MATHunesa
Jurnal Ilmiah Matematika
e-ISSN : 2716-506X | p-ISSN : 2301-9115

Volume 10 No 02
Tahun 2022

KOMPARASI METODE *SUPPORT VECTOR MACHINE* (SVM), *K-NEAREST NEIGHBORS* (KNN), DAN *RANDOM FOREST* (RF) UNTUK PREDIKSI PENYAKIT GAGAL JANTUNG

Socayo Adi
Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya
socayo.18057@mhs.unesa.ac.id

Atik Wintarti
Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya
atikwintarti@unesa.ac.id

Abstrak

Jantung adalah organ yang mempunyai peranan penting dalam kelangsungan hidup manusia karena fungsinya untuk mendistribusikan darah dari paru-paru ke seluruh bagian tubuh, yang dimana darah tersebut mengandung banyak sekali oksigen sehingga dapat membantu proses metabolisme di dalam tubuh manusia. Maka dari itu, organ jantung perlu dilindungi, dirawat, dan dijaga kondisinya untuk mencegah kerusakan pada jantung yang mengakibatkan penyakit gagal jantung. Begitu banyak orang yang meninggal karena penyakit gagal jantung, sehingga harus ada penelitian yang terbarukan untuk memprediksi penyakit gagal jantung tersebut. Peranan *Artificial Intelligence* (AI) dalam mendeteksi gagal jantung pada penelitian ini bertujuan untuk membandingkan prediksi akurasi dari tiga metode yang dipakai oleh peneliti, yaitu metode *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Random Forest* (RF). Dataset diambil dari laman *UCI Machine Learning* dengan judul *Heart Failure Clinical Records Dataset*. Akurasi paling tinggi yang dapat dihasilkan adalah menggunakan metode SVM dan RF dimana menghasilkan akurasi yang bernilai sama, yaitu 97% dengan rincian metode SVM menggunakan parameter *C* = 1, *gamma* = 0.01, *kernel* = linear dan dalam total waktu *running* program selama 2.82 detik sedangkan metode RF menggunakan parameter *n_estimators* = 30, *random_state* = 0, dan dalam total waktu *running* program selama 7.29 detik. Lalu untuk metode KNN menghasilkan akurasi yang bernilai 93%, dengan menggunakan parameter *n_neighbors* = 20 dan dalam total waktu *running* program selama 0.60 detik. Pada penelitian ini, peneliti juga membuat program Prediksi Gagal Jantung dengan *Graphical User Interface* (GUI) menggunakan ketiga metode yang telah diuji akurasinya.
Kata Kunci: gagal jantung, *support vector machine* (SVM), *k-nearest neighbors* (KNN), *random forest* (RF)

Abstract

The heart is an organ that has an important role in human lives because of its function to distribute blood from the lungs to all parts of the body, where the blood contains a lot of oxygen so that it can help metabolic processes in the human body. Therefore, the heart organ needs to be protected, cared for, and maintained in its condition to prevent damage of the heart that causes heart failure. So many people die of heart failure, so there must be a new study to predict heart failure. The role of artificial intelligence (AI) in detecting heart failure in this study aims to compare the prediction accuracy of the three methods used by researchers, namely the Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF). The dataset is taken from the UCI Machine Learning page with the title Heart Failure Clinical Records Dataset. The highest accuracy generated by using the SVM and RF methods which produce an accuracy of the same value, which is 97% with details of the SVM method using parameters C = 1, gamma = 0.01, kernel = linear and in total the program running time is 2.82 seconds while the RF method uses the parameter n_estimators = 30, random_state = 0, and the total program running time is 7.29 seconds. Then the KNN method produces an accuracy of 93%, using parameters n_neighbors = 20 and the total program running time is 0.60 seconds. In this study, we also made an application Heart Failure Prediction with Graphical User Interface (GUI) using three methods that have been tested for accuracy.
Keywords: *heart failure, support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF)*

PENDAHULUAN

Manusia memiliki organ yang selalu bekerja untuk bertahan hidup. Jantung adalah salah satu alat vital yang paling utama dalam kehidupan manusia. Jantung mempunyai peranan penting untuk memompa darah, sehingga darah dapat memasok oksigen dan menutrisi ke seluruh tubuh. Jumlah waktu yang dibutuhkan untuk jantung bekerja dan berdetak dalam tubuh manusia biasanya mengacu pada gerakan denyut jantung. (Rozie, Hadary, & Wigyarianto, 2016). Manusia tidak dapat mengontrol

denyut jantung mereka saat jantung mereka memompa darah atau saat jantung mereka bekerja karena manusia tidak mempunyai kemampuan untuk meningkatkan atau menurunkan kecepatan denyut jantung mereka. Jantung tidak akan pernah berhenti bekerja walaupun manusia sedang istirahat, sehingga jika manusia mempunyai kemampuan untuk mengontrol keadaan dan denyut jantung mereka, maka manusia sendiri akan kewalahan dalam mengatur jantung mereka sendiri dan mengakibatkan datangnya penyakit jantung bahkan bisa terjadi gagal jantung (Florencia, 2019)

Menurut data WHO (*World Health Organization*), sepertiga dari kematian secara global disebabkan oleh penyakit jantung. Meningkatnya angka penyakit gagal jantung dengan angka kematian yang tinggi menjadi beban yang cukup berat bagi sistem perawatan kesehatan di seluruh dunia. Penyakit gagal jantung menyebabkan kematian sekitar 17,9 juta orang setiap tahun di seluruh dunia dan memiliki prevalensi yang lebih tinggi di Asia (Ghosh et al., 2021). Rata-rata setengah dari semua pasien yang didiagnosis dengan penyakit gagal jantung meninggal hanya dalam waktu 1-2 tahun (Haq et al., 2018).

