

Klasifikasi *Cardiovascular Diseases* Menggunakan Algoritma *K-Nearest Neighbors* (KNN)

Classification of Cardiovascular Diseases using K-Nearest Neighbors (KNN) Algorithm

Vera Artanti¹, Muhammad Faisal², Fachrul Kurniawan³

^{1,2,3} Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Universitas Islam Negeri
Maulana Malik Ibrahim, Kota Malang, Indonesia

e-mail: ¹200605110039@student.uin-malang.ac.id, ²mfaisal@ti.uin-malang.ac.id,

³fachrulk@ti.uin-malang.ac.id

Abstrak

Penyakit Kardiovaskular (*Cardiovascular Diseases*) adalah faktor utama kematian global, dengan jumlah korban mencapai 17,9 juta jiwa setiap tahun atau sekitar 32% dari total kematian global (*World Health Organization*, 2021). Faktor risiko penyakit kardiovaskular diantaranya faktor usia, semakin bertambahnya usia seseorang, maka semakin tinggi risiko terkena penyakit kardiovaskular. Faktor lain yaitu memiliki riwayat penyakit kardiovaskular dalam keluarga, diabetes, tekanan darah tinggi, obesitas (kegemukan), Pola hidup tidak sehat, dan Stres. (Kemenkes RI, 2021). Masalah pada penelitian ini adalah bagaimana mengetahui model *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular. Tujuan peneliti ini adalah mengetahui kinerja *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular. Metode *K-Nearest Neighbor* (KNN) adalah salah satu metode klasifikasi yang memanfaatkan pola-pola data yang ada dalam *dataset* untuk mengklasifikasi kategori atau kelas dari suatu sampel yang belum diketahui. Hasil klasifikasi data pelatihan menunjukkan akurasi sebesar 85.49%, dengan *precision* 84,43%, *recall* 87,04%, dan *f1-score* 85,71%. Melalui uji coba menggunakan KNN, diperoleh hasil dengan akurasi sebesar 91% dan nilai presisi 90%, *recall* 93%, dan *f1-score* 92%. Kesimpulan dari penelitian ini adalah metode *K-Nearest Neighbor* (KNN) memiliki hasil yang baik untuk melakukan klasifikasi pada penyakit kardiovaskular yaitu akurasinya 91%.

Kata kunci: Penyakit Kardiovaskular, *K-Nearest Neighbors* (KNN), Kematian

Abstract

Cardiovascular diseases are a major factor in global mortality, with the number of victims reaching 17.9 million each year or around 32% of total global deaths (World Health Organization, 2021). Risk factors for cardiovascular disease include age, the older a person is, the higher the risk of developing cardiovascular disease. Other factors include having a family history of cardiovascular disease, diabetes, high blood pressure, obesity, unhealthy lifestyle, and stress. (Kemenkes RI, 2021). The problem in this study is how to determine the K-Nearest Neighbors (KNN) model well through the calculation of accuracy, recall, precision, and f1-score in the classification of cardiovascular disease. The purpose of this research is to determine the performance of K-Nearest Neighbors (KNN) well through the calculation of accuracy, recall, precision, and f1-score in the classification of cardiovascular disease. The K-Nearest Neighbor (KNN) method is a classification method that utilizes existing data patterns in a dataset to classify categories or classes of an unknown sample. The classification results of the training data show an accuracy of 85.49%, with a precision of 84,43%, recall of 87,04%, and f1-score of 85.71%. Through trials using KNN, results were obtained with an accuracy of 91% and a precision value of 90%, recall 93%, and f1-score 92%. The conclusion of this research is that the K-Nearest Neighbor (KNN) method has good results for classifying cardiovascular disease, namely its accuracy 91%.

Keywords: Cardiovascular Disease, K-Nearest Neighbors (KNN), Death

1. PENDAHULUAN

Penyakit Kardiovaskular, yang juga disebut sebagai penyakit gagal jantung dan pembuluh darah, merupakan permasalahan yang serius dalam ranah kesehatan global. Data menunjukkan bahwa penyakit ini merupakan faktor utama kematian di seluruh dunia. Pada tahun 2019, sekitar 17,9 juta kematian dikaitkan dengan penyakit kardiovaskular, menyumbang sekitar 32% dari total kematian global global (*World Health Organization, 2021*)[1]. Data pada tahun 2022, di Kementerian Kesehatan Republik Indonesia (Kemenkes RI) memberikan informasi bahwa faktor penyebab penyakit kardiovaskular melibatkan gangguan pada jantung dan sistem pembuluh darah, termasuk penyakit jantung koroner, gagal jantung, hipertensi, dan stroke (*American Heart Association, 2022*). Kematian di Indonesia akibat penyakit Kardiovaskular mencapai 651.481 penduduk per tahun, yang terdiri dari stroke 331.349 kematian, penyakit jantung koroner 245.343 kematian, Penyakit jantung hipertensi 50.620 kematian, dan penyakit kardiovaskular lainnya (*Institute for Health Metrics and Evaluation, 2019*). Faktor risiko penyakit kardiovaskular diantaranya faktor usia, semakin bertambahnya usia seseorang, maka semakin tinggi risiko terkena penyakit kardiovaskular. Faktor lain yaitu memiliki riwayat penyakit kardiovaskular dalam keluarga, diabetes, tekanan darah tinggi, obesitas (kegemukan), Pola hidup tidak sehat, dan Stres. Ketika seseorang stres, tubuh mereka mengeluarkan hormon kortisol yang mengakibatkan kakunya pembuluh darah (Kemenkes RI, 2021).

