JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM)

Siti Novianti Nuraini Arif, Amril Mutoi Siregar*, Sutan Faisal, Ayu Ratna Juwita

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
Email: [1]if20.sitinoviantinurainiarif@mhs.ubpkarawang.ac.id, [2,*]amrilmutoi@ubpkarawang.ac.id,
[3]sutanfaisal@ubpkarawang.ac.id, [4]ayurj@ubpkarawang.ac.id
Email Penulis Korespondensi: amrilmutoi@ubpkarawang.ac.id

Abstrak−Penyakit kardiovaskular (PKV) istilah umum untuk gangguan yang berhubungan dengan jantung, arteri koroner dan pembuluh darah. Paling umum penyakit ini disebabkan penyumbatan pembuluh darah, baik karena penumpukan lemak atau perdarahan internal. Menurut WHO, setiap tahun angka kematian mencapai 32% disebabkan oleh penyakit kardiovaskular, yaitu sekitar 17,9 juta orang setiap tahun. Banyaknya faktor penyebab PKV , menyulitkan dokter untuk mendiagnosis pasien berpeluang rendah atau lebih tinggi terkena serangan jantung. Diperlukan model machine learning untuk pengenalan sejak dini gejala serangan penyakit jantung. Model supervised learning seperti KNN dan SVM dilakukan pada penelitian sebelumnya, tanpa penggunaan seleksi fitur dengan dataset yang didapat dari Kaggle. PCA diterapkan guna mengurangi dimensi data menjadi variabel yang lebih kecil. Dengan digunakan evaluasi confusion matrix dan kurva ROC, didapatkan peningkatan hasil akurasi dari penelitian sebelumnya model KNN sebesar 83,6% menjadi 90,16%. Model SVM pun mengalami peningkatan akurasi dari sebelumnya 85,7% menjadi 86,88%. Digunakan seleksi fitur PCA, membuktikan adanya peningkatan akurasi pada penelitian tersebut. Model KNN dengan tingkat akurasi lebih tinggi yaitu 90,16% lebih baik untuk klasifikasi seseorang termasuk normal atau dapat terdiagnosis serangan jantung.

Kata Kunci: Machine learning; KNN; SVM; Klasifikasi; Serangan jantung

Abstract−Cardiovascular disease (CVD) is a general term for disorders related to the heart, coronary arteries, and blood vessels. These diseases are most commonly caused by blocked blood vessels, either due to fat buildup or internal bleeding. According to the WHO, each year, cardiovascular diseases account for 32% of all deaths, which translates to about 17.9 million people annually. The numerous factors causing CVD make it challenging for doctors to diagnose patients who are at low or higher risk of heart attacks. A machine learning model is needed for the early recognition of heart attack symptoms. Supervised learning models such as KNN and SVM were used in previous studies without feature selection, with datasets obtained from Kaggle. PCA was applied to reduce data dimensions to smaller variables. With the use of confusion matrix and ROC curve evaluations, the accuracy results of the previous KNN model improved from 83.6% to 90.16%. The SVM model also saw an accuracy increase from 85.7% to 86.88%. The use of PCA feature selection demonstrated an improvement in accuracy in the study. The KNN model, with a higher accuracy rate of 90.16%, is better for classifying individuals as normal or diagnosed with a heart attack.

Keywords: Machine learning; KNN; SVM; Classification; Heart Attack

1. PENDAHULUAN

Penyakit kardiovaskular (PKV) merupakan istilah umum untuk gangguan yang berhubungan dengan jantung, arteri koroner dan pembuluh darah (vena, arteri dan kapiler). Dalam istilah luas mencakup penyakit seperti penyakit jantung koroner, thrombosis vena dalam dan penyakit serebrovaskular yang menyebabkan serangan jantung atau stroke. Paling umum penyakit ini disebabkan penyumbatan pembuluh darah, baik karena penumpukan lemak atau perdarahan internal. Menurut World Health Organization (WHO), setiap tahun angka kematian mencapai 32% disebabkan oleh penyakit kardiovaskular, yaitu sekitar 17,9 juta orang setiap tahun [1].

Setiap tahun, di Indonesia penyakit kardiovaskular menyebabkan kematian sebanyak 651.481 orang. Dari jumlah tersebut, 331.349 orang meninggal akibat stroke, 54.343 orang karena penyakit jantung koroner, 50.60 orang akibat penyakit jantung hipertensi, dan sisanya disebabkan oleh berbagai penyakit kardiovaskular lainnya [2]. Dalam beberapa tahun terakhir penyakit kardiovaskular menjadi penyebab utama kematian di seluruh dunia baik di negara-negara berkembang maupun di negara-negara miskin [3] Beberapa aspek yang digunakan untuk mendiagnosis penyakit jantung berupa kadar kolesterol, tekanan darah tinggi, menurunnya olahraga, dan obesitas. Faktor-faktor tersebut kemungkinan dapat berubah ataupun tidak berubah. Ada beberapa faktor yang tidak dapat diubah seperti jenis kelamin, riwayat keluarga, dan usia. Sedangkan faktor yang dapat berubah termasuk perokok aktif maupun pasif, tekanan darah tinggi, kolesterol tinggi dan kurangnya aktivitas fisik. Selain itu analisis dokter menjadi penyebab sulitnya untuk mendiagnosis pasien tersebut berpeluang lebih rendah atau lebih tinggi terkena serangan jantung [4]. Junk food atau makanan olahan dapat menjadi faktor pemicu lainnya penyebab penyakit serangan jantung karena kandungan lemak jahat di dalamnya [5].

