Computatio: Journal of Computer Science and Information Systems, 8/1 (2024), 96 – 103

PREDIKSI SERANGAN JANTUNG DENGAN MENGGUNAKAN METODE LOGISTIC REGRESSION CLASSIFIER DAN ADABOOST

Steven Dharmawan[1] , Vincent Fernandes[2] , Hizkia Halim[3]
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Univertas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
Email: [1]steven.535180075@stu.untar.ac.id, [2]vincent.535190024@stu.untar.ac.id,
[3]hizkia.535190001@stu.ac.id

Abstrak

Serangan jantung merupakan penyebab kematian nomor 1 di dunia pada tahun 2019[1]. Hal ini mendorong kami untuk membuat sebuah aplikasi yang dapat memprediksi resiko terkena serangan jantung. Penelitian yang kami lakukan menggunakan berbagai metode dan mendapatkan akurasi sebesar 86.8421% menggunakan metode Logistic_Regression, 76.3158% menggunakan metode Decision Tree, 86.8421% menggunakan metode RandomForest Classification, 88.1579% menggunakan metode Bagging Classification, 90.7895% menggunakan metode AdaBoost Classification, 88,1579% menggunakan metode Voting_Classifier. Kami meningkatkan akurasi dari penelitian-penelitian yang dilakukan sebelumnya yaitu dari 88.6% menjadi 90.7895%.

***Kata-Kunci**--Logistic Regression, AdaBoost, Prediksi serangan jantung, heart attack dataset*

Abstract

Heart attack is the number 1 cause of death in the world in 2019. This prompted us to create an application that can predict the risk of having a heart attack. The research we did used various methods and got an accuracy of 86.8421% using the Logistic Regression method, 76.3158% using the Decision Tree method, 86.8421% using the Random Forest Classification method, 88.1579% using the Bagging Classification method, 90.7895% using the AdaBoost Classification method, 88.1579% using the Voting Classifier method. We increased the accuracy of the previous studies from 88.6% to 90.7895%.

***Keywords**: Logistic Regression, AdaBoost,* Heart attack prediction, *heart attack dataset*

1. PENDAHULUAN

Serangan jantung merupakan penyakit yang mematikan bagi seluruh manusia, tanpa jantung tidak ada manusia yang dapat bertahan hidup dan beraktivitas dengan normal. Di tahun 2019 berdasarkan dari *World Health Organization (WHO)*[1] tercatat 17,9 juta manusia meninggal karena *Cardiovascular diseases*. Dari seluruh kematian ini 85% meninggal karena serangan jantung dan *stroke*. Kebanyakan penyakit jantung dikarenakan pola hidup yang tidak sehat misalnya, merokok, pola tidur yang tidak teratur, pola makan yang tidak seimbang, dan mengonsumsi alkohol. Oleh karena itu, penulis bercita-cita menciptakan suatu aplikasi yang dapat membantu seseorang dalam menentukan apakah dirinya berpotensi terkena serangan jantung.

Aplikasi yang tim penulis buat merupakan suatu aplikasi yang akan memprediksi kemungkinan seseorang terkena serangan jantung. Tim penulis berharap aplikasi ini dapat

membantu orang-orang yang khawatir tentang kesehatan jantungnya. Dengan mengetahui potensi terkena serangan jantung, penulis berharap orang tersebut dapat memperbaiki pola hidupnya.

Setelah pencarian beberapa jurnal yang terkait dengan projek, didapatkan 5 buah jurnal. Kelima jurnal tersebut memuat topik yang sama yaitu *Heart Attack Prediction* dengan metode yang berbeda-beda. Jindall et al (2020) menggunakan metode KNN, Logistic Regression, dan Random Forest Classifier [2], Prasad et al (2019) menggunakan metode Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naïve Baiyes, dan Decision Tree [3], Bhat et al (2020) menggunakan metode Logistic Regression, Support Vector Machine, dan K-Nearest Neighbors [4], Galla et al (2020) menggunakan metode Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Support Vector Machine, Naïve Baiyes, Decision Tree dan Random Forest [5], Shah et al (2020) menggunakan metode Naïve Baiyes, Decision Tree, K-Nearest Neighbors, Random Forest [6]. Kelima Jurnal tersebut menggunakan dataset yang sama yaitu Heart Attack Analysis & Prediction Dataset [7].

