Perbandingan kinerja Algoritma Klasifikasi untuk mendeteksi Penyakit Jantung

Chepy Bagustian Sonjaya
Universitas Buana Perjuangan
Karawang, Indonesia
if19.chepysonjaya@mhs.ubpkarawang.ac.id

Anis Fitri Nur Masruriyah
Universitas Buana Perjuangan
Karawang, Indonesia
anis.masruriyah@ubpkarawang.ac.id

Dwi Sulistya Kusumaningrum
Universitas Buana Perjuangan
Karawang, Indonesia
dwi.sulistya@ubpkarawang.ac.id

***Abstract*— Penyakit jantung di Indonesia terutama pada usia produktif selalu terjadi kenaikan jumlah kasus. Adapun penyebab utama terjadinya kenaikan jumlah pasien jantung adalah gaya hidup dan pola makan yang tidak sehat. Meningkatnya pasien penyakit jantung juga berdampak pada penurunan taraf hidup. Dengan adanya hal tersebut, perlu adanya penelitian terkait membandingkan metode klasifikasi pada dataset penyakit jantung. Metode penelitian ini menggunakan model algoritma Support Vector Machine (SVM) dan Logistic Regression (LR). Agar penelitian mendapatkan hasil yang akurat digunakan teknik akuisisi data, pra-pemrosesan data dan transformasi data. Teknik evaluasi model yang digunakan yaitu K-Fold Cross Validation. Hasil analisis menunjukkan bahwa teknik validasi k-fold cross validation memberikan akurasi yang sama baiknya, tetapi hasil presisi relatif rendah. Algoritma SVM menghasilkan akurasi sebesar 91,57%, sedangkan LR menghasilkan akurasi sebesar 91,66%. Akan tetapi, SVM memiliki nilai presisi sebesar 61,20%, sedangkan LR memiliki presisi 54,31%.**

***Kata kunci — Cross Validation,* Klasifikasi*, Logistic Regression,* Penyakit jantung, SVM.**

I. PENDAHULUAN

Penyakit jantung sering disebut sebagai gangguan yang mempengaruhi kinerja jantung [1]. Ada berbagai jenis dan nama penyakit jantung, seperti penyakit jantung koroner, irama jantung yang tidak normal, penyakit jantung bawaan, kelainan pada katup, gagal jantung, dan serangan jantung [2]. Menurut data WHO tahun 2019, 17.9 juta orang meninggal karena penyakit jantung di seluruh dunia. WHO menyatakan bahwa penyakit jantung adalah salah satu penyebab utama kematian di Inggris, Amerika Serikat, Indonesia, dan hampir di seluruh negara lainnya [3].

Menurut data Kementerian Kesehatan Republik Indonesia [4], penyakit jantung mengalami peningkatan setiap tahun dan menjadi penyebab utama kematian di Indonesia, terutama pada masa usia produktif. Hal ini disebabkan oleh gaya hidup tidak sehat dan pola makan tidak seimbang yang meningkatkan prevalensi penyakit jantung dalam suatu populasi. Selama masa pandemi COVID-19, orang yang memiliki penyakit jantung bawaan memiliki risiko yang sangat tinggi, karena dikhawatirkan dapat memburuk kondisi mereka dan bahkan menyebabkan kematian. Tanpa penanganan yang baik, pasien penyakit jantung pada usia produktif bisa membahayakan nyawanya. Oleh karena itu, ada kebutuhan akan model klasifikasi menggunakan machine learning yang membantu layanan kesehatan dalam melakukan penanganan dan pencegahan terhadap penderita penyakit jantung.

Machine learning dapat membantu dokter dalam melakukan klasifikasi pada pasien penyakit jantung, karena penelitian medis selalu mengumpulkan data besar. Namun, pengolahan data besar yang salah dapat mempengaruhi akurasi hasil. Tugas pengolahan data besar yang kompleks membutuhkan keahlian dan pengalaman yang tinggi, seperti yang diterangkan dalam jurnal Utomo et al. [5].

