memiliki riwayat penyakit asma atau tidak
17	Kidney Disease	Kategorikal	Pasien memiliki riwayat penyakit diabetes atau tidak
18	Skin Cancer	Kategorikal	Pasien memiliki riwayat penyakit kanker kulit atau tidak

c. *Modelling*
Modelling akan dilakukan dengan beberapa algoritma yaitu decision tree, random forest, gradient boosting, dan logistic regression.

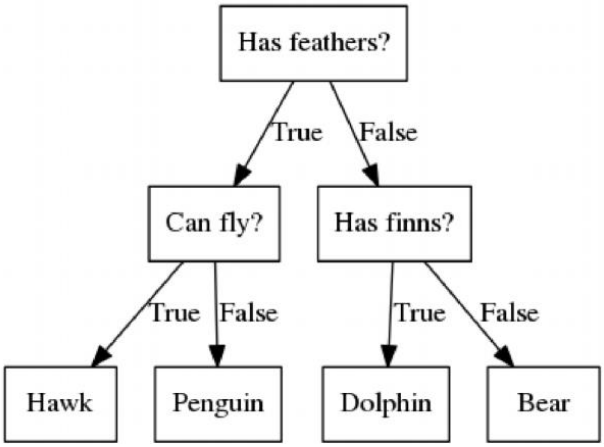
d. Evaluasi
Evaluasi model akan dilakukan dengan cara melihat *metrics score* dari masing-masing model yang telah dibuat, melakukan *cross validation* dari model dengan *metrics score* terbaik, dan melihat ROC *Curve* dari model.

e. Kesimpulan
Tahap terakhir yaitu kesimpulan dimana penulis akan membuat kesimpulan dari hasil penelitian yang telah dikerjakan.

2.2. Decision Tree

Decision tree atau pohon keputusan adalah algoritma yang biasanya digunakan untuk pengambilan keputusan dengan mencari solusi permasalahan berdasarkan kriteria sebagai node

yang terhubung satu sama lain dan membentuk struktur seperti pohon. Setiap pohon ini memiliki cabang yang mewakili suatu atribut wajib untuk dipenuhi agar dapat menuju cabang selanjutnya hingga berakhir di daun [7].



Gambar 3. Contoh Decision Tree

Gambar diatas merupakan contoh dari penggunaan *Decision Tree* dalam klasifikasi hewan dimana setiap atribut akan memiliki nilai tertentu hingga mendapat kesimpulan berupa label (*hawk, penguin, dolphin, bear*). Pada contoh diatas atribut pertama pada hewan yaitu apakah hewan tersebut memiliki bulu, jika iya apakah hewan tersebut dapat terbang, jika iya maka *decision tree* akan memberikan label berupa elang, jika tidak maka hewan tersebut adalah penguin.

2.3. Random Forest

Random forest merupakan kumpulan dari metode klasifikasi decision tree yang dikembangkan berdasarkan pemilihan atribut secara acak pada setiap node untuk menentukan klasifikasi. Proses klasifikasi dari random forest ini akan mengambil suara terbanyak dari pohon keputusan yang dikembalikan. Kelebihan dari random forest ini diantaranya yaitu akurasi yang dihasilkan bagus, cukup baik terhadap data yang memiliki outliers dan noise, dan sederhana serta mudah diparalelkan [8].

2.4. Gradient Boosting

Gradient boosting adalah salah satu teknik dari *machine learning* yang dapat digunakan untuk klasifikasi dan regresi. Model yang dibuat menggunakan algoritma *gradient boosting* ini akan menghasilkan model prediksi *ensemble* dari *weak prediction model* yang biasanya berupa pohon keputusan. Pohon keputusan ini akan dilatih dimana setiap pengamatan diberi bobot yang sama. Setelah melakukan evaluasi terhadap pohon keputusan yang pertama, bobot akan ditambah untuk pengamatan yang sulit diklasifikasikan dan menurunkan bobot untuk pengamatan yang mudah diklasifikasikan [9].