Manusia pasti tidak ingin organ jantungnya bermasalah agar dapat merasakan kelangsungan hidup yang panjang dan mengurangi kematian, maka dari itu dibutuhkan antisipasi sejak dini terkait penyakit gagal jantung. Untuk dapat memprediksi penyakit gagal jantung, beberapa tes sangat diperlukan. Kurangnya keahlian dari staf medis dapat menghasilkan prediksi yang salah (Pouriyeh et al., 2017). Salah satu cara untuk mengetahui penyakit gagal jantung adalah dengan menggunakan alat rekam medis karena alat tersebut dapat mengukur efek samping, sorotan tubuh yang dapat digunakan untuk melakukan analisis biostatistik terapi (Chicco & Jurman, 2020). Secara teknologi komputasi, *machine learning* juga dapat memprediksi informasi yang didapat dan mengidentifikasi komponen terpenting yang harus diingat untuk catatan klinis (Chicco & Jurman, 2020).

Dalam penelitian yang dilakukan oleh peneliti, pengklasifikasian bertujuan untuk mendeteksi kelangsungan hidup pasien penyakit gagal jantung dan penelitian ini diharapkan mampu memperbaharui hasil prediksi akurasi pada penelitian-penelitian yang sudah dilakukan sebelumnya. Contoh dari penelitian-penelitian yang sudah pernah dilakukan sebelumnya yaitu oleh Trianifa dkk. pada tahun 2019 dengan menggunakan metode *Support Vector Machine* (SVM), penelitian yang dilakukan oleh S. Rahayu dkk. pada tahun 2020 dan penelitian yang dilakukan oleh Newaz dkk. pada tahun 2021 dengan menggunakan metode *Random Forest* (RF).

Pada penelitian ini akan dilakukan komparasi dari metode *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Random Forest* (RF). Tiga metode tersebut digunakan oleh peneliti untuk membandingkan metode mana yang lebih cocok digunakan pada dataset. Dataset yang digunakan diambil dari laman *UCI Machine Learning* dengan judul *Heart Failure Clinical Records Dataset*.

KAJIAN TEORI

Penyakit Gagal Jantung

Jantung merupakan salah satu alat vital yang peranannya paling penting dalam tubuh manusia. Organ jantung mempunyai rongga yang berisi cairan dimana fungsinya untuk melumasi jantung selama berdenyut dan jantung terletak di tengah tulang rusuk. Darah di dalam jantung dialirkan melalui bagian kanan dan kiri jantung itu sendiri. Jantung memiliki dua katup dimana katup tersebut berfungsi untuk memastikan bahwa darah hanya akan mengalir dalam satu arah saja.

Peran paling utama dari jantung yaitu untuk mensuplai O_2 atau oksigen ke bagian-bagian tubuh manusia dan mengeluarkan metabolit yang sudah tidak terpakai dari dalam tubuh. Jantung bekerja dengan cara menjadikan kumpulan darah yang sudah mengandung CO_2 atau karbondioksida dari seluruh tubuh lalu mengirimkannya ke paru-paru, kemudian di dalam paru-paru akan dilakukan proses pengubahan dari karbondioksida menjadi oksigen. Lalu jantung akan mengirimkan kumpulan darah yang banyak mengandung oksigen dari paru-paru ke bagian-bagian tubuh manusia (Purbianto & Agustanti, 2017).

Gagal jantung adalah suatu kondisi dimana masih terdapat darah yang kembali ke jantung dalam keadaan normal, artinya darah tersebut belum terproses secara sempurna di dalam jantung akan tetapi jantung sudah tidak mampu lagi untuk memompa darah yang masih normal tersebut ke dalam jaringan tubuh untuk keperluan metabolisme.

Dengan kata lain, gagal jantung dapat terjadi jika jantung sudah tidak mampu lagi memompa darah yang kaya akan oksigen untuk memenuhi kebutuhan metabolisme di dalam tubuh (Agustina, Afiyanti, & Ilmi, 2017).

Gagal jantung merupakan gangguan multisistem dimana jantung mengalami disfungsi. Akibat dari tidak berfungsinya dari sistolik, maka akan menyebabkan penurunan curah jantung pada ventrikel sebelah kiri. Sedangkan akibat tidak berfungsinya dari diastolik, maka menyebabkan menurunnya komplians vertikel sebelah kanan yang terjadi karena gangguan relaksasi miokard (Hapsari A, 2021).

Support Vector Machine

Support vector machine (SVM) adalah mesin pembelajaran yang cara kerjanya dengan mencari fungsi pemisah terbaik untuk memisahkan kelas, dan SVM ini termasuk dalam pendekatan *supervised learning*. SVM menggunakan metode pembelajaran mesin berupa fungsi linier pada ruang fitur berdimensi tinggi yang menggunakan ruang maya dan dilatih dengan metode pembelajaran mesin berdasarkan teori optimasi yang diturunkan dari teori statistik (Putra & Rini, 2020). Pengenalan pola yang baik dapat digeneralisasi memakai metode ini.