Pengenalan sejak dini gejala serangan jantung memungkinkan pasien untuk mengelola beberapa risiko melalui perubahan diet atau obat, guna menghentikan perkembangan penyakit ke tahap yang lebih parah bahkan bisa kematian. Prinsip di balik machine learning guna melatih komputer untuk memahami informasi menggunakan data tanpa pemrograman yang ekstensif. Machine learning suatu teknik untuk membangun model analitik otomatis yang digunakan sistem untuk mempelajari data, mendeteksi tren dan mengurangi keterlibatan manusia dalam

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

proses pengambilan keputusan [6]. Pada penelitian sebelumnya digunakan teknik jaringan saraf tiruan (ANN), random forest, K-Nearest Neighbor (KNN), dan support vector machine. Disebutkan bahwa ANN menghasilkan akurasi tertinggi untuk prediksi penyakit jantung dibandingkan dengan algoritma kalsifikasi sebelumnya. Fitur invarian skala, Principle Component Analysis-K-Nearest Neighbor (PCA-KNN) digunakan dalam gambar medis untuk penskalaan guna mengembangkan pendekatan baru yang mencapai akurasi 83,6% [7]. Chepy et. al dalam penelitiannya menerapkan partisi k-fold cross validation untuk pengujian guna mengurangi bias yang terdapat pada data random [8].

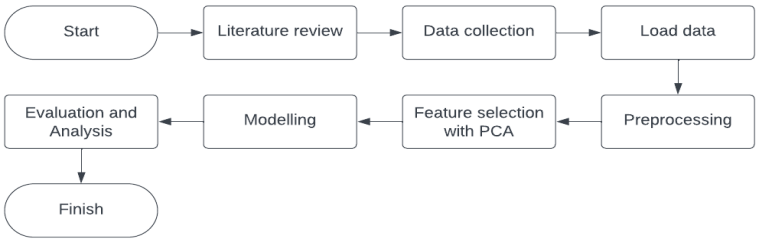
Dataset pada penelitian sebelumnya dikumpulkan dari Kaggle dengan 14 atribut, dataset tersebut berisi 1025 catatan pasien termasuk 713 laki-laki dan 312 perempuan dari berbagai usia. Di mana 499 (48,68%) pasien normal dan 526 (51,32%) pasien menderita penyakit jantung. Di antara pasien yang menderita penyakit jantung, 300 (57,03%) pasien adalah laki-laki dan 226 (42,97%) pasien adalah perempuan [9]. Pada penelitian lainnya menggunakan data yang berisi 13 atribut dan 299 orang menderita penyakit gagal jantung dengan model KNN dan didapatkan akurasi sebesar 94,92% [10]. Algoritma Sine Cosine Weighted K-Nearest Neighbour (SCA_WKNN) berbasis machine learning diusulkan untuk prediksi penyakit jantung dari data yang disimpan di blockchain. Karena data yang disimpan di blockchain tidak dapat diubah, ini berfungsi sebagai sumber otentik untuk data pembelajaran dan juga sebagai lingkungan penyimpanan yang aman untuk informasi pasien [11]. Pada penelitian lainnya digunakan uji Chi-Square guna menyortir fitur berdasarkan kelas dan menyaring fitur-fitur teratas yang bergantung pada label kelas. ChisqSelector (CHI) dari Apache Spark MLlib digunakan untuk seleksi fitur dalam konstruksi model [12]. Pada penelitian sebelumnya diterapkan juga teknik oversampling SMOTE + Random forest, yang dapat meningkatkan akurasi hingga 94,54% dan 98,4% kurva ROC [13]

Berbagai model diterapkan untuk prediksi risiko serangan jantung, probabilitas risiko serangan jantung ditampilkan melalui sebuah situs web. Model lainnya yang digunakan yaitu Support vector machine (SVM) mencapai nilai akurasi sebesar 85,7%. Nilai akurasi tersebut didapatkan dari beberapa atribut yang dianalisis berupa kolesterol, tekanan darah dan gula darah [14]. Dari beberapa penelitian terlihat adanya perbedaan yang signifikan antara model algoritma yang digunakan. Penelitian ini bertujuan untuk menguji teknik statistik dan mengombinasikannya dengan algoritma machine learning SVM dan KNN. Teknik statistik yang akan diterapkan adalah Principal Component Analysis (PCA). Tujuan utama dari penelitian ini adalah mengevaluasi keefektifan teknik statistik dalam memprediksi serangan jantung. Penelitian ini secara khusus akan mengukur peningkatan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score. Diharapkan bahwa penelitian ini dapat mengimplementasikan teknik statistik untuk meningkatkan performa model algoritma dalam memprediksi kasus serangan jantung dengan akurasi yang lebih tinggi, dibandingkan dengan model yang tidak menggunakan teknik statistik. Dengan demikian, diharapkan penelitian ini mampu memberikan pemahaman kepada tenaga kesehatan juga masyarakat dalam upaya pengendalian angka penyakit serangan jantung.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dimulai dengan pemeriksaan ilmiah. Tujuan pemeriksaan tersebut untuk menemukan dasar teoritis yang akan digunakan dalam penelitian dan untuk menemukan literatur ilmiah yang relevan untuk mendukung penelitian. Untuk menyimpulkan penelitian ini, tahapan atau prosedur penelitian akan dilaksanakan seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian

Pada gambar 1. penelitian dimulai dengan melakukan tinjauan literatur untuk memahami klasifikasi penyakit serangan jantung. Selanjutnya, data penelitian diambil dari dataset serangan jantung yang tersedia di platform Kaggle, yang dikenal sebagai sumber informasi dan kumpulan data terkemuka. Tahap berikutnya melibatkan preprocessing data untuk menghindari duplikasi, missing value, outlier, dan melakukan standardisasi. Setelah data berkualitas dan siap digunakan, data tersebut dibagi (splitting data) dengan rasio 80:20 untuk data latih dan data uji. Langkah selanjutnya adalah membuat model menggunakan algoritma machine learning KNN dan SVM untuk mengklasifikasi serangan jantung. Sebelumnya dilakukan seleksi fitur menggunakan PCA,

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

sebagai Langkah preprocessing untuk meningkatkan kinerja algoritma machine learning dengan mengurangi overfitting dan mengurangi waktu komputasi. Di proses akhir dilakukan evaluasi menggunakan confusion matrix dan Receiver Operating Characteristic (ROC) sebagai tahap penutup penelitian.