Strategi Kementerian kesehatan (Kemenkes) untuk mencegah penyakit kardiovaskular atau penyakit jantung yaitu yang pertama dengan mengupayakan informasi edukasi sebanyak mungkin kepada masyarakat melalui media-media. Kedua adalah deteksi dini, masyarakat diimbau melakukan deteksi dini lebih awal, dengan cara ukur tekanan darah, ukur gula darah, ukur indeks massa tubuh atau lingkaran perut. Ketiga yaitu perlindungan khusus, seperti menerapkan kawasan tanpa rokok. Cara terakhir yaitu pengobatan, yang merupakan fase terakhir ketika seseorang sudah terdeteksi penyakit jantung (Kemenkes RI, 2019). Diagnosis dini dan penanganan yang efektif bagi individu yang menderita jantung memiliki peran yang sangat signifikan. Kondisi kadar gula yang tidak terkontrol dapat menyebabkan faktor kerusakan pada organ tubuh dan meningkatkan risiko terjadinya penyakit jantung, gangguan ginjal, merusak saraf. Karena itu penting untuk mengembangkan metode klasifikasi yang akurat dan efektif dalam proses diagnosis penyakit kardiovaskular[2]. Tidak dapat dianggap remeh pentingnya deteksi dini dan penanganan yang efektif bagi penderita kardiovaskular. Ketidakstabilan kadar gula darah dapat memberikan dampak serius pada organ tubuh, meningkatkan risiko terkena penyakit kardiovaskular, merusak ginjal, dan menimbulkan berbagai komplikasi lainnya. Karena itu, merupakan suatu keharusan untuk mengembangkan metode klasifikasi yang akurat dan efektif dalam proses diagnosis penyakit kardiovaskular.[3].

Klasifikasi penyakit kardiovaskular berdasarkan faktor-faktor risiko tertentu dapat membantu dalam mengidentifikasi kelompok individu yang lebih rentan terhadap penyakit ini. Salah satu pendekatan yang dapat digunakan adalah algoritma *K-Nearest Neighbors* (KNN). Metode *K-Nearest Neighbor* (KNN) adalah salah satu metode klasifikasi yang memanfaatkan pola-pola data yang ada dalam *dataset* untuk mengklasifikasi kategori atau kelas dari suatu sampel yang belum diketahui. Pengklasifikasi *K-Nearest Neighbor* (KNN) telah membuktikan fleksibilitas yang tinggi dalam permasalahan klasifikasi yang kompleks. Penelitian ini mengarahkan fokusnya pada penerapan algoritma *K-Nearest Neighbors* (KNN) dalam mengklasifikasikan penyakit kardiovaskular[4]. Metode KNN digunakan karena Konsep dasar *K-Nearest Neighbors* (KNN) adalah sederhana, yaitu mencari "tetangga terdekat" dari suatu data berdasarkan kesamaan fitur atau atribut. Cara kerja *K-Nearest Neighbors* (KNN) melibatkan perbandingan data yang akan diklasifikasikan dengan data pelatihan yang sudah ada dalam *dataset*. Kelebihan dari algoritma *K-Nearest Neighbors* (KNN) adalah sederhana untuk dipahami dan mudah diimplementasikan.

Penelitian ini akan dilakukan untuk mengembangkan klasifikasi kematian akibat penyakit kardiovaskular dengan memanfaatkan algoritma KNN. Data diambil dari pasien yang terdiagnosa jantung dan yang tidak terdiagnosa akan dianalisis untuk mengidentifikasi pola yang berkaitan

dengan penyakit ini. Berbagai atribut seperti usia, tekanan darah, dan hasil tes laboratorium terkait gula darah akan diintegrasikan sebagai fitur-fitur analisis. Diharapkan bahwa dengan penerapan algoritma KNN, penelitian ini akan memberikan hasil mendiagnosis kardiovaskular yang lebih akurat dan memungkinkan penanganan lebih dini. Dengan demikian, penelitian ini memiliki potensi untuk memberikan sumbangan penting dalam meningkatkan pemahaman kita tentang penggunaan algoritma KNN dalam diagnosa penyakit kardiovaskular[5]. Sejumlah penelitian sebelumnya telah mengembangkan metode klasifikasi kardiovaskular dengan memanfaatkan algoritma KNN. Beberapa penelitian telah mengenali atribut-atribut yang memiliki peran krusial dalam diagnosis kardiovaskular, seperti usia, indeks massa tubuh (BMI), tekanan darah, dan hasil uji laboratorium terkait gula darah. Keefektifan algoritma KNN telah terbukti dalam menciptakan aturan-aturan yang dapat diterapkan untuk mengelompokkan pasien ke dalam kategori penderita kardiovaskular atau bukan kardiovaskular, berdasarkan atribut-atribut tersebut.[6].

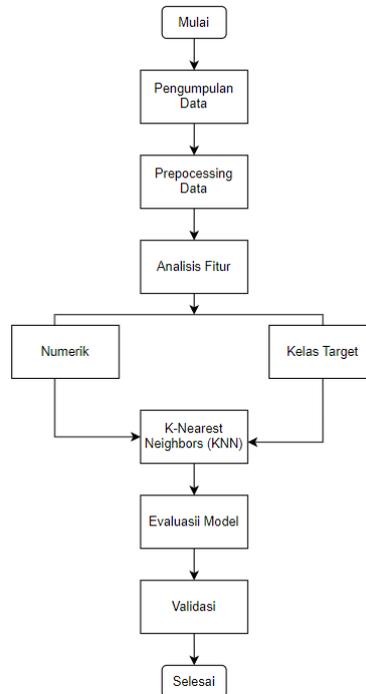
Penelitian oleh Gupta et al. (2018) menggunakan algoritma Decision Tree untuk mendiagnosis kardiovaskular berdasarkan atribut seperti usia, jenis kelamin, riwayat keluarga, indeks massa tubuh, dan riwayat penyakit lainnya. Penelitian ini mencapai tingkat akurasi yang baik dalam mengklasifikasikan pasien sebagai penderita kardiovaskular atau non-kardiovaskular. Pada penelitian ini mempunyai pembeda dengan penelitian sebelumnya, penelitian ini menggunakan algoritma KNN sebagai metode klasifikasi untuk mendiagnosis penyakit kardiovaskular. Algoritma KNN bekerja dengan membandingkan titik data yang baru dengan tetangga terdekatnya dalam ruang fitur dan mengklasifikasikannya berdasarkan mayoritas tetangga terdekat. Pemilihan algoritma ini berbeda dari penelitian sebelumnya yang menggunakan metode klasifikasi lainnya, seperti Decision Tree atau Random Forest[7]. Keputusan menggunakan algoritma KNN dapat memberikan hasil yang berbeda dalam hal akurasi dan performa klasifikasi. Penelitian ini menggunakan dataset terdiri dari 12 fitur yang relevan untuk mengklasifikasi kematian akibat gagal jantung pada penyakit kardiovaskular. Dataset ini memiliki komposisi dan karakteristik yang berbeda dari dataset yang digunakan dalam penelitian sebelumnya. Perbedaan dalam dataset, seperti jumlah dan jenis fitur yang digunakan, dapat mempengaruhi hasil dan keakuratan klasifikasi yang dicapai dalam penelitian ini. Dan dalam penelitian ini memberikan informasi tentang hasil klasifikasi yang dicapai menggunakan algoritma KNN. Hasil ini mencakup akurasi klasifikasi, serta *precision*, *recall*, dan *f1-score* untuk kedua kelas (pasien yang tidak akan meninggal dan pasien yang akan meninggal)[8]. Evaluasi ini memberikan gambaran tentang performa metode klasifikasi dalam penelitian ini. Perbandingan hasil dan evaluasi dengan penelitian sebelumnya dapat membantu dalam memahami keunggulan atau perbedaan dari pendekatan yang digunakan dalam penelitian ini[9], [10].