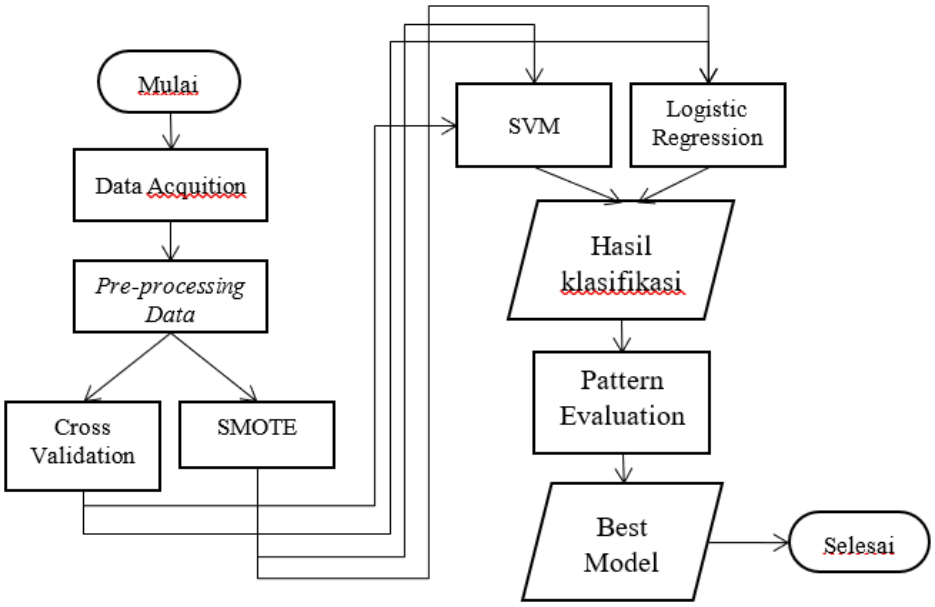
Menurut Mezzatesta et al. [6] melakukan penelitian untuk memprediksi tingkat kematian pada pasien Penyakit Jantung Koroner yang menjalani dialisis dengan menggunakan algoritma *Support Vector Machine* (SVM) dan *kernel Radial Basis Function* yang dioptimalkan dengan *GridSearch*. Hasil dari model prediksinya memiliki akurasi sebesar 95,25%. Ghosh et al. [7] juga melakukan penelitian dengan memanfaatkan pembelajaran mesin untuk memprediksi penyakit kardiovaskular dengan menerapkan algoritma ekstrasi fitur RELIEF dan LASSO. Hasil dari penelitiannya memiliki akurasi sebesar 99,05%. Terakhir, El-Hasnony et al. [8] membuat model pencegahan untuk penyakit stroke dan jantung dengan menerapkan active learning. Tujuannya adalah untuk menentukan faktor-faktor yang paling berpengaruh pada kedua penyakit tersebut sehingga tenaga medis dapat memberikan perawatan yang tepat untuk mencegah penyakit stroke dan jantung.

Beberapa referensi penelitian menjelaskan bahwa model klasifikasi dapat membantu tenaga medis dalam mengatasi masalah kesehatan jantung. Model klasifikasi juga terbukti dapat melakukan prediksi yang baik dalam bidang lain. Tujuan dari penelitian ini adalah untuk menemukan model terbaik dengan menerapkan beberapa algoritma dalam kasus penyakit jantung, serta menerapkan teknik ekstraksi fitur untuk mengidentifikasi variabel yang paling mempengaruhi penyakit jantung.

II. METODELOGI PENELITIAN

A. Alur Penelitian

Dalam penelitian ini, dataset penyakit jantung digunakan untuk melakukan klasifikasi menggunakan pembelajaran mesin. Metode yang digunakan termasuk *Support Vector Machine* (SVM) dan *Regresi Logistik* (LR). Dataset dibagi menjadi beberapa bagian dan dites menggunakan kedua metode untuk menentukan performa masing-masing. Alur penelitian dapat ditemukan pada Gambar 1.



Gambar 1. Alur Penelitian

B. Sumber data

Penelitian ini menggunakan data pasien penyakit jantung sebanyak 319.795 objek dengan 17 variabel dan 1 kelas target. Data yang digunakan adalah data kuantitatif yang berupa angka-angka yang akan diolah menggunakan metode statistika. Sumber data diambil dari catatan medis yang disetujui oleh Organisasi Kesehatan Dunia dan tersedia pada tahun 2022. Variabel-variabel yang dianalisis meliputi BMI, *Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer dan HeartDisease.*

Penelitian ini mengimplementasikan empat fase analitik (analitik kualitas data, analitik deskriptif, analitik diagnostik, dan analitik prediktif). Fase pertama meliputi pra-pemrosesan data termasuk analitik kualitas data dan analitik deskriptif. Setelah itu, hasil dari pra-pemrosesan data diproses untuk menghasilkan analitik diagnostik dan analitik prediktif.