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berasal dari Kaggle yaitu "Heart Attack Analysis & Prediction", berisi 303 baris 14 kolom yang ditemukan oleh Rashik Rahman seorang dosen di Universitas Asia Pasifik, Dhaka, Dhaka Vision, Bangladesh. Dataset tersebut dapat diakses melalui URL : https://www.Kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset. Informasi lebih detail mengenai dataset dapat dilihat pada Tabel 1.

Tabel 1. Informasi dataset

Nama fitur	Deskripsi
Age	Usia pasien
Sex	Jenis kelamin (1 = laki-laki, 0 = perempuan)
CP	Tipe nyeri dada (0-3)
Trestbps	Tekanan darah saat istirahat
Chol	Kolesterol serum (mg/dl)
Fbs	Gula darah puasa > 120 mg/dl (1 = benar, 0 = salah)
Restecg	Hasil elektrokardiografi istirahat (0-2)
Thalach	Detak jantung maksimal
Exang	Angina yang diinduksi oleh olahraga (1 = ya, 0 = tidak)
Oldpeak	Depresi ST yang diinduksi oleh olahraga relative terhadap istirahat
Slope	Kemiringan segmen STR puncak Latihan
Ca	Jumlah pembuluh utama (0-4) yang diwarnai oleh fluoroskopi
Thal	3 = normal; 6 = cacat tetap; 7 = cacat reversible
Target	1= penyakit jantung, 0 = normal

2.3 Data Preprocessing

Data preprocessing merupakan tahapan awal sebelum data dibuat menjadi sebuah model, agar data yang diolah terjaga kualitas dan konsistensinya. Langkah-langkah data preprocessing yaitu data cleaning, standardisasi dan seleksi fitur. Data cleaning merupakan proses mengoreksi atau menemukan adanya kesalahan atau anomali pada kumpulan data [15]. Langkah selanjutnya standardisasi merupakan transformasi data sehingga memiliki mean 0 dan standar deviasi 1 menggunakan Z-Score [16]. Seleksi fitur menggunakan PCA agar mengurangi noise dengan menghilangkan fitur-fitur yang tidak relevan agar meningkatkan akurasi prediksi [17].

2.4 Seleksi Fitur menggunakan PCA

Principal Component Analysis (PCA) teknik statistik yang digunakan untuk mengurangi dimensi data dengan mengubah data asli menjadi sejumlah kecil variabel atau komponen baru. PCA sangat berguna dalam analisis data yang memiliki banyak fitur. Langkah utama dalam PCA menghitung eigenvalue guna menunjukkan besarnya varian yang ditangkap oleh masing-masing komponen utama [18].

2.5 Algoritma Klasifikasi

a. K-Nearest Neighbors (KNN) adalah algoritma supervised learning yang digunakan untuk menyelesaikan masalah klasifikasi maupun regresi. Dalam konteks klasifikasi, algoritma KNN mengidentifikasi tetangga terdekat dari sebuah titik data baru seiring dengan variasi nilai K hasil prediksi akan berubah [19]. Algoritma KNN cocok untuk masalah klasifikasi dengan ukuran sampel yang besar [20]. Jarak Euclidean digunakan untuk menghitung dekat atau jauhnya jarak antar titik pada kelas K [21] . Formula untuk mencari jarak antara 2 titik dalam ruang dua dimensi [22]. Rumus untuk menghitung jarak Euclidean ada pada persamaan (1).

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

b. Support vector machine (SVM) algoritma yang menghasilkan model klasifikasi ke contoh baru dalam salah satu dari beberapa kategori. SVM mengawasi algoritma machine learning yang digunakan untuk masalah klasifikasi dan regresi [23]. Fungsi utama dari pengklasifikasi SVM adalah untuk menyelesaikan pemilihan subset fitur dengan penyesuaian parameter. Dalam penelitian ini algoritma diusulkan untuk mengoptimalkan dua parameter SVM, yaitu bobot C dan fungsi kernel [24]. Fungsi kernel mengimplementasikan model dalam ruang dimensi yang lebih tinggi menggunakan Gaussian radial basis function (RBF) [25]. Rumus untuk menghitung kernel RBF ada pada persamaan (2).

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2a^2}\right) \quad (2)$$

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

2.6 Evaluasi dan Analisis

Pada penelitian ini dilakukan confusion matrix dan Receiver Operating Characteristic (ROC) Curve guna mengevaluasi dan menganalisa kinerja model klasifikasi yang diterapkan. Confusion matrix metode evaluasi yang melibatkan perhitungan berbagai metrik termasuk, akurasi, presisi, recall dan F1-score [26].

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

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

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

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

Arti dari persamaan (3), (4), (5) dan (6) adalah sebagai berikut:

TP = True Positif (Benar positif : diklasifikasikan dengan benar sebagai penderita penyakit serangan jantung). TN = True Negatif (Benar negatif : diklasifikasikan dengan benar sebagai tidak menderita penyakit serangan jantung). FP = False Positif (Salah positif : diklasifikasikan sebagai penderita penyakit serangan jantung, padahal sebenarnya tidak menderita penyakit serangan jantung; kesalahan tipe I). FN = False negatif (Salah negatif :diklasifikasikan sebagai tidak menderita penyakit serangan jantung, padahal sebenarnya menderita penyakit serangan jantung; kesalahan tipe II).

ROC merupakan grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi. Grafik ROC menampilkan kurva yang menunjukkan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai threshold klasifikasi.

a. True Positiver Rate (TPR) atau Sensitivity adalah rasio positif yang benar teridentifikasi dengan benar dari semua kasus positif sebenarnya, seperti pada contoh persamaan (7).