Berdasarkan pada latar belakang yang telah dijelaskan, rumusan masalah yang diambil adalah bagaimana mengetahui model *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular. Tujuannya mengetahui kinerja *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular agar dapat meningkatkan keakuratan diagnosis penyakit kardiovaskular melalui penggunaan model *K-Nearest Neighbors* (KNN), serta deteksi lebih dini agar memudahkan dokter dalam penanganan pasien yang terkena penyakit kardiovaskular dan tidak. Deteksi lebih dini sangat bagus agar menghindari kematian pada penyakit kardiovaskular ini. Dokter agar lebih mudah melihat keakuratan hasil pemeriksaan.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *K-Nearest Neighbors* (KNN). Metode *K-Nearest Neighbors* (KNN) adalah klasifikasi yang digunakan dalam bidang pembelajaran mesin. Prinsip dasar dari KNN adalah untuk mengklasifikasikan suatu data baru berdasarkan mayoritas klasifikasi dari tetangga terdekatnya dalam ruang fitur. Metode KNN digunakan karena Konsep dasar *K-Nearest Neighbors* (KNN) adalah sederhana, yaitu mencari "tetangga terdekat" dari suatu data berdasarkan kesamaan fitur atau atribut. Cara kerja *K-Nearest Neighbors* (KNN) melibatkan

perbandingan data yang akan diklasifikasikan dengan data pelatihan yang sudah ada dalam *dataset*. Kelebihan dari algoritma *K-Nearest Neighbors* (KNN) adalah sederhana untuk dipahami dan mudah diimplementasikan. Algoritma ini cocok untuk masalah klasifikasi, di mana kita ingin mengklasifikasikan suatu data ke dalam salah satu kelas berdasarkan atribut-atribut yang dimilikinya[11]. Alur metode penelitian dapat dilihat pada gambar dibawah ini:



Gambar 1. Alur Penelitian

2.1 Pengumpulan data

Data untuk penelitian ini yaitu data sekunder dari *dataset* “*Heart Failure Prediction*” yang merupakan *open access* data web dari [BMC Medical Informatics and Decision Making](#)[12]. Data yang dibutuhkan mencakup informasi tentang pasien yang mengalami penyakit kardiovaskular dan fitur-fitur yang dapat digunakan untuk mengklasifikasi kematian akibat gagal jantung. Proses ini dapat melibatkan akses ke data medis elektronik, kerjasama dengan institusi kesehatan, atau penggunaan *dataset* publik yang tersedia. Penting untuk dipastikan bahwa data yang dikumpulkan ini memenuhi kriteria inklusi yang ditetapkan untuk penelitian ini. *Dataset* ini dikumpulkan oleh *Davide Chicco dan Giuseppe Jurman* dari rekam medis 299 pasien gagal jantung di Punjab, Pakistan. Data ini dikumpulkan di *Institut Kardiologi Faisalabad* yang bekerja sama dengan Rumah Sakit *Faisalabad*. Penelitian lain yang menggunakan data ini yaitu (*Sabilla Wilda Imama et al., 2022*) tentang Implementasi Multilayer Perceptron Untuk Memprediksi Harapan Hidup Pada Pasien Penyakit Kardiovaskula [13]. Penelitian ini juga menggunakan data yang sama dengan penelitian yang akan dilakukan sekarang, tetapi terdapat perbedaan pada metode klasifikasi yang akan digunakan yaitu dengan menggunakan metode *K-Nearest Neighbors* (KNN).

Tabel 1. Deskripsi Data

Fitur	Penjelasan	Pengukuran	Jangkauan
<i>Age</i>	Usia	Tahun	[40,...,95]
<i>Anaemia</i>	Penurunan sel darah merah	Boolean	0, 1
<i>High blood pressure</i>	Tekanan darah tinggi	Boolean	0, 1

<i>Creatinine phosphokinase</i>	Tingkat enzim dalam darah	mcg/L	[23,...,7861]
<i>Diabetes</i>	Jika pasien menderita diabetes	Boolean	0, 1
<i>Ejection fraction</i>	Persentase darah yang keluar dari jantung pada setiap kontraksi	Percentage	[14,..., 80]
<i>Platelets</i>	Trombosit didalam darah	kiloplatelets/mL	[25.01,..., 850.00]
<i>Serum creatinine</i>	Tingkat kreatinin dalam darah	mg/dL	[0.50,..., 9.40]
<i>Serum sodium</i>	Tingkat sodium dalam darah	mEq/L	[114,..., 148]
<i>Smoking</i>	Jika pasien merokok	Boolean	0, 1
(target) <i>death event</i>	Jika pasien meninggal selama masa tindak lanjut	Boolean	0, 1