C. *SVM (Support Vector Machine)*

SVM atau *Support Vector Machine* adalah metode yang digunakan untuk masalah klasifikasi, terutama pada analisis data. Algoritma ini membantu memisahkan data ke dalam dua kategori melalui pencarian *hyperplane* pemisah yang paling baik. *Hyperplane* ini bisa berupa garis pada dua dimensi, atau bidang datar pada dimensi tinggi. SVM memfokuskan analisis pada penentuan *hyperplane* yang memisahkan data dengan cara yang optimal [9].

Dalam algoritma SVM, pola dapat dikelompokkan menjadi kategori positif atau negatif dengan menggunakan garis pemisah yang berbeda. Margin adalah jarak antara hyperplane dan pola terdekat dalam setiap kelas. Pola yang disebut *support vector* membantu dalam memprediksikan hasil yang mungkin terjadi. Proses pembelajaran pada SVM berkonsentrasi pada menemukan lokasi *hyperplane* yang tepat [10].

Dalam algoritma SVM, ada dua jenis pemisahan kelas oleh *Hyperplane*, yaitu pemisahan sempurna oleh kelas yang disebut SVM linier dan pemisahan yang tidak sempurna oleh kelas disebut SVM nonlinier. SVM nonlinier merupakan solusi dari masalah SVM linier dengan menerapkan fungsi *kernel* pada ruang fitur dengan dimensi yang lebih tinggi [11]. Definisi SVM linier dan nonlinier dapat ditemukan pada Tabel 1.

Tabel 1. Definisi SVM linier dan nonlinier

SVM	Jenis Kernel	Definisi Rumus
Linier	Linier	K (x,y) = x.y
Non linier	*Polynomial*	K (x,y) = $(x.y + 1)^p$
	RBF	K (x,y) = $e^{-\|x.y\|^2/2\sigma^2}$
	Sigmoid	(x,y) = tanh (Kx.y- δ)

D. *LR (Logistic Regression)*

Logistic Regression adalah sebuah algoritma yang digunakan untuk memprediksi masalah regresi dan klasifikasi. Algoritma ini bisa digunakan untuk memprediksi data kategorikal dalam bentuk klasifikasi. *Logistic Regression* bisa digunakan untuk menganalisis variabel dependen yang bersifat dikotomi. Dalam hal ini, peristiwa dapat dikatakan terjadi atau tidak berdasarkan angka yang berkisar antara 0 sampai 1. Bila digunakan ambang keputusan, hal tersebut bisa menjadi masalah klasifikasi dan membedakan jenis yang berbeda. Algoritma ini juga dikenal dengan sebutan regresi logistik, pengklasifikasi log-linear, atau klasifikasi entropi maksimum (MaxEnt) [11].

Regresi logistik adalah jenis regresi yang tidak mengasumsikan adanya hubungan linier antara variabel independen dan dependen. Sebaliknya, regresi logistik memiliki pola kurva linier. Proses ini melibatkan kombinasi linear dari variabel independen menjadi variabel prediktor. Variabel prediktor ini kemudian dikonversi menjadi probabilitas melalui fungsi logistik. [12].

$$p = \frac{1}{1+e^{-(a+bX)}} \quad (1)$$

Dalam persamaan (1), "p" mewakili probabilitas, "a" dan "b" adalah parameter model, dan "X" adalah variabel faktor.

E. *Confussion Matrix dan K-Fold*

Evaluasi dilakukan untuk memilih dataset dan metode klasifikasi yang dapat memberikan tingkat akurasi yang optimal. Dalam penelitian ini, evaluasi dilakukan dengan memperhatikan *Confussion Matrix*. *Confussion Matrix* adalah alat untuk menentukan sejauh mana performa pengklasifikasi dalam mengidentifikasi atau memprediksi kelas dalam data [15]. Matriks konfusi dihitung menggunakan model yang bertujuan untuk mengetahui seberapa baik performa model ketika menggunakan data uji. Evaluasi didasarkan pada presisi dan akurasi. Perhitungan ini dapat dilihat dalam tabel 2 dan diterangkan dalam persamaan 2 dan 3.