2.5. Logistic Regression

Logistic Regression yaitu teknik analisis data dalam statistika yang dibangun untuk mengetahui relasi dari setiap variabel. *Logistic regression* menggunakan probabilitas untuk memprediksi data kategorikal. Pada *logistic regression* terdapat fungsi sigmoid yang digunakan untuk menggabungkan nilai input secara linear dan nilai koefisien yang digunakan untuk memprediksi hasilnya.

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

Keterangan :

ln(p/1-p)	:	logit dari variabel dependen (yaitu probabilitas sukses dibagi dengan probabilitas gagal)
β0	:	konstanta (intercept) dari model
β1, β2, …, βn	:	koefisien regresi dari masing-masing variabel independen (X1, X2, …, Xn)
X1, X2, …, Xn	:	variabel independen yang digunakan dalam model

2.6. Evaluasi

Evaluasi dari penelitian dibagi menjadi beberapa bagian, yang pertama yaitu melihat dari *metrics score* dari masing-masing model. Adapun beberapa persamaan dari metrics yang akan digunakan yaitu:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

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

$$FPR = \frac{FP}{FP + TN} \quad (4)$$

$$AUC = \frac{\sum(TPR[i-1] - TPR[I])\ X\ (FPR[i] + FPR[i+1])}{2} \quad (5)$$

Keterangan:

TP : *True Positives* (model memprediksi pasien memiliki penyakit jantung dengan benar)
TN : *True Negatives* (model memprediksi pasien tidak memiliki penyakit jantung dengan benar)
FN : *False Negatives* (model memprediksi pasien tidak memiliki penyakit jantung namun kenyataannya pasien memiliki penyakit jantung)
FP : *False Positives* (model memprediksi pasien memiliki penyakit jantung namun kenyataannya pasien tidak memiliki penyakit jantung)

Metriks yang dipilih ini berdasarkan ketidakseimbangan dari hasil prediksi yang menyebabkan diperlukannya metriks selain akurasi untuk menilai performa. Metriks recall digunakan karena pada metriks ini lebih peneliti lebih memilih *false positive* terjadi daripada *false negative*. Jika di analogikan, peneliti lebih memilih membuat model yang memprediksi pasien positif penyakit jantung padahal kenyataannya tidak daripada model memprediksi pasien tidak memiliki penyakit jantung padahal sebenarnya memiliki penyakit jantung yang dapat menyebabkan kematian jika diabaikan.

Selain menggunakan metrik diatas, model juga akan dievaluasi menggunakan teknik *Cross validation. Cross validation* merupakan teknik yang digunakan untuk memvalidasi model dan menilai keakuratan hasil analisis [10]. Dengan menggunakan *cross validation* dapat membantu untuk mengetahui seberapa stabil performa dari model. Sedangkan ROC Curve atau kurva ROC yaitu kurva yang digunakan untuk menggambarkan kinerja dari model dengan melihat prediksi yang benar saja (*True Positive* & *False Positive*). Terdapat cara yang biasanya digunakan untuk menghitung daerah dibawah ROC *curve* ini yaitu dengan *Area Under Curve* (AUC) [11].

3. Hasil dan Pembahasan

Penelitian ini menggunakan algoritma *Random Forest, Decision Tree, Gradient Boosting,* dan *Logistic Regression.* Dataset yang digunakan akan dibagi terlebih dahulu menjadi dua yaitu data pelatihan dan pengujian dengan pembagian skema 80% untuk data latih dan 20% untuk data pengujian. Sebelum dilakukannya pelatihan, data latih akan di *resampling* terlebih dahulu. Pembuatan model akan dilatih berdasarkan data yang tidak di *resampling* dan data di *resampling.* Kemudian, masing-masing dari model akan dilakukan *hyperparameter tuning* untuk

meningkatkan performa dari model. Berikut merupakan hasil performa dari masing – masing model:

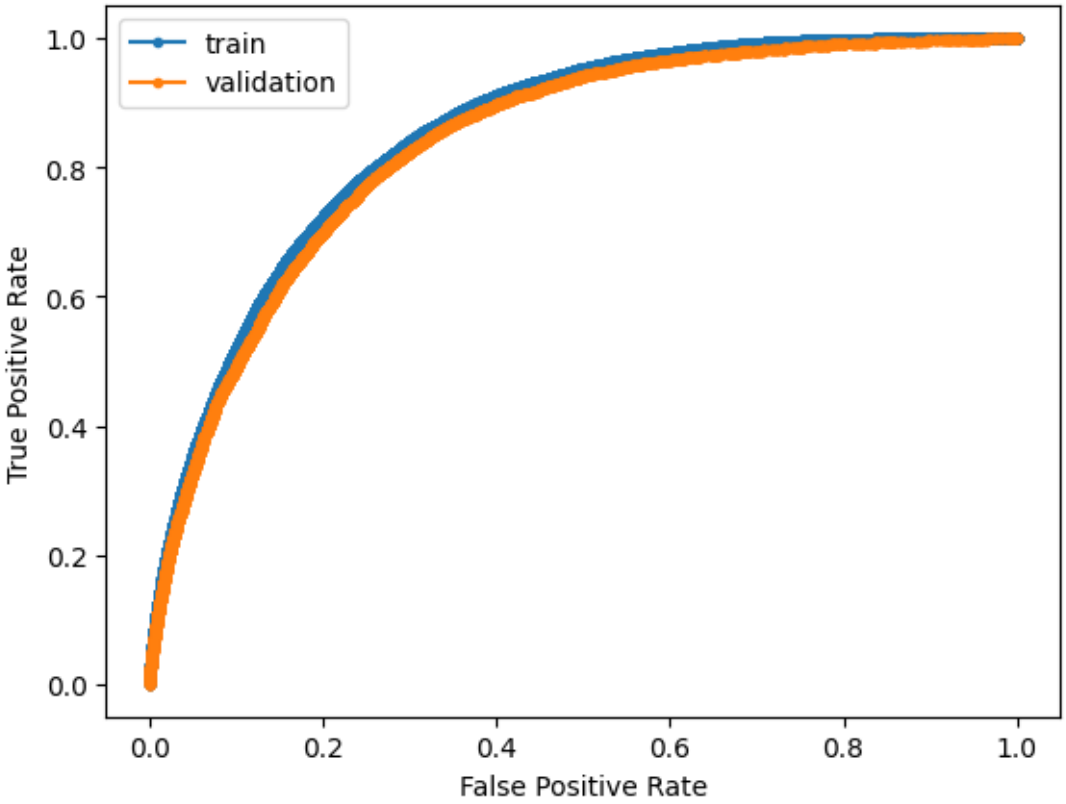
Tabel 3. Metriks Skor dari Masing-masing Model

Model	Recall Train	ROC AUC Train	Recall Test	ROC AUC Test
DT	0.966601	0.983301	0.253886	0.583878
DT_Resampled	0.996702	0.998337	0.301813	0.588921
RF	0.971131	0.985306	0.118986	0.547726
RF_Resampled	0.998756	0.998324	0.304589	0.607655
LR	0.108981	0.549926	0.105477	0.54817
LR_Resampled	0.783834	0.756072	0.778127	0.754826
GBC	0.096171	0.54465	0.089193	0.54092
GBC_Resampled	0.844995	0.853967	0.541821	0.70411
DT_tuning	0.806835	0.762442	0.793856	0.757731
DT_tuning_resampled	0.815344	0.767067	0.77054	0.74509
RF_tuning	0.825411	0.772014	0.806625	0.763451
RF_tuning_resampled	0.82558	0.776804	0.768505	0.748832
GBC_tuning	0.420232	0.660299	0.40322	0.65211
GBC_tuning_resampled	0.811854	0.779775	0.749075	0.749173
LR_tuning	0.779338	0.759305	0.776647	0.760175
LR_tuning_resampled	0.783619	0.755542	0.780163	0.755489