2.2 Preprocessing Data

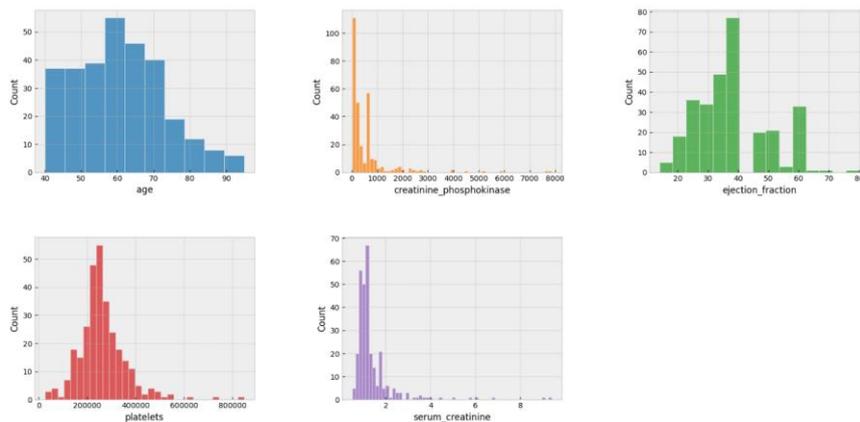
Pada tahapan ini adalah pra-pemrosesan data, yang melibatkan beberapa tahap untuk membersihkan dan mempersiapkan data sebelum diterapkan dalam analisis. Tahap *preprocessing* mencakup langkah-langkah seperti penghapusan data yang tidak lengkap atau tidak valid, penyesuaian skala fitur-fitur yang berbeda agar memiliki rentang yang serupa, dan penghapusan outlier yang mungkin mempengaruhi hasil analisis. Selain itu, data juga akan dipisahkan menjadi subset pelatihan dan pengujian untuk memvalidasi kinerja model. *Preprocessing* yang hati-hati dan teliti sangat penting untuk keakuratan dan keandalan analisis yang akan dilakukan selanjutnya. Hal tersebut juga dapat melibatkan teknik seperti imputasi data yang hilang atau pengkodean ulang variabel kategorikal menjadi bentuk yang dapat dianalisis lebih lanjut[14].

2.3 Analisis Fitur

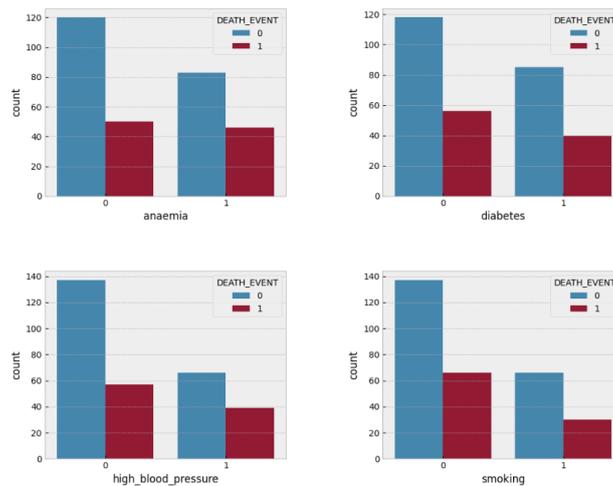
Pada tahap ini, akan di analisis setiap fitur yang ada dalam dataset untuk mengidentifikasi korelasi dan hubungan dengan kematian akibat gagal jantung. Metode analisis statistik dan teknik eksplorasi data dapat digunakan, seperti uji korelasi, analisis regresi, atau visualisasi data. Tujuan dari analisis fitur adalah untuk memahami karakteristik setiap fitur dan memilih fitur-fitur yang paling berpengaruh dalam mengklasifikasi kematian akibat gagal jantung.

Fitur-fitur yang tidak memiliki korelasi yang signifikan atau memberikan kontribusi yang rendah dalam klasifikasi dapat dieliminasi untuk memfokuskan analisis pada fitur-fitur yang lebih informatif. Terdapat dua pilihan fitur yaitu fitur numerik dan fitur Kelas target. Fitur numerik dapat memvisualisasikan histogram untuk setiap fitur numerik dalam dataset secara bersamaan, dan dengan menggunakan warna yang berbeda untuk setiap histogram. Sedangkan Kelas target dapat memvisualisasikan countplot untuk setiap fitur kategori dalam dataset dengan memperhatikan kategori target '*DEATH_EVENT*'.

Histogram fitur Numerik



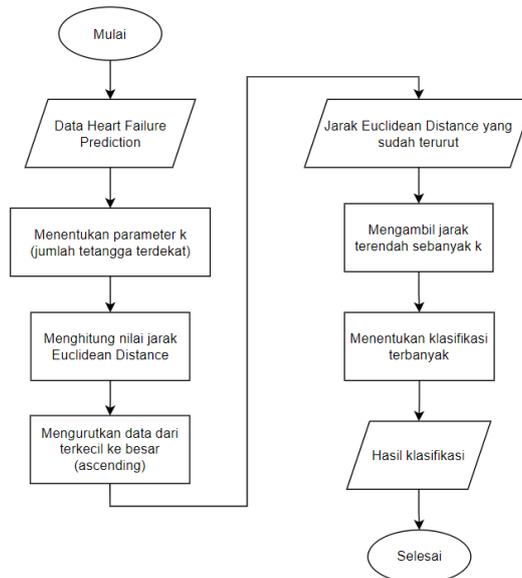
Gambar 2. Fitur Numerik



Gambar 3. Fitur Kelas Target

2.4 Implementasi Metode *K-Nearest Neighbors* (KNN).

Metode *K-Nearest Neighbors* (KNN) yaitu algoritma didalam machine learning yang digunakan untuk mengklasifikasi dan regresi. Algoritma ini didasarkan pada konsep bahwa objek yang memiliki atribut serupa cenderung memiliki label atau nilai target yang sama. Selanjutnya akan menerapkan algoritma *K-Nearest Neighbors* (KNN) sebagai metode klasifikasi dalam penelitian ini. Model KNN akan dibangun dengan menggunakan subset data yang telah dipilih sebagai data pelatihan[15].



Gambar 4 Flowchart Implementasi Metode KNN

Pada gambar 4 *Flowchart* tersebut terdapat beberapa tahap sebelum mendapatkan klasifikasi. Pada tahap awal yaitu *input* data *Heart Failure Prediction*. Sebelum memulai proses perhitungan nilai jarak, langkah berikutnya yaitu menentukan parameter *k*, yang merupakan jumlah tetangga terdekat yang akan digunakan dalam klasifikasi. Setelah itu, langkah selanjutnya adalah menghitung nilai jarak menggunakan metode *Euclidean Distance*.