Tabel 2. Confussion Matrix

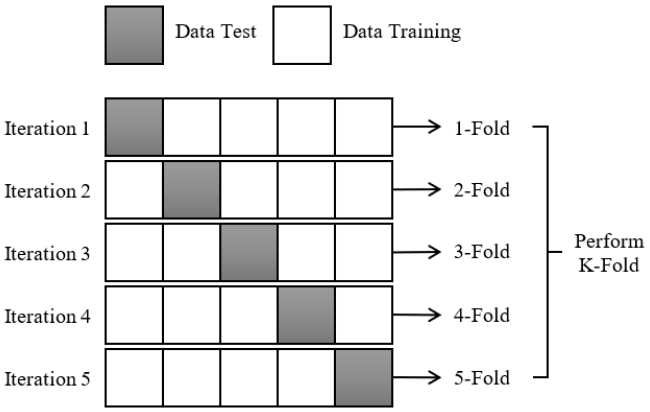
	Aktual	
Prediksi	A	B
A	TP (*True Positive*)	FP (*False Positive*)
B	FN (*False Negative*)	TN (*True Negative*)

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Presisi} = \frac{True\ Positive}{True\ Positive + False\ positive} \quad (3)$$

Pada persamaan (2) bertujuan untuk mengukur tingkat akurasi dalam memprediksi *true positive* dan *true negative* dari keseluruhan data. Ia mengukur seberapa baik prediksi dilakukan dalam menentukan pasien terkena atau tidaknya penyakit jantung. Persamaan (3) memfokuskan pada presisi dengan membandingkan *true positive* dengan jumlah data yang diprediksi positif. Ini memberikan pernyataan tentang seberapa besar persentase pasien yang terdiagnosis menderita penyakit jantung dari seluruh pasien yang diprediksi menderita penyakit jantung.

Selanjutnya, evaluasi pada penelitian dilakukan melalui Teknik *K-Fold Cross Validation*. Proses ini melibatkan pembagian data menjadi K bagian data uji dan data latih. Ilustrasi Teknik *K-Fold Cross Validation* dapat dilihat pada Gambar 2.



Gambar 2. Ilustrasi Teknik *K-Fold Cross Validation*

III. HASIL DAN PEMBAHASAN

A. Pra-pemrosesan Data

Penelitian ini menggunakan dataset dari catatan medis yang disetujui oleh Organisasi Kesehatan Dunia dan diambil pada tahun 2022 dengan jumlah data sebanyak 319.795 dan 17 atribut. Data tersebut dibersihkan untuk menghilangkan duplikat dan data bernilai null. Data kategori dikonversikan menjadi angka, dengan 1 = yes/male dan 0 = no/female untuk mempermudah pemrosesan oleh mesin yang hanya dapat membaca data numerik. Sementara data numerik dik normalisasi untuk menyamakan skalanya sehingga mempermudah proses pengujian.

B. Pemodelan Data

Pada tahap *modelling*, dilakukan proses klasifikasi pada data yang telah dikumpulkan. Data tersebut terlebih dahulu dipartisi menjadi data uji dan data latih menggunakan teknik *k-fold cross validation*. Setelah melalui proses *pre-processing*, data tersebut dilanjutkan dengan proses pembelajaran *supervised learning* menggunakan metode SVM dan *Regresi Logistik*. Hasil dari kedua metode tersebut kemudian akan dibandingkan untuk menentukan metode yang paling baik untuk data tersebut.

C. Pengujian

Dalam penelitian ini, metode *k-fold cross validation* digunakan untuk pengujian data untuk mengurangi bias yang mungkin terdapat dalam data acak. Dalam hal ini, dataset dibagi menjadi 10 bagian yang berukuran sama dan dilakukan proses pelatihan dan pengujian pada model sebanyak 10 kali. Kemudian, klasifikasi dilakukan menggunakan algoritma *Support Vector Machine* dan *Logistic Regression*, dan hasilnya dalam bentuk nilai akurasi dan presisi dapat ditemukan pada Tabel 3, dengan hasil k-fold 1 sampai k-fold 10.