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

b. False Positive Rate (FPR) rasio negatif yang salah diidentifikasi sebagai positif dari semua kasus negatif sebenarnya, seperti pada contoh persamaan (8).

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \quad (8)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dataset “Heart Attack Analysis & Prediction” terdiri dari 13 tipe data integer dan 1 tipe data float. Pada gambar 2. menampilkan 10 baris pertama pada dataset.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Gambar 2. Info dataset

Pada gambar 3. dijelaskan tidak ada missing value pada dataset tersebut sehingga dapat dipastikan setiap kolom terdapat nilai, data clean dan siap untuk dilakukan analisis data.

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

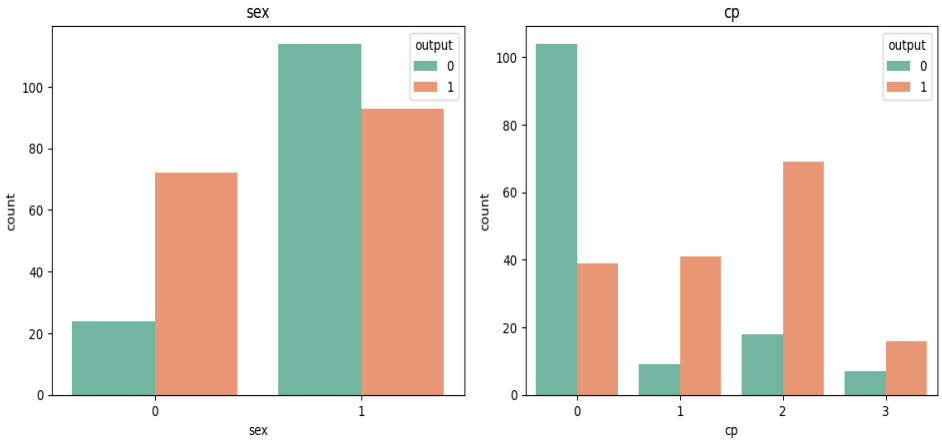
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

	Total Missing Values
age	0
sex	0
cp	0
trtbps	0
chol	0
fbs	0
restecg	0
thalachh	0
exng	0
oldpeak	0
slp	0
caa	0
thall	0
output	0

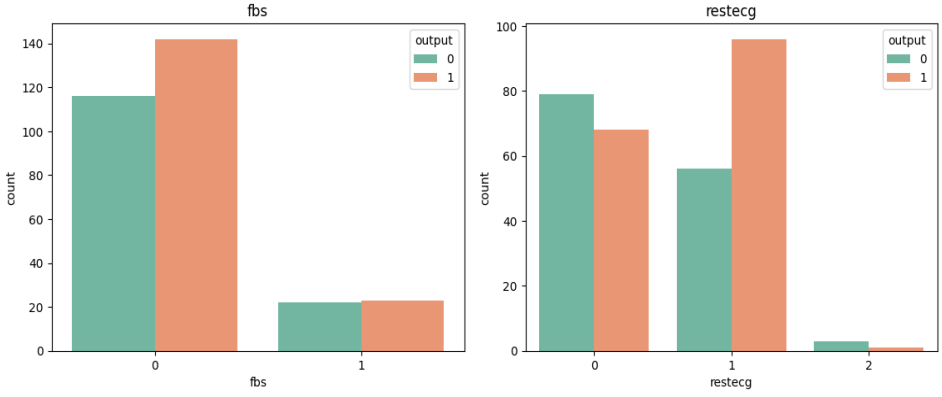
Gambar 3. Missing value dataset

3.2 Data Preprocessing

Setelah didapatkan info detail dari dataset, fitur dengan tipe kategori di visualisasikan agar terlihat perbedaan dari masing-masing variabel terhadap output seperti pada gambar 4. untuk fitur sex dan chest pain. Gambar 5. visualisasi untuk data fasting blood sugar (fbs) dan resting electrocardiographic results (rest_ecg). Gambar 6. visualisasi data exercise induced angina (exang) dan slope (slp). Gambar 7. visualisasi data number of major vessels (caa) dan thal rate (thall). Gambar 8 memvisualisasikan output.



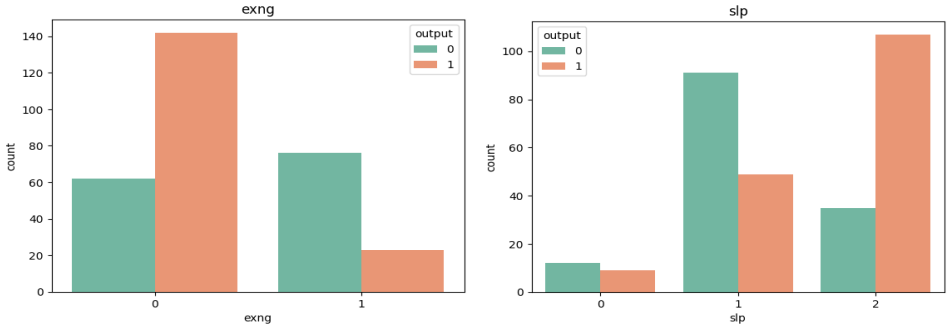
Gambar 4. Visualisasi data sex dan cp



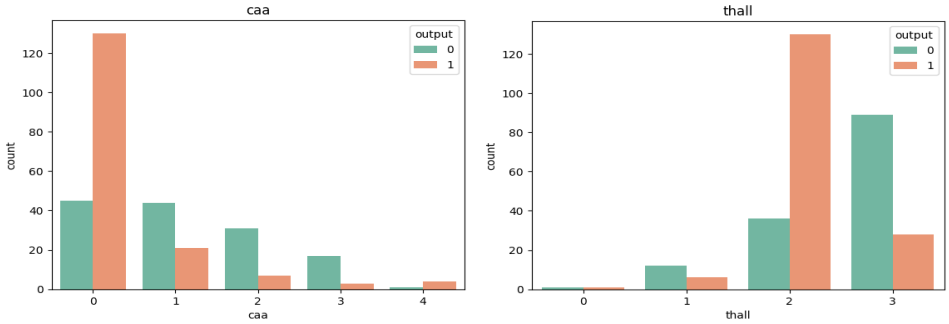
Gambar 5. Visualisasi data fbs dan restecg