Setelah proses perhitungan jarak selesai, langkah berikutnya adalah menyusun hasil perhitungan tersebut dalam urutan dari nilai terkecil ke terbesar, atau secara *ascending*. Setelah data diurutkan, langkah selanjutnya adalah mengambil sejumlah *k* data terkecil, sesuai dengan jumlah tetangga terdekat yang sudah ditetapkan sebelumnya.

Untuk mencari pendekatan KNN dapat menggunakan rumus (1):

$$distance(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \tag{1}$$

Hasil dari rumus ini adalah jarak Euclidean antara dua titik data dalam ruang fitur. Dalam konteks KNN, kita akan menggunakan rumus ini untuk mengukur jarak antara titik data yang akan diklasifikasi (x_1) dan titik data pelatihan lainnya (x_2). Kita kemudian akan mengambil k titik data pelatihan terdekat dengan x_1 berdasarkan jarak ini untuk melakukan klasifikasi atau regresi. KNN memungkinkan kita untuk mengklasifikasikan atau mengklasifikasi nilai titik data x_1 berdasarkan mayoritas kelas atau nilai dari k tetangga terdekatnya[11].

Algoritma KNN bekerja dengan mencari k -terdekat tetangga dari setiap titik data yang baru, di mana nilai k adalah parameter yang perlu ditentukan. Langkah-langkah penerapan KNN meliputi menentukan parameter k yang optimal, menghitung jarak antara titik data menggunakan metrik jarak (seperti *Euclidean* atau *Manhattan distance*), dan menentukan kelas target berdasarkan mayoritas kelas tetangga terdekat[16]. Dalam tahap ini, menggunakan implementasi algoritma KNN yang tersedia dalam perpustakaan atau menggunakan kode yang telah dikembangkan sebelumnya[17].

```

GridSearchCV
GridSearchCV(cv=10, estimator=KNeighborsClassifier(n_jobs=-1), n_jobs=-1,
             param_grid={'n_neighbors': array([ 2,  3,  4,  5,  6,  7,  8,  9, 10, 11])},
             scoring='f1')
└─ estimator: KNeighborsClassifier
  KNeighborsClassifier(n_jobs=-1)
  └─ KNeighborsClassifier
    KNeighborsClassifier(n_jobs=-1)
    
```

Gambar 4. GridSearchCV

Pada gambar 4 diatas, adalah GridSearch yang digunakan untuk mencari nilai k . dengan menggunakan itu akan mempermudah mencari nilai k yang terbaik untuk klasifikasi algoritma KNN ini.

2.5 Evaluasi Model

Setelah membangun model KNN, selanjutnya akan dilakukan evaluasi kinerja model untuk mengukur sejauh mana model dapat mengklasifikasi kematian akibat gagal jantung dengan akurat[18]. Evaluasi model dilakukan menggunakan *Confusion matrik model* yang meliputi metrik evaluasi seperti akurasi, presisi, *recall*, F1-score. Selain itu juga akan di evaluasi menggunakan kurva *receiver operating characteristic* (ROC-AUC). Data pengujian yang terpisah dari dataset awal akan digunakan untuk menguji model yang telah dilatih. Dengan membandingkan klasifikasi model dengan label kelas aktual, dan dapat mengevaluasi seberapa baik model dapat mengklasifikasikan kasus kematian akibat gagal jantung[19]. Hasil evaluasi ini akan memberikan wawasan mengenai keandalan dan kualitas model yang dikembangkan. *Confusion matrix model* dapat dilihat pada tabel 2 dibawah ini.

Tabel 2. Confusion matrix model

		Nilai Sebenarnya	
		True	False
Nilai Prediksi	True	FP (False Positive Correct Result)	FP (False Positive Correct Result)
	False	FN (False Negative Correct Result)	FN (False Negative Correct Result)

2.6 Validasi

Tahapan ini, melakukan validasi terhadap model yang dikembangkan. Validasi ini bertujuan untuk memastikan bahwa model dapat menghasilkan klasifikasi yang konsisten dan generalisasi yang baik di data yang belum pernah ada sebelumnya. Metode validasi yang umum digunakan adalah validasi silang, di mana data dibagi menjadi beberapa subset dan model dievaluasi secara berulang menggunakan subset pelatihan dan pengujian yang berbeda. Setelah validasi, akan dibahas hasil dan temuan yang didapatkan dari penelitian ini, termasuk interpretasi terhadap kemampuan model dalam mengklasifikasi kemungkinan pasien meninggal karena gagal jantung. Diskusi juga dapat mencakup keterbatasan model, potensi pengembangan lebih lanjut, dan implikasi praktis dari penelitian ini[20].

3. HASIL DAN PEMBAHASAN

Penelitian ini melakukan klasifikasi penyakit kardiovaskular menggunakan algoritma *K-Nearest Neighbors* (KNN). *Dataset* ini dikumpulkan oleh *Davide Chicco dan Giuseppe Jurman* berdasarkan rekam medis 299 pasien gagal jantung di Punjab, Pakistan. Data yang digunakan terdiri dari 12 kelas fitur yang relevan untuk mengklasifikasikan kematian akibat gagal jantung. [19]. Dalam data tersebut terdapat kelas target yaitu *death event*, kelas target ini untuk mengklasifikasi antara pasien meninggal dan tidak meninggal. Berdasarkan laporan klasifikasi data pelatihan ini mampu memberikan hasil yang cukup akurat dalam mengklasifikasi kematian akibat gagal jantung. Penelitian ini dilakukan split data yaitu dengan rasio 60:40, yang mana data latih 60% dan data uji 40%.

	precision	recall	f1-score	support
not gonna die	0.8662	0.8395	0.8527	162
will die	0.8443	0.8704	0.8571	162
accuracy			0.8549	324
macro avg	0.8553	0.8549	0.8549	324
weighted avg	0.8553	0.8549	0.8549	324

Gambar 5. Hasil klasifikasi data pelatihan

Tahap klasifikasi data pelatihan diatas mendapat akurasi, presisi, recall, F1-score sebesar 85.49%. Selain itu, precision 84,43%, recall 87,04%, dan f1-score 85,71%. Hasil ini memberikan nilai yang baik untuk kedua kelas, menunjukkan kemampuan model dalam mengklasifikasikan dengan benar pasien akibat gagal jantung.