Tabel 3. Hasil Akurasi dan presisi K-fold 10

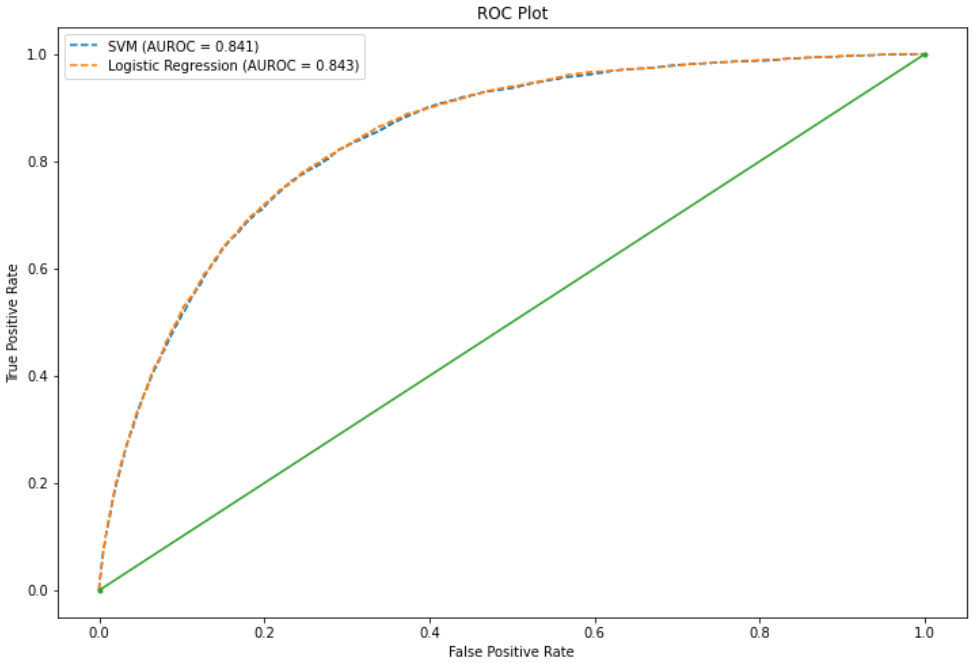
K-Fold	SVM		LR	
	Akurasi	Presisi	Akurasi	Presisi
1	91.56%	60.70%	91.58%	54.39%
2	91.54%	61.12%	91.56%	54.61%
3	91.57%	60.80%	91.59%	54.28%
4	91.58%	60.39%	91.60%	54.19%
5	91.53%	60.46%	91.56%	54.58%
6	91.57%	61.12%	91.60%	54.77%
7	91.60%	61.02%	91.61%	54.40%
8	91.57%	60.36%	91.57%	53.71%
9	91.57%	60.40%	91.59%	54.41%
10	91.57%	61.20%	91.66%	54.31%

Tabel 3 menunjukkan bahwa performa terbaik dalam hal akurasi diperoleh oleh algoritma SVM dan LR saat melakukan validasi dengan metode K-*fold* 4, dengan hasil akurasi SVM sebesar 91,58% dan LR 91,60%. Sedangkan, hasil presisi terbaik ditemukan pada K-*fold* 10 dengan algoritma SVM memiliki presisi 61,20%.

D. Evaluasi

Evaluasi dilakukan dengan mempertimbangkan akurasi, presisi, sensitivitas, dan spesifisitas. Proses ini menggunakan data pengujian yang telah dipisahkan sebelumnya dan hasil evaluasi ditampilkan dalam bentuk matriks konfusi yang ditunjukkan dalam tabel 2.

Hasil perhitungan menggunakan Confusion Matrix dan 10-Fold Cross Validation menunjukkan bahwa model yang dibangun dengan menggunakan algoritma SVM memiliki tingkat akurasi sebesar 91,57%. Namun, ketika algoritma Logistic Regression digunakan, model klasifikasi yang dihasilkan memiliki tingkat akurasi yang lebih tinggi, yaitu sebesar 91,66%. Oleh karena itu, dapat disimpulkan bahwa Logistic Regression lebih baik dalam membangun model klasifikasi dibandingkan SVM untuk dataset penyakit jantung. Indeks ROC Area digunakan sebagai metrik untuk mengevaluasi kinerja model dalam memprediksi. Indeks ROC (Receiver Operating Characteristics) digunakan untuk menganalisis kinerja suatu model algoritma [13].



Gambar 3. Evaluasi SVM dan LR

Evaluasi menggunakan matriks konfusi dan ROC curve untuk menentukan performa model berdasarkan nilai AUC telah terbukti bahwa hasil uji algoritma Regresi Logistik memiliki akurasi yang lebih tinggi dibandingkan dengan hasil prediksi algoritma SVM. Tampilan tersebut menunjukkan bahwa performa model Regresi Logistik jauh lebih baik berdasarkan nilai Area Under the Curve (AUC) sebesar 84,3%.