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

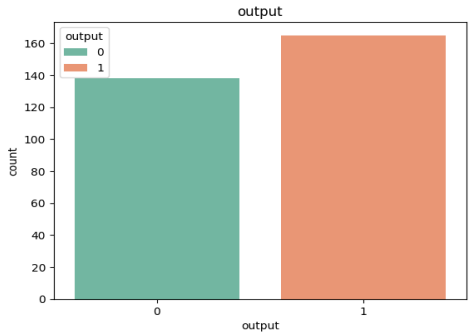
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844



Gambar 6. Visualisasi data exng dan slp

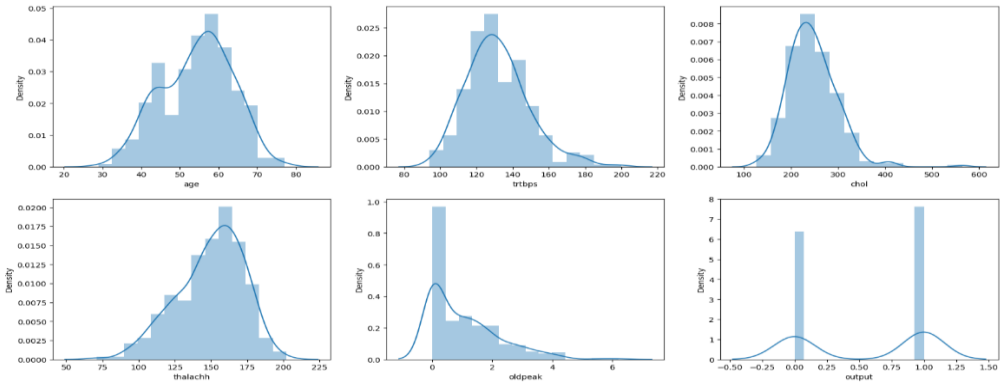


Gambar 7. Visualisasi data caa dan thall



Gambar 8. Visualisasi data output

Selanjutnya untuk memaksimalkan kinerja machine learning terhadap model, dilakukan normalisasi agar penskalaan fitur berada dalam rentang tertentu 0 dan 1 seperti pada gambar 9. Normalisasi data menggunakan Z-Score standardization, mengubah nilai sehingga distribusi dengan rata-rata 0 dan standar deviasi 1.



Gambar 9. Normalisasi data numerik

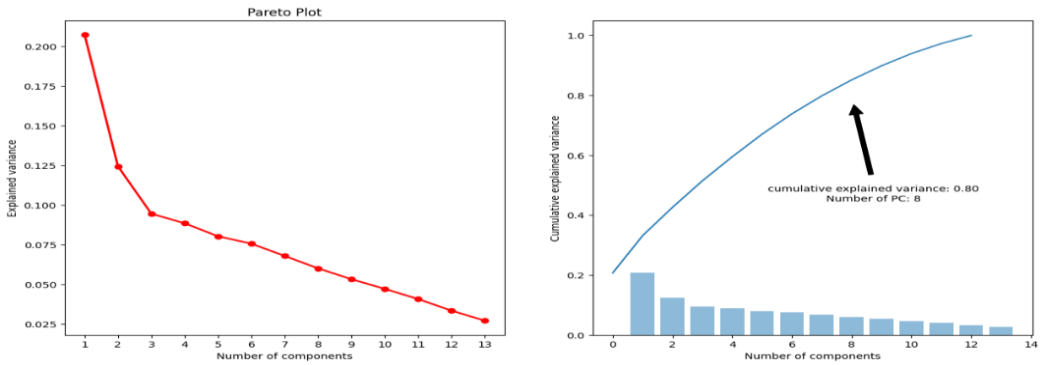
3.3 Seleksi Fitur menggunakan PCA

Dataset yang digunakan dalam penelitian ini memiliki dimensi yang tinggi, dengan 14 fitur. Jumlah fitur yang besar menghambat pencapaian hasil terbaik dan menyebabkan overfitting. Oleh karena itu, PCA diterapkan pada

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

dataset ini untuk mengubah 14 fitur menjadi 8 fitur, sehingga meningkatkan kinerja hasilnya. Plot pareto digunakan dalam penelitian ini, seperti pada gambar 10. untuk memeriksa eigenvalue dan menentukan jumlah optimal komponen utama. Diagram pareto dari kurva PCA, sumbu x menampilkan komponen utama dalam urutan menurun berdasarkan kontribusinya. Sumbu y menampilkan varian yang dijelaskan oleh setiap komponen. Plot pareto menunjukkan bahwa jumlah optimal PC adalah 8, dengan kumulatif varian sebesar 0.80. Konsep varian yang dijelaskan kumulatif merupakan aspek mendasar dari PCA, yang merupakan teknik pengurangan dimensi yang digunakan dalam analisis multivariat.