Proses *hyperparameter* yaitu untuk mencari parameter (nilai k) optimal dengan menggunakan *GridSearchCV*. *GridSearchCV* adalah teknik yang digunakan untuk mencari *hyperparameter* terbaik guna membentuk model yang optimal sehingga menghasilkan skor akurasi tertinggi. *GridSearchCV* dirancang untuk menentukan *hyperparameter* optimal yang akan menghasilkan performa terbaik untuk model *K-Nearest Neighbors*. Parameter yang digunakan peneliti yaitu:

Tabel 2. Parameter

Parameter	Keterangan
$n_neighbors=list(range(2,12))$	Nilai jarak k yang dicari mulai dari 2-11
$weight_options = [“uniform”, distance”]$	Berfungsi mengontrol bagaimana tetangga diberi "bobot" saat melakukan klasifikasi

Hasil parameter nilai k pada penelitian ini didapatkan melalui teknik *GridSearchCV* sebagai berikut:

Tabel 3. Hasil rata-rata skor dan parameter (k)

<i>mean_test_score</i>	<i>params</i>
0.762976	{'n_neighbors': 2}
0.869284	{'n_neighbors': 3}
0.839076	{'n_neighbors': 4}
0.875669	{'n_neighbors': 5}
0.823318	{'n_neighbors': 6}
0.859506	{'n_neighbors': 7}
0.824845	{'n_neighbors': 8}
0.855720	{'n_neighbors': 9}
0.836012	{'n_neighbors': 10}
0.844570	{'n_neighbors': 11}

Tabel 3 menunjukkan hasil parameter yang didapatkan, kemudian akan diambil satu parameter yang terbaik yaitu nilai *k* (nilai *n_neighbor*) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah $k = 5$ dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah $= 0.875669$.

Pemilihan nilai *k* yang optimal juga perlu diperhatikan, karena sangat berpengaruh terhadap hasil akurasi yang akan didapatkan. Semakin baik nilai *k* maka nilai akurasi akan semakin baik dan lebih akurat. Dengan terus mengembangkan dan meningkatkan model klasifikasi kematian akibat gagal jantung, diharapkan dapat memberikan kontribusi yang signifikan dalam diagnosis dini, penanganan yang tepat, dan pencegahan penyakit kardiovaskular, serta meningkatkan kualitas hidup pasien.

Hasil klasifikasi pada data uji seperti pada gambar 6 dibawah.

```

Accuracy: 0.91
Precision: 0.90
Recall: 0.93
F1 Score: 0.92

Classification Report:
              precision    recall  f1-score   support

     0       0.93      0.90      0.91         41
     1       0.90      0.93      0.92         41

   accuracy          0.91
  macro avg          0.91
 weighted avg          0.91

```

```

Confusion Matrix:
[[37  4]
 [ 3 38]]

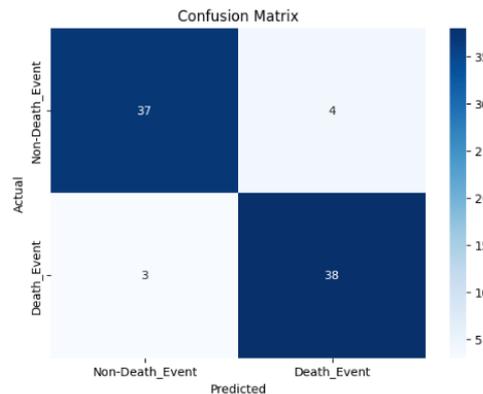
```

Gambar 6. Hasil Klasifikasi Data Uji

Hasil klasifikasi data uji dengan metode *K-Nearest Neighbors* (KNN) pada penelitian disini menghasilkan model yang dapat mengklasifikasi kematian akibat gagal jantung dengan tingkat akurasi yang baik. Evaluasi model menunjukkan bahwa KNN dapat mengklasifikasikan pasien dengan presisi, *recall*, dan *f1-score* yang tinggi. Untuk pasien yang tidak meninggal akibat gagal jantung, model memiliki presisi sebesar 93%. Selain itu, *recall* sebesar 90% menunjukkan kemampuan model dalam mengidentifikasi sebagian besar pasien yang sebenarnya meninggal akibat gagal jantung.

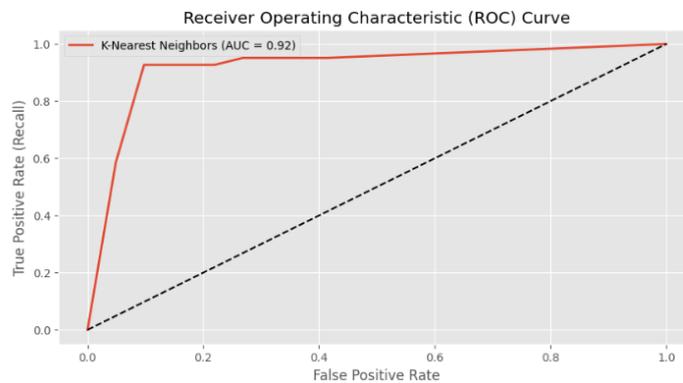
Untuk pasien yang meninggal akibat gagal jantung, model memiliki presisi 90% dan *recall* 93%, menunjukkan kemampuan model dalam mengklasifikasikan dengan benar pasien yang meninggal. Seluruh performa ini menghasilkan *f1-score* sebesar 0.91 untuk pasien yang tidak meninggal dan 0.92 untuk pasien yang meninggal akibat gagal jantung. Dengan tingkat akurasi sebesar 91%, model KNN memberikan klasifikasi yang cukup akurat dalam mendiagnosis kemungkinan kematian akibat gagal jantung.

Menurut [22] 90%-100% merupakan kategori klasifikasi sangat baik, 80%-90% merupakan kategori klasifikasi baik, 70%-80% merupakan kategori klasifikasi cukup baik, 60%-70% merupakan kategori klasifikasi kurang baik, dan 50%-60% merupakan kategori klasifikasi gagal. Berdasarkan pernyataan tersebut maka hasil penelitian ini termasuk kategori klasifikasi sangat baik karena menghasilkan nilai akurasi 91%.