E. Perbandingan Model

Berdasarkan hasil penelitian sebelumnya [14], pembanding yang dipakai adalah K-*Fold* 10 karena banyak penelitian menunjukkan bahwa ini merupakan pilihan terbaik untuk mencapai estimasi yang akurat. Hasil evaluasi yang didapatkan dibandingkan dengan tabel 5.

Tabel 4. Perbandingan Performa

K-Fold Cross Validation			
Metode	**Akurasi**	**Presisi**	**ROC**
SVM	91,57%	61.20%	84.1%
LR	91,66%	54.31%	84.3%

Hasil penelitian menunjukkan bahwa nilai presisi lebih rendah dibandingkan dengan akurasi saat menggunakan *K-Fold Cross Validation*. Walaupun hasil perulangan dari masing-masing percobaan tidak selalu tepat atau kurang presisi, kedua model algoritma klasifikasi yaitu SVM dan *Logistic Regression* terbukti memberikan hasil yang sangat akurat. Tabel 5 memperlihatkan bahwa *logistic regression* lebih unggul dibandingkan SVM berdasarkan evaluasi *K-Fold Cross Validation*. Nilai akurasi yang didapatkan oleh

logistic regression sebesar 91.66% dengan presisi 54.31% serta ROC 84.3%. Sedangkan, SVM memperoleh akurasi 91.57% , presisi 61.20% dan Evaluasi ROC 84.1%.

IV. PENUTUP

A. Kesimpulan

Hasil penelitian menunjukkan bahwa metode membandingkan menggunakan *K-Fold Cross Validation* memberikan nilai akurasi tertinggi. *Logistic Regression* memperoleh hasil lebih baik dibandingkan dengan *Support Vector Machine* (SVM), dengan nilai akurasi sebesar 91.76%, presisi sebesar 54.82% dan ROC 84.3%. SVM memperoleh akurasi 91.57% , presisi 61.20% dan Evaluasi ROC 84.1%.

B. Saran

Pada penelitian ini, hanya ada perbandingan antara dua algoritma klasifikasi yaitu *Support Vector Machine* (SVM) dan *Logistic Regression* (LR). Pada penelitian yang akan datang, mungkin akan ada penambahan algoritma klasifikasi lain yang akan memperbaiki performa dengan dataset penyakit jantung.

PENGAKUAN

Publikasi atau naskah ilmiah ini merupakan salah satu komponen kajian milik Chepy Bagustian Sonjaya dengan judul *The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease* yang dibimbing oleh Ibu Anis Fitri Nur Masruriyah, M. Kom dan Ibu Dwi Sulistya Kusumaningrum, M. Pd.

DAFTAR PUSTAKA

[1] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[2] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[3] "Cardiovascular diseases (CVDs)." https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Oct. 02, 2022).

[4] "Kementerian Kesehatan Republik Indonesia," 2021. https://www.kemkes.go.id/article/view/21093000002/penyakit-jantung-koroner-didominasi-masyarakat-kota.html (accessed Aug. 04, 2022).

[5] D. P. Utomo, P. Sirait, and R. Yunis, "Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0," *J. Media Inform. Budidarma*, vol. 4, no. 4, pp. 994–1006, 2020, doi: 10.30865/mib.v4i4.2355.

[6] S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.005.

[7] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[8] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, "Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031184.

[9] Trivusi, "Penjelasan Lengkap Algoritma Support Vector Machine (SVM)," 2022. https://www.trivusi.web.id/2022/04/algoritma-svm.html (accessed Aug. 06, 2022).

[10] M. Aminullah, *Perbandingan Performa Klasifikasi Machine Learning dengan Teknik Resampling pada Dataset Tidak Seimbang*. 2021. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/57648/1/MUHAMMAD AMINULLAH-FST.pdf

[11] F. Handayani, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, p. 329, 2021, doi: 10.26418/jp.v7i3.48053.

[12] W. Willy, D. P. Rini, and S. Samsuryadi, "Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1720, 2021, doi: 10.30865/mib.v5i4.3177.

[13] R. A. Nurdian, Mujib Ridwan, and Ahmad Yusuf, "Komparasi Metode SMOTE dan ADASYN dalam Meningkatkan Performa Klasifikasi Herregistrasi Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 1, 2022, doi: 10.28932/jutisi.v8i1.4004.

[14] U. Amelia *et al.*, "IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK PREDIKSI PENYAKIT STROKE DENGAN ATRIBUT BERPENGARUH," vol. III, pp. 254–259, 2022.