Jindall et al (2020) melakukan pengujian menggunakan K-Nearest Neighbors, Logistic Regression dan Random Forest [2]. Jindall et al (2020) ini menyimpulkan K-Nearest Neighbors mempunyai akurasi tertinggi dengan 88.52% [2]. Prasad et al (2019) melakukan pengujian menggunakan Decision Tree dengan akurasi 78.69%, K-Nearest Neighbors dengan akurasi 77.05%, Logistic Regression dengan akurasi 86.89% [3]. Bhat et al (2020) melakukan pengujian menggunakan Logistic Regression, Support Vector Machine, dan K-Nearest Neighbors. Bhat et al (2020) mendapatkan akurasi tertinggi menggunakan metode Logistic Regression dengan akurasi 85% [4]. Galla et al (2020) melakukan pengujian menggunakan Decision Tree dengan akurasi 80.33%, Logistic Regression dengan akurasi 86.89%, K-Nearest Neighbors dan Random Forest dengan akurasi 88.6% [5]. Shah et al (2020) melakukan pengujian menggunakan Naïve Baiyes dengan akurasi 88.157%, 90.789%, Decision Tree dengan akurasi 80.263%, Random Forest dengan akurasi 86.84% [6].

2. METODE PENELITIAN

2.1. Data

Data yang digunakan merupakan data riwayat kesehatan yang berhubungan dengan serangan jantung. Data yang digunakan didapat dari Kaggle [7]. Terdapat 303 data dan 14 feature seperti yang dijelaskan pada Tabel 1.

Tabel 1 Dataset

Data	Description
age	Age of the patient
sex	Sex of the patient
exng	Exercise induced angina (1 = yes; 0 = no)
caa	number of major vessels (0-4) colored by flourosopy
CP	Chest Pain type chest pain type • Value 1: typical angina • Value 2: atypical angina • Value 3: non-anginal pain • Value 4: asymptomatic
trtbps	resting blood pressure (in mm Hg)
chol	cholestoral in mg/dl fetched via BMI sensor
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
oldpeak	ST depression induced by exercise relative to rest

Data	Description
restecg	resting electrocardiographic results • Value 0: normal • Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) • Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved
slp	Slope of peak exercise ST segment • Value 0: upsloping • Value 1: flat • Value 2: downsloping
thall	A blood disorder called thalassemia • Value 0: NULL (dropped from the dataset previously • Value 1: fixed defect (no blood flow in some part of the heart) • Value 2: normal blood flow
target	0 = less chance of heart attack, 1 = more chance of heart attack

Pada data yang kami peroleh, value 0 pada fitur thall merupakan data *NULL* sehingga kami mengganti value 0 tersebut dengan mode dari fitur thall. Selanjutnya, kami melakukan pengecekan terhadap data duplikat dan kami menemukan adanya satu buah data duplikat. Setelah menemukannya kami melakukan pembersihan kepada data duplikat.

2.2 *Algoritma*

Pada projek yang kami kerjakan, kami menggunakan metode *Logistic Regression* dengan boosting menggunakan AdaBoost.

2.2.1 *Logistic Regression*

Logistic regression sangat sering digunakan dalam pengaplikasian machine learning. Model ini mengambil variabel vektor dan mengevaluasi koefisien atau bobot dari setiap variabel input dan memprediksi kelas target [9]. Secara matematik fungsi logistic regression didefinisikan menggunakan persamaan (1)[9]:

$$\text{Logit(S)} = b_0 + b_1 M_1 + b_2 M_2 + b_3 M_3 \ldots . b_k M_k \quad (1)$$

S merupakan probabilitas dari fitur pilihan, M merupakan nilai yang digunakan untuk memprediksi dan b merupakan *intercept* dari model.