Konsep pengklasifikasian SVM bertujuan untuk menemukan *hyperplane* (bidang pemisah) yang terbaik sehingga dapat digunakan untuk memisahkan dua kelas data pada ruang input. Cara untuk menemukan *hyperplane* adalah dengan mencari titik yang maksimum dan menghitung *margin hyperplane*. Margin merupakan jarak antara *hyperplane* dengan data yang paling dekat dari setiap kelas, sedangkan data yang terdekat dengan *hyperplane* disebut *support vector* (Kramar et al., 2018).

Beberapa hal yang perlu diperhatikan saat menggunakan metode SVM adalah mengenal *hyperplane* dan *support vector*. Fungsi dari *hyperplane* adalah sebagai alat pembantu untuk membatasi titik data yang diklasifikasi. *Support vector* adalah titik data yang dapat mempengaruhi orientasi dan posisi *hyperplane*.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) adalah pembelajaran mesin dengan pendekatan *supervised learning* yang sangat mudah untuk dipahami karena metode ini termasuk klasifikasi *lazy learner* dimana data latih akan disimpan terus-menerus dan akan dilakukan pemrosesan ketika data uji sudah muncul sehingga metode ini dikatakan metode yang mudah dipahami (Lutimath, Chethan & Pol, 2019). Objek yang diklasifikasikan menggunakan metode ini akan diklasifikasi berdasarkan data latih yang tetangganya paling dekat dengan objek atau selisih dari objek yang terkecil. KNN adalah sebuah metode yang dimonitori dengan hasil *instance query* yang diklasifikasikan berdasar pada sebagian besar kategori dalam KNN.

Metode KNN ini bertujuan untuk mengklasifikasi objek baru dengan atribut dan *training sample*. Prinsip umum dari metode ini adalah mencari $k-$ data pelatihan untuk menentukan tetangga terdekat berdasarkan besar kecilnya jarak (Andiani, Sukemi, & Rini, 2020). Selain itu, metode ini dapat digunakan untuk memperkirakan dan memprediksi data, tetapi metode ini lebih sering digunakan untuk mengklasifikasikan teknik data *mining* karena metode ini hanya memerlukan waktu *running* program yang sangat singkat.

Random Forest

Random forest (RF) adalah metode untuk mengklasifikasikan sejumlah besar data. Klasifikasi RF dilakukan dengan menggabungkan beberapa pohon (*trees*) dengan pelatihan pada data sampel. Penggunaan *trees* yang *massive* dapat mempengaruhi nilai akurasi yang dihasilkan. Keputusan klasifikasi RF didasarkan pada jumlah voting dari pohon yang telah dibentuk sehingga dari pohon yang sudah dibentuk akan ditentukan suara yang paling banyak (S. Rahayu et al., 2020).

Pohon yang dibangun pada metode ini harus mencapai ukuran maksimum dari $k-$ data. Pengembangan dilakukan dengan menggunakan pemilihan fitur secara acak untuk meminimalkan kesalahan.

RF merupakan cara untuk menerapkan pendekatan identifikasi stokastik untuk proses klasifikasi. Proses klasifikasi akan berjalan ketika semua pohon terbentuk. Ketika proses klasifikasi sudah selesai, akan dilakukan inisialisasi dengan data sebanyak mungkin berdasarkan skor akurasi (Andiani, Sukemi, & Rini, 2020).

Confusion Matrix

Confusion matrix adalah tabel untuk menggambarkan kinerja sampel klasifikasi yang digunakan untuk mengukur kinerja pada kumpulan data. *Confusion matrix* memberikan informasi mengenai klasifikasi aktual dan prediksi yang telah dibuat oleh kinerja sistem sehingga hasil dari suatu sistem sering dievaluasi menggunakan tes *confusion matrix* (Narkhede, 2018).

Metode ini digunakan untuk menghitung keakuratan konsep data *mining*. *Confusion matrix* dapat dinyatakan seperti Tabel 1.

Tabel 1. *Confusion Matrix*

Actual	***Predicted***	
	Gagal jantung	**Tidak gagal jantung**
Gagal jantung	*True Positive*	*False Negative*
Tidak gagal jantung	*False Positive*	*True Negative*

Keterangan:

- *True Positive* (TP) adalah prediksi hasil yang bernilai positif, data sebenarnya yang positif.
- *True Negative* (TN) adalah prediksi hasil yang bernilai negatif, data yang sebenarnya negatif.
- *False Positive* (FP) adalah perdiksi hasil yang bernilai positif, data yang sebenarnya negatif.
- *False Negative* (FN) adalah prediksi hasil yang bernilai negatif, data yang sebenarnya positif.

Dengan metode *confusion matrix* dapat dilakukan perhitungan dan menghasilkan empat nilai, yaitu: *recall*, *precision*, *F1 score, accuracy*.