Gambar 7. *Confusion Matrix*

Berdasarkan hasil *Confusion Matrix* diatas, terdapat 37 jumlah pasien yang benar-benar tidak meninggal, dan model berhasil mengidentifikasi mereka dengan benar. Terdapat 38 mencerminkan pasien meninggal. Di sisi lain, angka 4 menunjukkan kesalahan di mana pasien yang sebenarnya tidak meninggal diklasifikasi meninggal, sedangkan angka 3 mencerminkan pasien yang sebenarnya meninggal, tetapi diklasifikasi tidak meninggal.



Gambar 7. Kurva ROC

Terlihat bahwa model KNN memiliki jarak yang signifikan dari garis putus-putus yang mewakili kurva ROC dari pengklasifikasi, menunjukkan kinerja yang baik dalam membedakan pasien yang meninggal akibat gagal jantung. Selain itu, area di bawah kurva (AUC) sebesar 0.92 yang dihasilkan oleh model KNN adalah yang tertinggi, mendekati nilai 1. Nilai AUC ketika mendekati nilai 1 maka hasil tersebut sangat baik, sedangkan ketika jauh dari nilai 1 maka hasilnya kurang baik. Hal ini menunjukkan bahwa model KNN adalah model terbaik dalam melakukan klasifikasi pada dataset ini, dengan kemampuan yang superior dalam mengklasifikasi kematian akibat gagal jantung.

4. KESIMPULAN DAN SARAN

Penelitian ini menerapkan algoritma *K-Nearest Neighbors* (KNN) klasifikasi pada pasien dengan penyakit kardiovaskular. Berdasarkan hasil uji coba dengan dilakukan split data yaitu dengan rasio 60:40, yang mana data latih 60% dan data uji 40%. Mendapatkan hasil akurasi 91% serta nilai presisi 90%, recall 93%, dan f1-score 92%. Selain didapatkan hasil akurasi, presisi, recall, dan f1-score didapatkan juga hasil AUC sebesar 0,92 atau 92%. Hasil tersebut adalah sangat baik karena nilai AUC mendekati nilai 1.

Hasil dari penelitian sebelumnya “Implementasi *K-Nearest Neighbor* Untuk *Prototype* Sistem Pakar Identifikasi Dini Penyakit Jantung” [23] menunjukkan tingkat akurasi *k-nearest neighbor* sebesar 70.04%. Hasil dari “*Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification*” menunjukkan tingkat akurasi 80.6% yang menggunakan *dataset Heart Failure Prediction* dengan 299 data. Hasil penelitian dari [22] menunjukkan bahwa pada jurnal ini mengklasifikasi pasien berdasarkan *serum creatinine* dan *ejection fraction*, yang menggunakan beberapa metode dan evaluasi *ROC*. *Dataset* yang digunakan adalah *Heart Failure Prediction* dengan 299 data. Dari penelitian ini dihasilkan *Random forests* memiliki akurasi 74% dan *AUC* 80%. Selain akurasi, penelitian ini juga mengevaluasi metrik penting lainnya seperti presisi (90%), recall (93%), dan f1-score (92%).

Kesimpulannya penelitian ini menunjukkan bahwa memberikan kontribusi yaitu metode KNN dapat mencapai akurasi yang sangat tinggi yaitu 91%. Ini adalah peningkatan yang signifikan dibandingkan dengan penelitian sebelumnya yang menggunakan algoritma serupa atau berbeda, seperti yang ditunjukkan dengan akurasi 70.04% pada penelitian sebelumnya menggunakan KNN [23] dan akurasi 80.6% menggunakan *Naive Bayes Classifier* pada dataset yang sama [22]. Saran untuk penelitian ini adalah dapat menggunakan metode klasifikasi yang berbeda sehingga dapat membandingkan nilai akurasi yang didapatkan agar mendapatkan metode klasifikasi yang terbaik, dan dapat menggunakan proses *hyperparameter* yang berbeda selain *GridSearchCV* untuk *K-Nearest Neighbor*.

DAFTAR PUSTAKA

- [1] A. Lianou, D. Rallis, M. Baltogianni, A. Vlahos, H. Milionis, And V. Giapros, “Biomarkers Of Cardiovascular Disease Risk In The Neonatal Population,” *Journal Of Developmental Origins Of Health And Disease*, Vol. 14, No. 2. Cambridge University Press, Pp. 155–165, Apr. 03, 2023. Doi: 10.1017/S2040174422000459.
- [2] H. Hachimi, École Nationale Des Sciences Appliquées De Kénitra, Institute Of Electrical And Electronics Engineers, And Institute Of Electrical And Electronics Engineers. Morocco Section, *2019 International Conference On Optimization And Applications (Icoa) : April 25-26, 2019, Ensa Of Kenitra, Ibn Tofail University, Kenitra, Morocco*. 2019.
- [3] K. M. Mohi Uddin, R. Ripa, N. Yeasmin, N. Biswas, And S. K. Dey, “Machine Learning-Based Approach To The Diagnosis Of Cardiovascular Vascular Disease Using A Combined Dataset,” *Intell Based Med*, Vol. 7, Jan. 2023, Doi: 10.1016/J.Ibmed.2023.100100.
- [4] W. Sangkatip And J. Phuboon-Ob, “Non-Communicable Diseases Classification Using Multi-Label Learning Techniques,” In *Incit 2020 - 5th International Conference On*