2.2.2 AdaBoost

Ketika melakukan pengujian model menggunakan *Logistic Regression*, kami mendapatkan akurasi yang tidak memuaskan sehingga kami memutuskan untuk melakukan peningkatan akurasi menggunakan AdaBoost. Metode AdaBoost dapat meningkatkan akurasi dengan cara membangkitkan kombinasi dari suatu model, tetapi hasil klasifikasi atau prediksi yang dipilih adalah model yang memiliki nilai bobot yang paling besar. Jadi, setiap model yang dibangkitkan memiliki atribut berupa nilai bobot. Dataset yang telah seimbang akan divalidasi dengan menggunakan 10-*fold cross validation*. Hasil dari validasi akan menghasilkan data yang diukur yaitu AUC dan Akurasi [8].

Algoritma AdaBoost

1. Inisialisasi menggunakan persamaan (2) [8]

$$D_1(i) = \frac{1}{m} \quad (2)$$

2. for t = 1, … , T:
3. Pengujian terhadap distribusi D_t.
4. Mendapatkan *hypothesis* menggunakan persamaan (3) [8]

$$h_i : X \rightarrow \{-1, +1\} \tag{3}$$

dengan error yang didapatkan dari persamaan (4) [8]

$$\varepsilon_t = Pr_{x \sim D_t}[h_t(X_i) \neq y_i]. \tag{4}$$

5. Memilih α_t dengan persamaan (5) [8]

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right). \tag{5}$$

6. Memperbaharui D menggunakan persamaan (6) [8]

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} X \begin{cases} e^{-\alpha_t} & if\ h_t(X_i) = y_i \\ e^{\alpha_t} & if\ h_t(X_i) \neq y_i \end{cases} \tag{6}$$

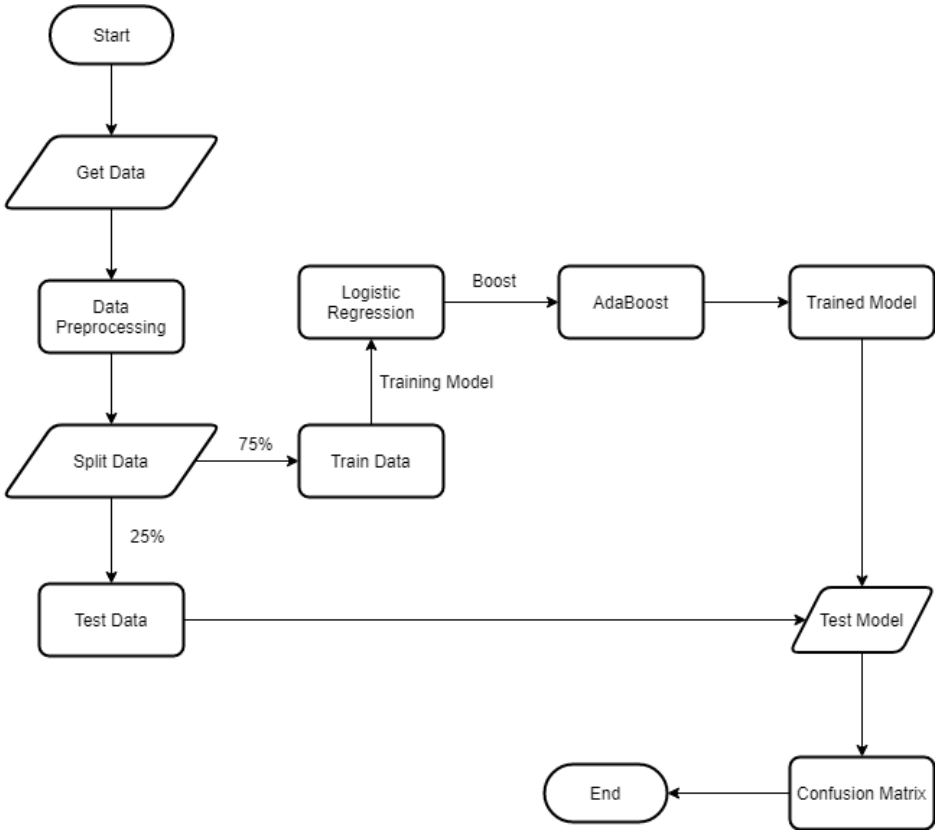
$$= \frac{D_t(i) e^{-\alpha_t y_i h_i(X_i)}}{Z_t}$$