(1) *Recall* adalah tingkat probabilitas dengan hasil yang positif dan ditentukan dengan benar. Rumus dari *recall* adalah :

$$recall = \frac{TP}{TP+FN} \times 100\% \quad (1)$$

(2) *Precision* adalah tingkat probabilitas dengan hasil yang positif dan ditentukan dengan benar. Rumus dari *precision* adalah :

$$precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

(3) *Accuracy* adalah nilai hasil untuk menentukan keakuratan model. Rumus dari *accuracy* adalah :

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (3)$$

(4) *F1 Score* adalah nilai hasil untuk mengetahui nilai rata-rata dari perbandingan keluaran *precision* maupun keluaran *recall*. Rumus dari F1 *Score* adalah :

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall+Precisioin} \times 100\% \quad (4)$$

METODE

Peneliti menggunakan tiga metode dalam melakukan prediksi ini, yaitu metode SVM, KNN, dan RF. Banyak sekali metode untuk melakukan klasifikasi pada pengolahan data dan pembelajaran mesin. Tetapi ketiga metode ini memiliki keunggulan yang sangat signifikan saat peneliti menggunakannya, yaitu metode SVM, KNN, dan RF.

Tiga metode tersebut dipilih peneliti karena memiliki keunggulannya masing-masing. Metode SVM memiliki keunggulan yaitu dapat menentukan *hyperplane* dengan cara memilih bidang yang memiliki optimal margin maka generalisasi SVM dapat terjaga dengan sendirinya (Wibisono & Fahrurozi, 2019). Metode KNN memiliki keunggulan yaitu mudah diimplementasikan karena cukup mendefinisikan fungsi untuk menghitung jarak antar kelas. Sedangkan keunggulan metode RF yaitu mampu mengklasifikasikan data yang memiliki atribut tidak lengkap dan dapat digunakan pada data sampel yang banyak (Pouriyeh et al., 2017).

Support vector machine (SVM)

SVM merupakan metode pengklasifikasian yang bisa digunakan secara linier atau non-linier. Cara untuk menemukan *hyperplane* yang terbaik adalah dengan mencari titik yang maksimum dan menghitung *margin hyperplane*. Tujuan dari metode SVM adalah untuk menghitung hasil dari *support vector* sehingga peneliti hanya perlu mengetahui fungsi kernelnya, dan fungsi non-liniernya tidak perlu untuk diketahui (Putra & Rini, 2020).

Persamaan *support vector machine* (SVM) :

$$f(x) = w^t \Phi(x) + b \quad (5)$$

Keterangan :

b = Bias
$x = (x_1, x_2, \dots, x_D)^T$ = Variabel iput
$w = (w_0, w_1, \dots, w_D)^T$ = Parameter bobot
$\Phi(x)$ = Fungsi transformasi fitur

K-nearest neighbors (KNN)

KNN adalah metode yang menggunakan klasifikasi tetangga (*neighbors*) sebagai prediktor untuk *instance query* yang baru. Proses klasifikasi pada metode ini akan terjadi didasarkan pada kesamaan antara data yang baru dan data sampel yang meningkat (Andiani, Sukemi, & Rini, 2020). Jarak antara data terbaru dan data yang sudah lama dihitung dengan rumus sebagai berikut :

$$d = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (6)$$

Keterangan:

$X1$ = Data sampel
$X2$ = Data *training* atau *testing*
i = Variabel data
d = Jarak
p = Dimensi data

***Random forest* (RF)**

RF adalah metode untuk mengklasifikasikan sejumlah besar data. Klasifikasi random forest dilakukan dengan menggabungkan pohon dengan pelatihan pada data sampel. Akar adalah *node* yang berada di bagian atas pohon keputusan atau biasa disebut dengan *root*. *Internal node* adalah cabang yang memiliki setidaknya dua pintu keluar dan satu pintu masuk. *Leaf node* atau simpul terminal adalah simpul terakhir dengan hanya satu pintu masuk dan tidak ada jalan keluar. Cara menghitung *entropy* untuk menentukan tingkat polusi atribut dan nilai perolehan informasi adalah dengan dimulainya pohon keputusan (S. Rahayu et al., 2020). Untuk menghitung *entropy* digunakan rumus berikut :

$$Entropy\ (Y) = -\sum_i p(c|Y) \log^2 p(c|Y) \quad (7)$$

Keterangan:

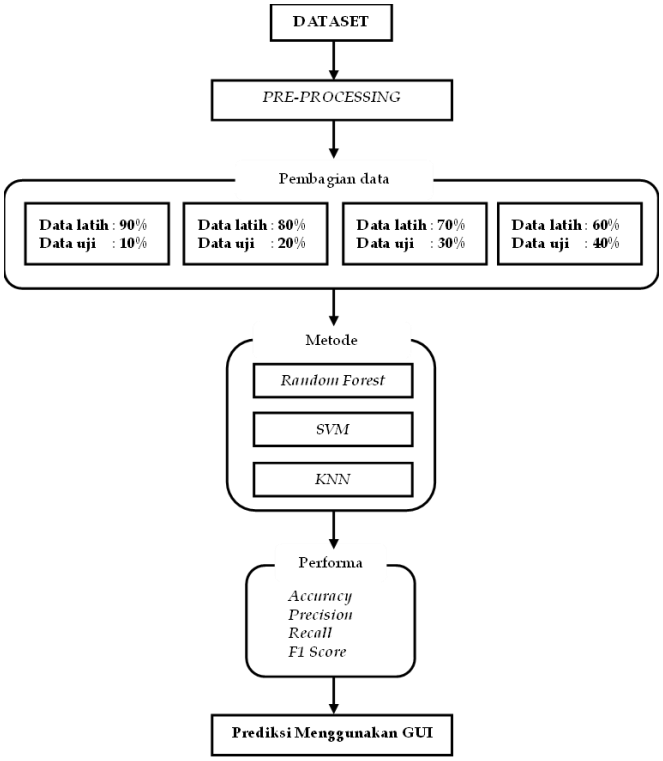
Y = Himpunan kasus
$P(c|Y)$ = Proporsi nilai Y terhadap kelas c