Gambar 10. Pareto plot eugenvalue dalam PCA

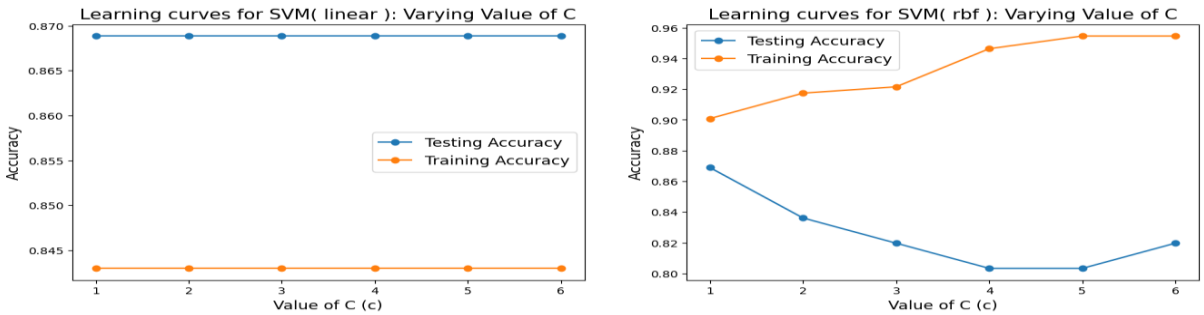
3.4 Penerapan Algoritma Klasifikasi

3.3.1 K-Nearest Neighbors (KNN)

Ada beberapa perpustakaan yang perlu diimpor guna mengimplementasikan KNN, seperti pembagian data, pengaturan parameter dan evaluasi model. Kumpulan data yang sudah ada, dibagi menjadi data uji dan data latih dengan rasio 80:20. Setelah diterapkan PCA untuk reduksi dimensi sehingga didapatkan data latih dan data uji yang sudah ditransformasi. Akurasi model KNN dapat ditingkatkan seiring bertambahnya jumlah tetangga terdekat yang ditentukan oleh nilai K. Dalam penelitian ini, kami menggunakan nilai K sebesar 7, yang dalam persamaan (1) adalah jarak Euclidean (d) antara dua titik x dan ya. Jarak dihitung menggunakan rumus di atas untuk setiap tetangga, dan kelas mayoritas di antara tetangga ini menentukan klasifikasi instance baru. Nilai K dalam tahap ini guna melatih model KNN agar mendapatkan hasil evaluasi dari data uji.

3.3.2 Support vector machine (SVM)

Model pendekatan supervised learning SVM digunakan dalam klasifikasi pola untuk meningkatkan keamanan dan kualitas layanan. Fungsi kernel linier dan RBF banyak digunakan karena berasal langsung dari produk dalam fitur asli. Mereka sangat bermanfaat dalam skenario di mana mereka menawarkan keuntungan seperti parameter yang lebih sedikit dan pemrosesan yang cepat. Dalam kasus seperti ini, memilih fungsi parameter alternatif menjadi penting untuk kesesuaian yang lebih baik, sehingga kami memilih parameter C. Kernel RBF dengan nilai c =1 menunjukkan kinerja yang paling optimal dalam hal akurasi pelatihan dan pengujian, sepeti pada gambar 11. Parameter C dalam fungsi SVM sebagai parameter penalty untuk kesalahan klasifikasi yang ada dalam data pelatihan. Ini mengatur sejauh mana penalty diberikan kepada titik data yang berada di sisi yang salah dari hyperlane pemisah. Nilai C yang lebih tinggi menghasilkan penalty yang lebih besar, sehingga membuat model lebih ketat dalam menangani kesalahan klasifikasi selama pelatihan. Dengan memilih nilai C yang tepat, kita dapat mencapai keseimbangan optimal antara menyesuaikan model dengan data pelatihan dan mencegah overfitting.



Gambar 11. Kernel SVM linear dan RBF untuk nilai c terbaik

3.5 Evaluasi dan Analisis

Setelah melakukan preprocessing data, kinerja klasifikasi ditampilkan secara visual menggunakan berbagai metrik. Preprocessing dataset melibatkan pengecekan missing value dan normalisasi data menggunakan Z-Score,

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

mengubah nilai sehingga distribusi dengan rata-rata 0 dan standar deviasi 1. Selain itu, penskalaan data dengan pengurangan dimensi menggunakan Principal Component Analysis (PCA) diterapkan pada semua algoritma machine learning yang digunakan dalam penelitian ini. Evaluasi kinerja model menggunakan confusion matrix dan Receiver Operating Characteristic (ROC).

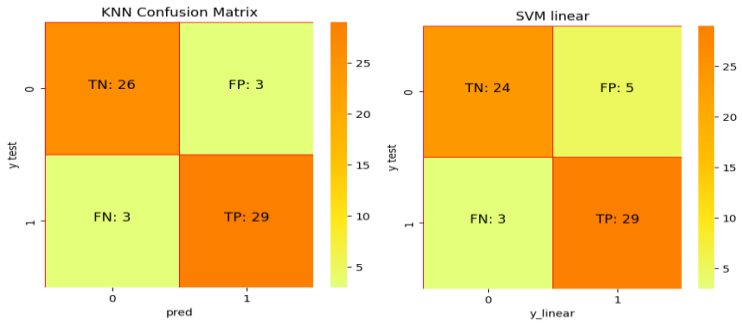
3.4.1 Kinerja dengan Confusion Matrix

Penelitian kali ini membandingkan 2 model algoritma yaitu KNN dan SVM untuk mengetahui performa dari masing-masing model menggunakan metrik seperti akurasi, presisi, recall dan F1-score. Rumus untuk mengetahui masing-masing performa dari metrik tersebut ada pada persamaan (3), (4), (5) dan (6).

Akurasi mengacu pada jumlah titik data yang diprediksi dengan benar oleh model machine learning dari total titik data, dan dapat dihitung pada persamaan (3). Presisi adalah persentase elemen relevan yang diprediksi dengan benar oleh model dan dapat dihitung menggunakan persamaan (4). Sementara Recall adalah persentase elemen relevan yang diklasifikasikan dengan benar oleh model dari semua elemen relevan dan dapat dihitung menggunakan persamaan (5). F1-Score adalah rata-rata harmonis dari presisi dan recall, ukuran seimbang yang berguna ketika distribusi kelas tidak seimbang seperti pada persamaan (6). Hasil dari model KNN dan SVM ditunjukkan pada tabel 2, dan confusion matrix dengan label x mewakili data uji dan label y mewakili prediksi model ditunjukkan dalam gambar 12.