7. Mengeluarkan Output menggunakan persamaan (7) [8]

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{7}$$

2.3 Rancangan Eksperimen

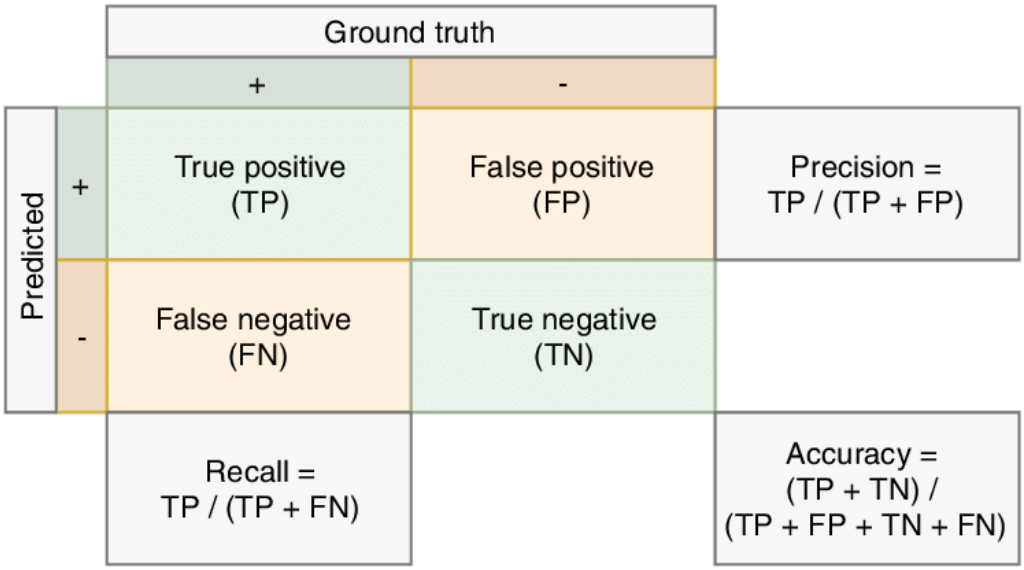
Dalam pembuatan program untuk memprediksi tingkat kemungkinan terkena penyakit serangan jantung, kami membuat program berjalan sesuai alur Flowchart pada Gambar 1.



Gambar 1 FlowChart

2.4 Metode evaluasi

Kami menggunakan metode evaluasi *confusion matrix* dan *classification report. Confusion Matrix* merupakan tentang informasi aktual dan prediksi klasifikasi yang dilakukan oleh sistem klasifikasi [10]. Kinerja atau performa sistem klasifikasi tersebut biasanya dievaluasi menggunakan data dalam matriks [10]. Pada confusion matrix terdapat 4 klasifikasi hasil prediksi yaitu *True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).* Dari *Confusion Matrix* dapat dihitung untuk Akurasi, Presisi, dan Recall. Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih [10]. Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia [10]. Untuk data dua kelas classifier ditampilkan pada Gambar 2 [11].