$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{v \varepsilon Values(a)} \frac{|Y_v|}{|Y_a|} Entropy(Y_v) \quad (8)$$

Keterangan:

$Values(a)$ = Nilai yang terdapat di dalam himpunan kelas a
Y_v = Subkelas dari Y dengan kelas v yang berelasi dengan kelas a
Y_a = Nilai-nilai yang mempunyai kesesuaian dengan kelas a

Alur penelitian ini menggunakan tiga metode yang dapat dilihat pada Gambar 1. berikut :



Gambar 1. Alur penelitian

Gambar 1. menjelaskan bahwa dataset akan melalui tahap *pre-processing* terlebih dahulu untuk normalisasi data agar data yang digunakan pada ketiga metode tersebut tidak memiliki penyimpangan yang besar. Tahap *pre-processing* ini dilakukan dengan menggunakan *library MinMaxScaler*. Setelah itu dataset akan dibagi menurut data latih dan data uji, kemudian diproses menggunakan tiga metode yaitu SVM, KNN, dan *random forest*. Proses klasifikasi menggunakan ketiga metode tersebut akan menghasilkan empat nilai dari *confusion matrix* yaitu *accuracy, precision, recall, F1 score*. Tahap terakhir pada penelitian ini, peneliti juga membuat program GUI untuk memudahkan prediksi pasien mengidap penyakit gagal jantung atau tidak. Program GUI yang dibuat menggunakan ketiga metode dengan hasil akurasi yang terbaik dan waktu *running* program yang paling singkat.

HASIL DAN PEMBAHASAN

Kumpulan data yang digunakan dalam penilitian ini dipublikasikan pada tahun 2020. Dataset tersebut diambil dari laman *UCI Machine Learning* dengan judul *Heart Failure Clinical Records Dataset* (https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records). Dataset terdiri dari 299 data pasien penyakit gagal jantung dan masing-masing

data memiliki 13 atribut sebagaimana dideskripsikan pada Tabel 2. berikut.

Tabel 2. Artibut dalam penelitian

No	Atribut	Keterangan	Tipe data	Contoh data
1	Age	Umur dari pasien penyakit gagal jantung	*Integer*	55 (tahun)
2	Anaemia	Menurunnya hemoglobin atau sel darah merah dalam tubuh	*Boolean*	1 = Ya 0 = Tidak
3	High Blood Pressure	Hipertensi	*Boolean*	1 = Ya 0 = Tidak
4	Creatinine Phosphokinase (CPK)	Enzim CPK dalam darah	*Integer*	7861 (mcg/L)
5	Diabetes	Pasien menderita diabetes atau tidak	*Boolean*	1 = Ya 0 = Tidak
6	Ejection Fraction	Volume darah yang mengalir meninggalkan jantung setiap jantung berkontraksi	*Integer*	38 (%)
7	Platelets	Jumlah trombosit dalam tubuh	*Double*	263358.03 (kiloplatelets /mL)
8	Sex	Jenis Kelamin	*Boolean*	1 = Laki-laki 0 = Perempuan
9	Serum Creatinine	Jumlah kreatinin serum yang terdapat pada darah	*Double*	1.1 (mg/dL)
10	Serum Sodium	Jumlah natrium serum yang terdapat pada darah	*Integer*	136 (mEq/L)
11	Smoking	Perokok atau tidak perokok	*Boolean*	1 = Ya 0 = Tidak
12	Time	Periode tindak lanjut	*Integer*	6 (hari)
13	[Target] Death Event	Pasien yang telah meninggal dalam masa tindak lanjut	*Boolean*	1 = Gagal jantung 0 = Tidak gagal jantung

Data yang telah terkumpul kemudian dilakukan tahap *preprocessing* untuk mengubah data-data tersebut menjadi data yang lebih normal dan tidak memiliki penyimpangan yang besar, dalam artian semua data akan memiliki skala dan rentang nilai yang sama. Pada penilitian ini, peniliti menggunakan *library MinMaxScaler* untuk melakukan *preprocessing* karena *library* tersebut bekerja dengan cara menyesuaikan nilai data dalam rentang tertentu yaitu antara 0 sampai 1 (Shaheen, Agarwal, & Ranjan, 2020). Berikut adalah persamaan dari *library MinMaxScaler* :

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (9)$$

Pada Gambar 2. peneliti juga menampilkan hasil dari tahap *preprocessing* menggunakan *library MinMaxScaler* :

```
#Preprocessing
scaler = MinMaxScaler().fit(X)
standardized_data = scaler.transform(X)
X = standardized_data
print(X)
```

```
[[0.63636364 0.         0.07131921 ... 1.         0.         0.        ]
 [0.27272727 0.         1.         ... 1.         0.         0.00711744]
 [0.45454545 0.         0.01569278 ... 1.         1.         0.01067616]
 ...
 [0.09090909 0.         0.25988773 ... 0.         0.         0.97508897]
 [0.09090909 0.         0.30492473 ... 1.         1.         0.98220641]
 [0.18181818 0.         0.02207196 ... 1.         1.         1.        ]]
```

Gambar 2. Hasil *preprocessing* dataset

Dapat diperhatikan pada Gambar 2. menjelaskan bahwa *library MinMaxScaler* tersebut bekerja dengan cara menyesuaikan data dengan rentang antara 0 sampai 1.