Berdasarkan tabel diatas didapatkan bahwa model Random Forest dengan hyperparameter tuning memiliki performa terbaik yaitu 80,6% pada recall dan 76,3% pada ROC AUC dibandingkan dengan model lainnya.

```
Maximum cross-validation: 0.8194921070693205
Minimum cross-validation: 0.796203110704483
Overall cross-validation: 0.8082083569596922
Standar Deviation: 0.008327737344904655
```

Gambar 4. Hasil Cross Validation



Gambar 5. Kurva ROC

Kemudian dilakukan cross validation dengan pembagian data 5 kali dan menghasilkan nilai recall terbaik sebesar 81.9%, recall terkecil sebesar 79.6%, dan rata-rata dari recall yang didapat sebesar 80.8%. Hal ini menandakan stabilnya performa dari model yang telah dibuat.

4. Kesimpulan

Berdasarkan hasil dari penelitian yang telah dilaksanakan, random forest merupakan algoritma yang cukup baik untuk klasifikasi penyakit jantung dengan nilai recall sebesar 80.6% dan ROC AUC sebesar 76.3%. Dengan melakukan *hyperparameter tuning* pada model mampu membuat performa model menjadi lebih baik. Selain itu, penggunaan teknik *cross validation* membuktikan stabilnya performa dari model yang mendapatkan nilai recall terbaik sebesar 81.9%, recall terkecil sebesar 79.6%, dan rata-rata dari recall yang didapat sebesar 80.8%. Pada penelitian selanjutnya diharapkan dapat memperbanyak fitur atau *independent variable* sehingga mendapatkan performa dari model yang lebih baik.

Daftar Pustaka

[1] D. Derisma, “Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining,” *Journal of Applied Informatics and Computing*, vol. 4, no. 1, 2020, doi: 10.30871/jaic. v4i1.2152.

[2] M. Aminullah, *Perbandingan Performa Klasifikasi Machine Learning dengan Teknik Resampling pada Dataset Tidak Seimbang*. 2021.

[3] M. A. Bianto, K. Kusrini, and S. Sudarmawan, “Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes,” *Creative Information Technology Journal*, vol. 6, no. 1, 2020, doi: 10.24076/citec.2019v6i1.231.

[4] R. Annisa, “Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung,” *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 3, no. 1, 2019.

[5] A. Nurmasani and Y. Pristyanto, “Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class,” *Pseudocode*, vol. 8, no. 1, 2021, doi: 10.33369/pseudocode.8.1.21-26.
[6] K. Nur Azizah, “Menkes Buka-bukaan Singgung ‘Biang Kerok’ Jumlah Dokter Spesialis RI Mandek,” *detikhealth*, Feb. 23, 2023.
[7] F. Y. Pamuji and V. P. Ramadhan, “Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy,” *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 1, 2021, doi: 10.26905/jtmi. v7i1.5982.
[8] L. Ratnawati and D. R. Sulistyaningrum, “Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel,” *Jurnal Sains dan Seni ITS*, vol. 8, no. 2, 2020, doi: 10.12962/j23373520.v8i2.48517.
[9] J. Petrus, “Klasifikasi Mamalia Menggunakan Extreme Gradient Boosting Berdasarkan Fitur Histogram of Oriented Gradient,” 2022.
[10] D. A. Nasution, H. H. Khotimah, and N. Chamidah, “Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN,” *Computer Engineering, Science and System Journal*, vol. 4, no. 1, 2019, doi: 10.24114/cess. v4i1.11458.

Halaman ini sengaja dibiarkan kosong