Gambar 2 Confusion Matrix

Perhitungan pada *Confusion Matrix* berdasarkan pada Gambar 2 sebagai berikut:

Perhitungan Akurasi didapatkan menggunakan persamaan (8) [11]:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

Perhitungan Presisi didapatkan menggunakan persamaan (9) [11]:

$$Presisi = \frac{TP}{TP+FP} \quad (9)$$

Perhitungan Recall didapatkan menggunakan persamaan (10) [11]:

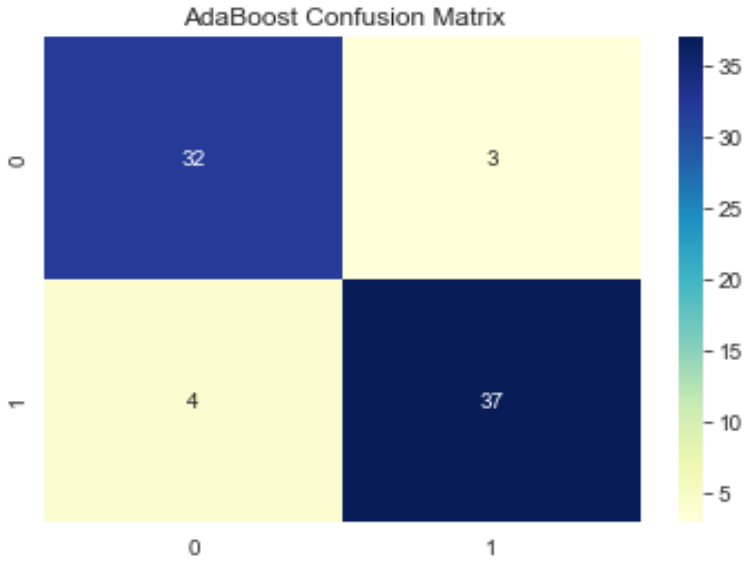
$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1 Score merupakan rata-rata tertimbang dari presisi dan daya ingat untuk kelas itu [12]. Secara umum memberikan gambaran yang lebih besar tentang bagaimana kinerja model untuk label itu dan jelas semakin tinggi angka ini semakin baik [12]. F1 Score dapat dihitung menggunakan persamaan (11) [11]:

$$F_1 = \frac{2}{Recall^{-1}+Precision^{-1}} \quad (11)$$

2. HASIL DAN PEMBAHASAN

Pada pengujian yang kami lakukan, kami menggunakan 226 data training dan 76 data testing yang terdiri dari 35 data dengan value 0 dan 41 data dengna value 1. Setelah melakukan pengujian dengan menggunakan metode yang berbeda-beda. Kami mendapatkan akurasi sebesar 86.8421% menggunakan metode *Logistic_Regression*, 76.3158% menggunakan metode *Decision Tree,* 86.8421% menggunakan metode *RandomForest Classification*, 88.1579% menggunakan metode *Bagging Classification*, 90.7895% menggunakan metode *AdaBoost Classification*, 88,1579% menggunakan metode *Voting_Classifier*. Dari akurasi metode-metode diatas kami mendapatkan metode *AdaBoost* adalah metode dengan akurasi terbaik. *Confusion Matrix* dari metode *AdaBoost* dapat dilihat pada Gambar 3 dan *Classification Matrix* dapat dilihat pada Gambar 4.