Kelas dari hasil klasifikasi dataset terdiri dua kelas, 1 untuk gagal jantung dan 0 untuk tidak gagal jantung. Pengujian yang akan dilakukan didasarkan pada distribusi data pelatihan (*training*) dan pengujian (*testing*) menggunakan prosedur pembagian dataset sebagai berikut.

A. *Training* 90% dan *testing* 10%

Tabel 3. Hasil uji coba *training* 90% dan *testing* 10%

Metode	*Precision*	*Recall*	*F1 Score*	*Accuracy*
SVM	90%	98%	93%	97%
RF	90%	98%	93%	97%
KNN	96%	75%	81%	93%

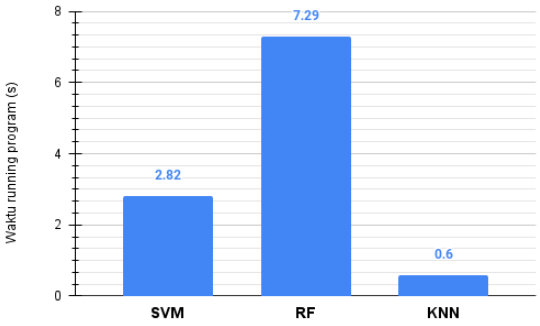
Berdasarkan Tabel 3, akurasi terbaik dihasilkan oleh metode SVM dan RF dengan nilai yang sama yaitu sebesar 97%. Berikut parameter-parameter yang diperoleh pada setiap metode :

- *Support vector machine* (SVM)
  - $C = 1$
  - $gamma = 0.01$
  - $kernel = linear$
- *Random forest* (RF)
  - $n\_estimators = 30$

- $random\_state = 0$

- *K-nearest neighbors* (KNN)
  - $n\_neighbors = 20$

Pada Gambar 3. peneliti juga merepresentasikan waktu *running* program melalui grafik diagram batang.



Gambar 3. Hasil waktu *running* program

Sehingga dapat disimpulkan bahwa untuk hasil uji coba *training* 90% dan *testing* 10%, metode terbaik menurut akurasi dan waktu *running* program yang paling singkat adalah memakai metode SVM.

B. *Training* 80% dan *testing* 20%

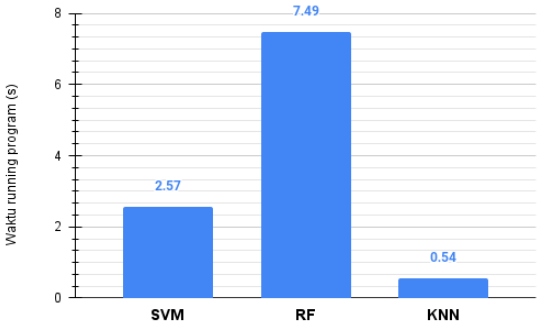
Tabel 4. Hasil uji coba *training* 80% dan *testing* 20%

Metode	*Precision*	*Recall*	*F1 Score*	*Accuracy*
SVM	86%	80%	82%	88%
RF	90%	81%	84%	90%
KNN	91%	64%	67%	83%

Pada Tabel 4. terlihat bahwa akurasi yang terbaik dihasilkan oleh metode RF dengan nilai sebesar 90%. Parameter-parameter yang diperoleh setiap metode :

- *Support vector machine* (SVM)
  - $C = 100$
  - $gamma = 0.01$
  - $kernel = rbf$
- *Random forest* (RF)
  - $n\_estimators = 30$
  - $random\_state = 0$
- *K-nearest neighbors* (KNN)
  - $n\_neighbors = 10$

Berikut adalah waktu *running* program dalam diagram batang.



Gambar 4. Hasil waktu *running* program

Jadi kesimpulannya adalah bahwa untuk hasil uji coba *training* 80% dan *testing* 20%, metode yang memiliki akurasi tinggi adalah metode RF tetapi waktu *running* program metode RF lebih lama dibandingkan waktu *running* program metode SVM dan KNN.

C. *Training* 70% dan *testing* 30%

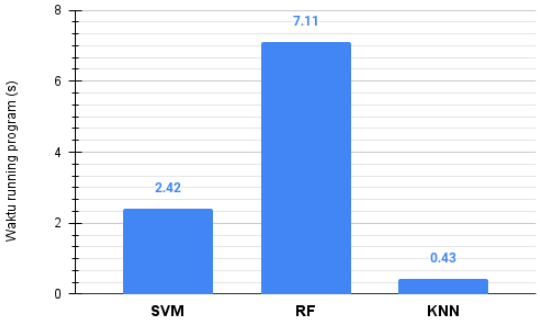
Tabel 5. Hasil uji coba *training* 70% dan *testing* 30%

Metode	*Precision*	*Recall*	*F1 Score*	*Accuracy*
SVM	85%	81%	82%	87%
RF	90%	84%	87%	90%
KNN	80%	59%	59%	77%

Tabel 5. diatas menjelaskan bahwa akurasi terbaik dihasilkan oleh metode RF dengan nilai sebesar 90%. Berikut parameter-parameter yang diperoleh pada setiap metode :

- *Support vector machine* (SVM)
  - $C = 100$
  - $gamma = 0.01$
  - $kernel = rbf$
- *Random forest* (RF)
  - $n\_estimators = 90$
  - $random\_state = 0$
- *K-nearest neighbors* (KNN)
  - $n\_neighbors = 10$

Gambar 5. merupakan representasi waktu *running* program yang dihasilkan dalam bentuk grafik diagram batang.