- Information Technology*, Institute Of Electrical And Electronics Engineers Inc., Oct. 2020, Pp. 17–21. Doi: 10.1109/Incit50588.2020.9310978.
- [5] M. Swathy And K. Saruladha, “A Comparative Study Of Classification And Prediction Of Cardio-Vascular Diseases (Cvd) Using Machine Learning And Deep Learning Techniques,” *Ict Express*, Vol. 8, No. 1, Pp. 109–116, Mar. 2022, Doi: 10.1016/J.Icte.2021.08.021.
- [6] T. S. H. Jørgensen, M. K. Wium-Andersen, M. B. Jørgensen, And M. Osler, “The Impact Of Mental Vulnerability On The Relationship Between Cardiovascular Disease And Depression,” *European Psychiatry*, Vol. 63, No. 1, 2020, Doi: 10.1192/J.Eurpsy.2020.20.
- [7] Savita, G. Sharma, G. Rani, And V. Singh Dhaka, “A Review On Machine Learning Techniques For Prediction Of Cardiovascular Diseases,” In *Pdgc 2020 - 2020 6th International Conference On Parallel, Distributed And Grid Computing*, Institute Of Electrical And Electronics Engineers Inc., Nov. 2020, Pp. 237–242. Doi: 10.1109/Pdgc50313.2020.9315747.
- [8] B. E. Alekseev, A. I. Skorobogatova, D. A. Egorov, And A. A. Anisimov, “Baroreflex Function Analysis For Assessing The Cardiovascular System’s Health,” In *Proceedings Of The 2021 Ieee Conference Of Russian Young Researchers In Electrical And Electronic Engineering, Elconrus 2021*, Institute Of Electrical And Electronics Engineers Inc., Jan. 2021, Pp. 1713–1717. Doi: 10.1109/Elconrus51938.2021.9396321.
- [9] M. El Bekkali, Institute Of Electrical And Electronics Engineers. Morocco Section, F. École Nationale Des Sciences Appliquées, Jāmi‘at Sīdī Muḥammad Ibn ‘abd Allāh, And Institute Of Electrical And Electronics Engineers, *The 7th Mediterranean Congress Of Telecommunications 2019 : Cmt 2019 : October 24-25, 2019, Ensa Of Fez - Transmission And Information Processing Laboratory, Usmba University, Fez, Morocco*. 2019.
- [10] R. Hagan, C. J. Gillan, And F. Mallett, “Comparison Of Machine Learning Methods For The Classification Of Cardiovascular Disease,” *Inform Med Unlocked*, Vol. 24, Jan. 2021, Doi: 10.1016/J.Imu.2021.100606.
- [11] Z. Li And H. Zhang, “Fusing Deep Metric Learning With Knn For 12-Lead Multi-Labelled Ecg Classification,” *Biomed Signal Process Control*, Vol. 85, Aug. 2023, Doi: 10.1016/J.Bspc.2023.104849.
- [12] A. B. Wibisono And A. Fahrurrozi, “Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner,” *Jurnal Ilmiah Teknologi Dan Rekayasa*, Vol. 24, No. 3, Pp. 161–170, 2019, Doi: 10.35760/Tr.2019.V24i3.2393.
- [13] Sabilla Wilda Imama, Vista Candra Bella, And Suryani Dhebys, “Implementasi Multilayer Perceptron Untuk Memprediksi Harapan Hidup Pada Pasien Penyakit Kardiovaskular,” 2022.
- [14] T. Magrupov, Y. Talatov, M. Magrupova, And D. Ripka, “A Technique For Classifying The Ecg Signal Into Various Possible States Of The Cardiovascular System,” In *Proceedings Of The 2020 Ieee International Conference On Electrical Engineering And Photonics, Eexpolytech 2020*, Institute Of Electrical And Electronics Engineers Inc., Oct. 2020, Pp. 127–131. Doi: 10.1109/Eexpolytech50912.2020.9243864.
- [15] S. Alim, “Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes Orange Data Mining Implementation For Student Graduation Classification Using K-Nearest Neighbor, Decision Tree And Naive Bayes Models,” 2021.
- [16] T. A. Munandar And A. Q. Munir, “Implementasi K-Nearest Neighbor Untuk Prototype Sistem Pakar Identifikasi Dini Penyakit Jantung K-Nearest Neighbor For Prototype Expert System For Early Identification Of Heart Disease,” *Jurnal Teknologi Informasi*, Vol. Xvii, No. 2, 2022.
- [17] H. W. Dhany, *Seminar Nasional Informatika (Senatika) Prosiding Senatika 2021 Performa Algoritma K-Nearest Neighbour Dalam Memprediksi Penyakit Jantung*. 2021. [Online]. Available: <https://www.kaggle.com/>

- [18] M. Ary, D. Ayu, And F. Rismiati, “Satin-Sains Dan Teknologi Informasi Ukuran Akurasi Klasifikasi Penyakit Mesothelioma Menggunakan Algoritma K-Nearest Neighbor Dan Backward Elimination,” Vol. 5, No. 1, 2019, [Online]. Available: [Http://Jurnal.Stmik-Amik-Riau.Ac.Id](http://jurnal.stmik-amik-riau.ac.id)
- [19] A. Samosir, M. Hasibuan, W. E. Justino, And T. Hariyono, “Komparasi Algoritma Random Forest, Naïve Bayes Dan K-Nearest Neighbor Dalam Klasifikasi Data Penyakit Jantung,” *Cardiovascular Diseases* , Vol. 15, No. 2, Pp. 189–194, 2019.
- [20] M. Sholeh, D. Andayati, R. Yuliana Rachmawati, P. Studi Informatika, And F. Teknologi Informasi Dan Bisnis, “Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes Data Mining Model Classification Using Algorithm K-Nearest Neighbor With Normalization For Diabetes Prediction,” 2022.
- [21] R. Arisandi, “Perbandingan Model Klasifikasi Random Forest Dengan Resampling Dan Tanpa Resampling Pada Pasien Penderita Gagal Jantung,” *Jurnal Gaussian*, Vol. 12, No. 1, Pp. 136–145, May 2023, Doi: 10.14710/J.Gauss.12.1.136-145.
- [22] P. Subarkah, W. R. Damayanti, And R. A. Permana, “Comparison Of Correlated Algorithm Accuracy Naive Bayes Classifier And Naive Bayes Classifier For Classification Of Heart Failure,” *Ilkom Jurnal Ilmiah*, Vol. 14, No. 2, Pp. 120–125, Aug. 2022, Doi: 10.33096/Ilkom.V14i2.1148.120-125.
- [23] T. A. Munandar And A. Q. Munir, “Implementasi K-Nearest Neighbor Untuk Prototype Sistem Pakar Identifikasi Dini Penyakit Jantung K-Nearest Neighbor For Prototype Expert System For Early Identification Of Heart Disease,” *Jurnal Teknik Informatika*, 2022.
- [24] D. Chicco And G. Jurman, “Machine Learning Can Predict Survival Of Patients With Heart Failure From Serum Creatinine And Ejection Fraction Alone,” *Bmc Med Inform Decis Mak*, Vol. 20, No. 1, Feb. 2020, Doi: 10.1186/S12911-020-1023-5.