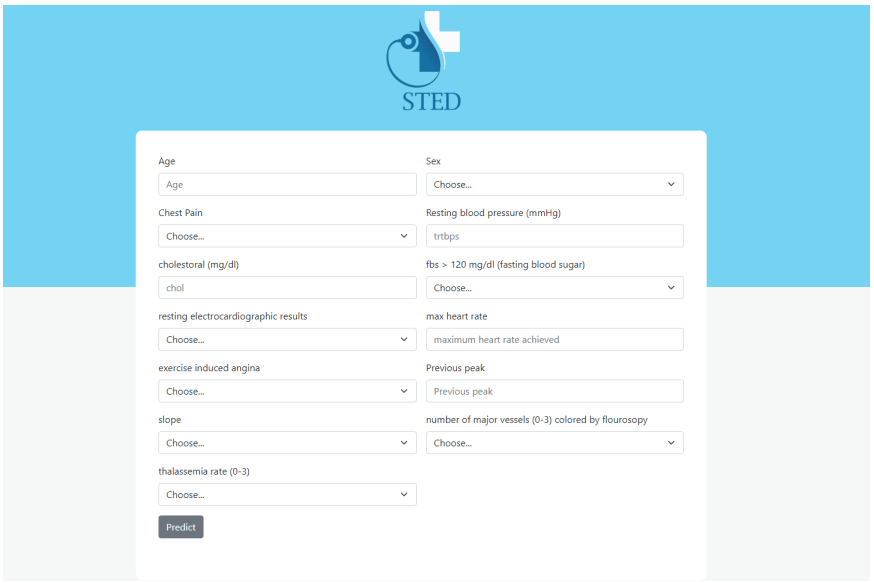


Gambar 3 Confusion Matrix AdaBoost

	precision	recall	f1-score	support
0	0.86	0.89	0.87	35
1	0.90	0.88	0.89	41
accuracy			0.88	76
macro avg	0.88	0.88	0.88	76
weighted avg	0.88	0.88	0.88	76

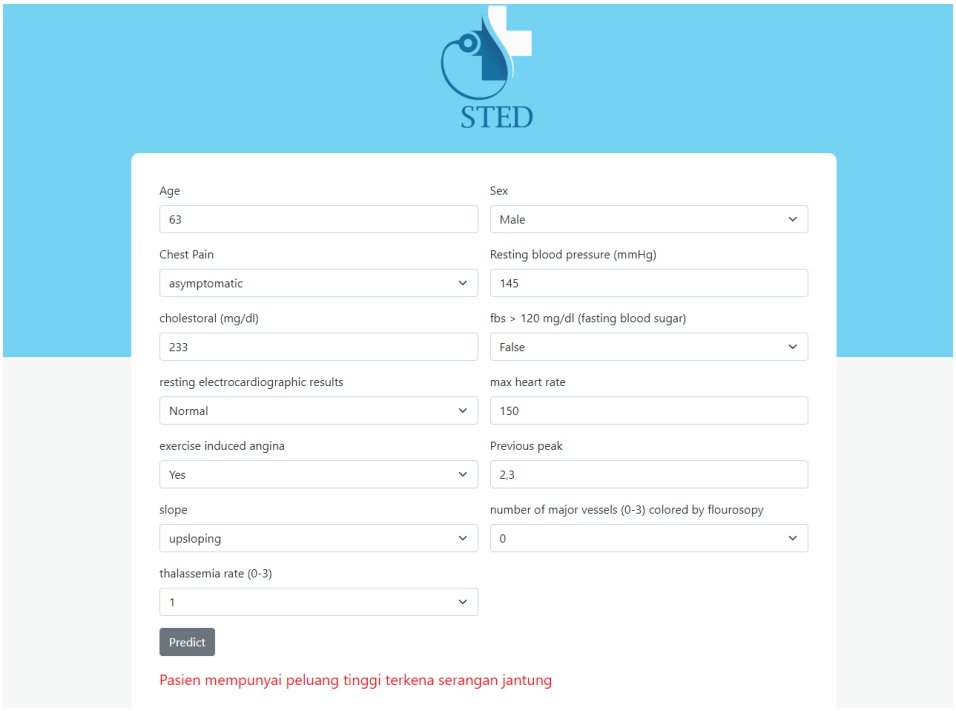
Gambar 4. *Classification Report AdaBoost*

Setelah kami selesai melakukan pengujian, kami selanjutnya mengaplikasikan hasil pengujian kedalam bentuk website. Website yang kami buat bertujuan agar user dapat melakukan prediksi serangan jantung dengan mudah. Website yang kami buat menggunakan metode AdaBoost sebagai metode untuk melakukan prediksi serangan jantung. Tampilan UI pada website dapat dilihat pada Gambar 5.



Gambar 5. *UI Heart Attack Predictor*

Seperti yang dapat dilihat pada Gambar 5 input yang dibutuhkan ada 13 data yaitu *Age, Sex, Chest Pain, Resting blood pressure, Cholestoral*, *fbs, resting electrocardiographic results, Max heart rate, Exercise induced angina, Previous peak, Slope, Number of major vessels colored by fluorospy, thalassemia rate*. Hasil output dari prediksi dapat dilihat pada Gambar 6.