Gambar 5. Hasil *confusion matrix* terhadap tiga metode

Hasil dari Tabel 5. dan Gambar 5. yang telah dilakukan oleh peneliti, dapat diambil kesimpulan bahwa untuk hasil uji coba pada data *training* 70% dan *testing* 30%, metode yang memiliki akurasi paling tinggi adalah metode RF, tetapi waktu *running* program yang dibutuhkan oleh metode RF lebih lama dibandingkan waktu *running* program metode SVM dan KNN

D. *Training* 60% dan *testing* 40%

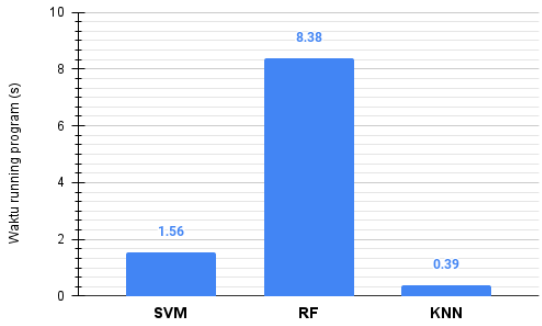
Tabel 6. Hasil uji coba *training* 60% dan *testing* 40%

Metode	*Precision*	*Recall*	*F1 Score*	*Accuracy*
SVM	88%	84%	86%	88%
RF	89%	85%	87%	89%
KNN	73%	53%	49%	71%

Dari Tabel 6. dapat dijelaskan bahwa akurasi terbaik dihasilkan oleh metode RF dengan nilai sebesar 89%. Berikut parameter yang diperoleh pada setiap metode :

- *Support vector machine* (SVM)
  - $C = 100$
  - $gamma = 0.01$
  - $kernel = rbf$
- *Random forest* (RF)
  - $n\_estimators = 40$
  - $random\_state = 0$
- *K-nearest neighbors* (KNN)
  - $n\_neighbors = 10$

Berikut peneliti juga merepresentasikan waktu *running* program melalui grafik diagram batang.



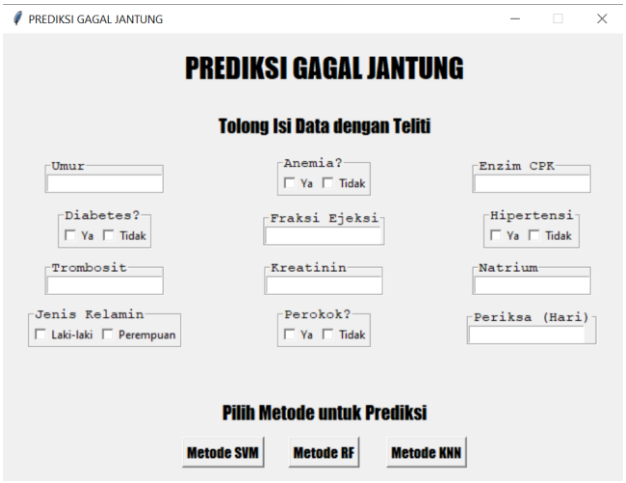
Gambar 6. Hasil *confusion matrix* terhadap tiga metode

Sehingga dapat diambil kesimpulan bahwa untuk hasil uji coba *training* 60% dan *testing* 40%, metode yang memiliki akurasi tertinggi adalah metode RF tetapi waktu *running* program metode RF lebih lama dibandingkan waktu *running* program metode SVM dan KNN.

Pembagian data latih (*training*) dan data uji (*testing*) dilakukan oleh peneliti pada bagian A (*training* 90% dan *testing* 10%), B (*training* 80% dan *testing* 20%), C (*training* 70% dan *testing* 30%), dan D (*training* 60% dan *testing* 40%). Dari semua uji coba, dapat diambil kesimpulan bahwa akurasi yang terbaik dihasilkan dari pembagian data pada bagian A yaitu *training* 90% dan *testing* 10%, dimana menghasilkan akurasi sebesar 97% untuk metode SVM dan RF, sedangkan untuk metode KNN menghasilkan akurasi sebesar 93%. Dapat dilihat hasil akurasi pada setiap pembagian data, diketahui bahwa semakin besar data yang diuji maka akurasi pada setiap metode akan semakin menurun sehingga peneliti tidak membagi dataset dengan rincian data uji lebih besar dari 40% karena data uji tidak digunakan untuk melatih dataset yang diberikan. Jika data uji yang digunakan lebih besar daripada data latih, maka tidak akan diketahui *outcome* dari dataset tersebut dan menyebabkan akurasi semakin menurun atau *out-of-sample testing* (Afifah, 2020). Untuk mendapatkan hasil klasifikasi yang baik, maka data latih harus lebih besar daripada data uji karena semakin banyak data yang dilatih maka model yang dibentuk juga akan semakin bagus dan setelah proses pelatihan akan dilanjutkan pengujian untuk menentukan keakurasian model (Nugroho, 2020).