Tabel 2. Performa algoritma berdasarkan confusion matrix

Model dan Akurasi	Normal (0)			Penyakit Jantung (1)		
	Presisi	Recall	F1-Score	Presisi	Recall	F1-Score
KNN (90,16%)	90%	90%	90%	91%	91%	91%
SVM (86,88%)	89%	83%	86%	85%	91%	88%

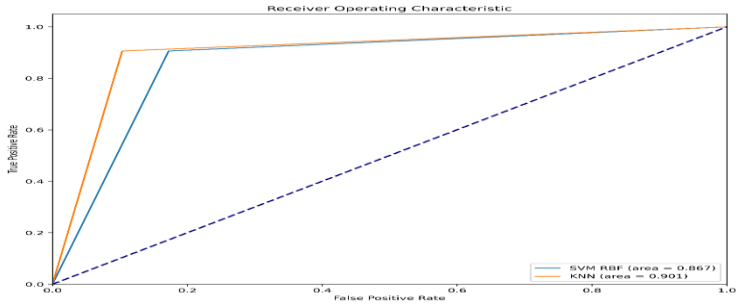


Gambar 12. Confusion matrix model

3.4.2 Kinerja dengan ROC Curve

Kami telah mengimplementasikan Receiver Operating Characteristic (ROC) yang berfungsi sebagai alat konvensional untuk pemulihan dan penilaian model dalam masalah yang melibatkan klasifikasi dua kelas. Kurva ROC dapat dihitung dengan menggunakan True Positive Rate (TPR) dan False Positive Rate (FPR) yang diperoleh dari perhitungan confusion matrix seperti pada persamaan (3) dan (4). Terlihat jelas bahwa terdapat konvergensi di antara dua model klasifikasi machine learning. Akurasi tertinggi dicapai oleh KNN dengan akurasi 90,16%. Pada gambar 13. terlihat kurva ROC untuk kedua model berada di atas garis diagonal (garis dari (0,0) ke (1,1), yang menunjukkan bahwa kedua model lebih baik daripada tebakan acak dalam membedakan antara kelas positif dan negatif. Kurva ROC KNN (diwakili oleh garis orange) berada di atas kurva ROC SVM RBF (diwakili oleh garis biru) untuk sebagian besar nilai FPR menunjukkan bahwa KNN memiliki performa klasifikasi yang lebih baik daripada SVMRBF dalam konteks ini.

Nilai AUC untuk KNN 0.901, lebih baik dibandingkan dengan nilai AUC SVM RBF 0.867 yang artinya AUC lebih mendekati 1 menunjukkan model tersebut memiliki performa yang lebih baik, seperti pada gambar 13.



Gambar 13. ROC model KNN dan SVM

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

4. KESIMPULAN

Penelitian ini mengeksplorasi klasifikasi penyakit serangan jantung menggunakan dua algoritma supervised learning, yaitu K-Nearest Neighbors (KNN) dan Support vector machine (SVM) dengan kernel RBF. Setelah melakukan preprocessing data yang mencakup pengecekan missing value, normalisasi menggunakan Z-Score, dan pengurangan dimensi menggunakan Principal Component Analysis (PCA), kinerja masing-masing model dievaluasi menggunakan metrik akurasi, presisi, recall, F1-Score serta kurva ROC. Hasil evaluasi menunjukkan bahwa model KNN memiliki akurasi tertinggi sebesar 90,16% dengan nilai presisi dan recall yang juga lebih tinggi dibandingkan dengan model SVM RBF. KNN menunjukkan performa lebih baik dalam klasifikasi penyakit serangan jantung, dengan nilai F1-Score yang lebih konsisten. Selain itu kurva ROC KNN yang berada di atas kurva ROC SVM RBF untuk sebagian besar nilai FPR menunjukkan bahwa KNN memiliki kemampuan diskriminasi yang lebih baik antara kelas positif dan negatif. Nilai AUC untuk KNN adalah 0.901, lebih tinggi dibandingkan dengan AUC SVM RBF yang sebesar 0.867, mengindikasikan bahwa KNN lebih andal dalam membedakan antara pasien yang benar-benar memiliki penyakit serangan jantung dan yang tidak. Dengan demikian, dalam konteks penelitian ini, model KNN lebih efektif dan akurat dalam kalsifikasi penyakit serangan jantung dibandingkan dengan model SVM RBF. Ini menunjukkan pentingnya memilih algoritma yang tepat dan melakukan preprocessing data yang optimal untuk meningkatkan kinerja model machine learning dalam aplikasi medis.

REFERENCES

[1] M. Rizwan, S. Arshad, H. Aijaz, R. A. Khan, dan M. Z. U. Haque, "Heart Attack Prediction using Machine Learning Approach," dalam 2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), IEEE, Nov 2022, hlm. 1–8. doi: 10.1109/INTELLECT55495.2022.9969395.

[2] Rokom, "Cegah Penyakit Jantung dengan Menerapkan Perilaku CERDIK dan PATUH." Diakses: 22 Juni 2024. [Daring]. Tersedia pada: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230925/4943963/cegah-penyakit-jantung-dengan-menerapkan-perilaku-cerdik-dan-patuh/#:~:text=Kematian%20di%20Indonesia%20akibat%20penyakit,Matrics%20and%20Evaluation%2C%202019).

[3] J. N, D. P, M. E, R. Santhosh, R. Reshma, dan D. Selvapandian, "Heart Attack Prediction using Machine Learning," dalam 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, Sep 2022, hlm. 854–860. doi: 10.1109/ICIRCA54612.2022.9985736.

[4] A. Jain, A. Chandra Sekhara Rao, P. Kumar Jain, dan Y.-C. Hu, "Optimized levy flight model for heart disease prediction using CNN framework in big data application," Expert Syst Appl, vol. 223, hlm. 119859, Agu 2023, doi: 10.1016/j.eswa.2023.119859.