Gambar 6 Tampilan hasil output dari prediksi

Pada website yang kami buat, kami menggunakan data uji seperti yang terlihat pada Gambar 6 dan menunjukkan hasil pasien mempunyai peluang tinggi terkena serangan jantung. Website yang kami buat dapat diakses pada halaman *https://sted-final.herokuapp.com/*.

4. KESIMPULAN

Kesimpulan dari penelitian yang kami lakukan adalah metode Logistic Regression menggunakan AdaBoost mempunyai kecocokan yang tinggi dengan dataset yang kami gunakan. Hal ini dapat dilihat dari tingkat akurasi yang tinggi pada model yang kami buat yaitu sebesar 90.7895%. Penelitian yang kami buat memiliki tingkat akurasi yang lebih tinggi dari penelitian-penelitian sebelumnya namun penelitian yang kami buat memiliki kekurangan yaitu jumlah data yang tergolong sedikit sehingga penambahan data dapat berisiko mengurangi akurasi yang didapatkan saat ini. Penelitian yang kami buat masih memungkinkan untuk dikembangkan lebih lanjut dengan cara penambahan jumlah data yang digunakan, penyesuaian metode yang lebih baik, dan hasil output yang dapat dibuat kedalam bentuk persentase agar user dapat mengetahui secara detail tingkat resiko serangan jantung yang dimiliki.

DAFTAR PUSTAKA

[1] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Jindall, H., Agrawall, S., Kheral, R., Jain, R., dan Nagrath, P., 2020, Heart disease prediction using machine learning algorithms, *https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf,* diakses tanggal 1 November 2021.

[3] Prasad, R., Anjali, P., Adil, S., dan Deepa, N., 2019, Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning, *https://www.ijeat.org/wp-content/uploads/papers/v8i3S/C11410283S19.pdf*, diakses tanggal 1 November 2021.

[4] Bhat, A., Pragathi., Pranamya, M., dan Smitha., 2020, Prediction of Heart Disease Using Logistic Regression, *https://www.irjet.net/archives/V7/i6/IRJET-V7I6310.pdf*, diakses tanggal 1 November 2021.

[5] Galla, S, S, B., Munaga, M., Manchuri, S, R., Rajalakshmi, 2020, Heart Disease Prediction Using Machine Learning Techniques, *https://www.researchgate.net/publication/344557562_Heart_Disease_Prediction_Using_Machine_Learning_Techniques*, diakses pada tanggal 1 November 2021.

[6] Shah, D., Patel, S., dan Santosh, K. B., 2020, Heart Disease Prediction using Machine Learning Techniques*, https://link.springer.com/article/10.1007/s42979-020-00365-y*, diakses pada tanggal 1 November 2021.

[7] *www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?select=heart.csv*

[8] Bisri, A., Romi, S, W., 2015, Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree, *https://media.neliti.com/media/publications/243690-penerapan-adaboost-untuk-penyelesaian-ke-f7fdc8fc.pdf, diakses pada 8 November 2021.*

[9] Prabhat, A., dan Kullar, V., 2017, Sentiment classification on Big Data using Naïve Bayes and Logistic Regression.

[10] Yohana, T, U., Dewi, A, S., Heningtyas, Y., 2020, Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi, *http://repository.lppm.unila.ac.id/25817/1/2647-6378-1-PB.pdf*, diakses pada 9 November 2021.

[11] Jacob, H, J., Rune, H, J., Inceoglu, F., Thomas, S, T., 2019, A Cloud Detection Algorithm for Satellite Imagery Based on Deep Learning, *https://www.researchgate.net/publication/334840641_A_cloud_detection_algorithm_for_satellite_imagery_based_on_deep_learning*, diakses pada 9 November 2021.

[12] Lingga, A, A., Pratiwi, A, M, A., Respatiwulan, 2019, Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier, *https://jurnal.uns.ac.id/ijas/article/viewFile/29998/21230*, diakses pada 9 November 2021.