[5] D. Ismafillah, T. Rohana, dan Y. Cahyana, "Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke," Jurnal Riset Komputer), vol. 10, no. 1, hlm. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.

[6] K. Tn, S. C. P, M. S, A. Kodipalli, T. Rao, dan S. Kamal, "Prediction of Early Heart Attack Possibility Using Machine Learning," dalam 2023 2nd International Conference for Innovation in Technology (INOCON), IEEE, Mar 2023, hlm. 1–5. doi: 10.1109/INOCON57975.2023.10100993.

[7] N. Nandal, L. Goel, dan R. TANWAR, "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis," F1000Res, vol. 11, hlm. 1126, Sep 2022, doi: 10.12688/f1000research.123776.1.

[8] C. B. Sonjaya, A. Fitri, N. Masruriyah, D. S. Kusumaningrum, dan A. R. Pratama, "The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease," INTERNAL (Information System Journal, vol. 5, no. 2, hlm. 166–175, 2022, doi: 10.32627.

[9] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, dan M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput Biol Med, vol. 136, hlm. 104672, Sep 2021, doi: 10.1016/j.compbiomed.2021.104672.

[10] D. A. Muhammad, R. Amril, dan M. Siregar, "Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung," vol. III, no. 1, 2022, [Daring]. Tersedia pada: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

[11] H. Hasanova, M. Tufail, U.-J. Baek, J.-T. Park, dan M.-S. Kim, "A novel blockchain-enabled heart disease prediction mechanism using machine learning," Computers and Electrical Engineering, vol. 101, hlm. 108086, Jul 2022, doi: 10.1016/j.compeleceng.2022.108086.

[12] A. K. Gárate-Escamila, A. Hajjam El Hassani, dan E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Inform Med Unlocked, vol. 19, hlm. 100330, 2020, doi: 10.1016/j.imu.2020.100330.

[13] A. F. N. Masruriyah, H. Y. Novita, C. E. Sukmawati, S. N. N. Arif, dan A. R. Ramadhan, "Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling," MIND (Multimedia Artificial Intelligent Networking Database) Journal, vol. 8, no. 2, hlm. 242–253, Des 2023.

[14] A. A. Shanbhag, C. Shetty, A. Ananth, A. S. Shetty, K. Kavanashree Nayak, dan B. R. Rakshitha, "Heart Attack Probability Analysis Using Machine Learning," dalam 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), IEEE, Nov 2021, hlm. 301–306. doi: 10.1109/DISCOVER52564.2021.9663631.

[15] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, dan A. Raad, "Cardiovascular Events Prediction using Artificial Intelligence Models and Heart Rate Variability," Procedia Comput Sci, vol. 203, hlm. 231–238, 2022, doi: 10.1016/j.procs.2022.07.030.

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 8, Nomor 3, Juli 2024, Page 1617-1626
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v8i3.7844

[16] S. P. Barfungpa, H. K. Deva Sarma, dan L. Samantaray, "An intelligent heart disease prediction system using hybrid deep dense Aquila network," Biomed Signal Process Control, vol. 84, hlm. 104742, Jul 2023, doi: 10.1016/j.bspc.2023.104742.
[17] F. Ali dkk., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Information Fusion, vol. 63, hlm. 208–222, Nov 2020, doi: 10.1016/j.inffus.2020.06.008.
[18] Takio Kurita, "Principal Component Analysis (PCA)." Diakses: 20 Juni 2024. [Daring]. Tersedia pada: https://link.springer.com/referenceworkentry/10.1007/978-3-030-03243-2_649-1, 2020.
[19] R. R. Sanni dan H. S. Guruprasad, "Analysis of performance metrics of heart failured patients using Python and machine learning algorithms," Global Transitions Proceedings, vol. 2, no. 2, hlm. 233–237, Nov 2021, doi: 10.1016/j.gltp.2021.08.028.
[20] M. Wang, X. Yao, dan Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," IEEE Access, vol. 9, hlm. 25394–25404, 2021, doi: 10.1109/ACCESS.2021.3057693.
[21] S. Faisal, "Implementation of K-Nearest Neighbor Algorithm for Customer Satisfaction," Buana Information Tchnology and Computer Sciences (BIT and CS, vol. 1, no. 2, 2020.
[22] UNTUNG JAMARI, "PENJELASAN CARA KERJA ALGORITMA K-NEAREST NEIGHBOR (KNN)." Diakses: 19 Juni 2024. [Daring]. Tersedia pada: http://labdas.si.fti.unand.ac.id/2022/03/20/penjelasan-cara-kerja-algoritma-k-nearest-neighbor-knn/
[23] V. Chang, V. R. Bhavani, A. Q. Xu, dan M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthcare Analytics, vol. 2, hlm. 100016, Nov 2022, doi: 10.1016/j.health.2022.100016.
[24] S. P. Patro, G. S. Nayak, dan N. Padhy, "Heart disease prediction by using novel optimization algorithm: A supervised learning prospective," Inform Med Unlocked, vol. 26, hlm. 100696, 2021, doi: 10.1016/j.imu.2021.100696.
[25] E. P. P. Kendrew Huang, "Support Vector Machine Algorithm." Diakses: 19 Juni 2024. [Daring]. Tersedia pada: https://sis.binus.ac.id/2022/02/14/support-vector-machine-algorithm/
[26] E. R. Lidinillah, T. Rohana, dan A. R. Juwita, "Analisis sentimen twitter terhadap steam menggunakan algoritma logistic regression dan support vector machine," TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika, vol. 10, no. 2, hlm. 154–164, Jul 2023, doi: 10.37373/tekno.v10i2.440.

Submitted: **28/06/2024**; Accepted: **23/07/2024**; Published: **27/07/2024**