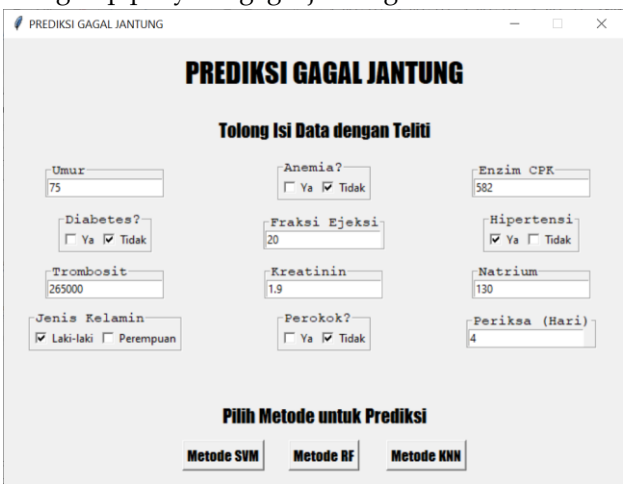
Pada tahap terakhir, peneliti juga membuat *Graphical User Interface* (GUI) untuk memudahkan jika terdapat *user* yang ingin mencoba program yang dibuat peneliti untuk memprediksi pasien mengidap penyakit gagal jantung atau tidak mengidap penyakit gagal jantung. Program ini dibuat menggunakan parameter-parameter terbaik yang telah diperoleh dari hasil uji coba pembagian data. Peneliti menggunakan pembagian data *training* 90% dan *testing* 10% dengan parameter dari metode SVM adalah $C = 1$, $gamma = 0.01$, dan $kernel = linear$. Parameter dari metode KNN adalah $n\_neighbors = 30$. Dan parameter dari metode RF adalah $n\_estimators = 30$ dan $random\_state = 0$.

Berikut adalah tampilan dari GUI yang dibuat oleh peneliti.



Gambar 7. Tampilan awal GUI

Gambar 7. adalah tampilan awal dari GUI yang dibuat peneliti. Pada tampilan awal ini, *user* akan diminta untuk mengisi data-data secara lengkap. Setelah data-data yang diisikan oleh *user* sudah lengkap, maka *user* dapat memilih salah satu metode yang ingin digunakan untuk memprediksi apakah mengidap penyakit gagal jantung atau tidak.



Gambar 8. Tampilan *input* data dari *user*

Gambar 8. memperlihatkan *user* telah mengisi data-data yang diperlukan untuk prediksi secara lengkap. Dalam Gambar 8, data yang diisikan adalah salah satu data dari *Heart Failure Clinical Records Dataset* yang diambil dari laman *UCI Machine Learning*.

Gambar 9. Tampilan deteksi tidak gagal jantung

Gambar 9. adalah contoh tampilan jika *user* memilih metode SVM dan hasil deteksi pasien tidak mengidap penyakit gagal jantung.

Gambar 10. Tampilan deteksi gagal jantung

Gambar 10. adalah contoh tampilan jika *user* memilih metode SVM dan hasil deteksi pasien mengidap penyakit gagal jantung.

PENUTUP

SIMPULAN

Pada penelitian ini, peneliti menggunakan dataset dari *UCI Machine Learning* dengan judul *Heart Failure Clinical Records Dataset*. Tujuan penelitian ini adalah untuk mengomparasi akurasi dari metode *support vector mahince* (SVM), *random forest,* dan *k-nearest neighbors* (KNN). Dalam mengomparasi ketiga metode tersebut, peneliti juga mengomparasi pembagian data latih dan data uji dengan rincian pembagiannya adalah 90% : 10%, 80% : 20%, 70% : 30%, 60% : 40% sehingga menghasilkan *confusion matrix* yang berguna untuk melihat *precision, recall, F1 score, accuracy* yang dihasilkan.

Hasil terbaik menurut akurasi dan waktu yang dibutuhkan untuk *running* program, diperoleh metode SVM dan RF dengan pembagian data *training* 90% dan *testing* 10%, dimana akurasi dari kedua metode tersebut adalah 97% dan *running* program yang dibutuhkan untuk metode SVM adalah 2.82 detik sedangkan untuk metode RF adalah 7.29 detik. Parameter yang digunakan oleh metode SVM adalah $C = 1, gamma = 0.01, kernel = linear$. Sedangkan parameter yang digunakan oleh metode RF adalah $n\_estimators = 30$ dan $random\_state = 0$.

Jika dibandingkan dengan penelitian yang sudah pernah dilakukan sebelumnya, didapatkan prediksi akurasi yang dihasilkan dalam penelitian ini adalah meningkat dalam artian prediksi akurasi dalam penelitian ini lebih baik daripada penelitian sebelumnya. Menurut penelitian yang dilakukan oleh (Trianifa, 2019), hasil prediksi akurasi menggunakan metode SVM adalah sebesar 95%, menurut penelitian yang dilakukan oleh (S. Rahayu

et al., 2020) hasil prediksi akurasi menggunakan metode RF adalah sebesar 94.31%, dan menurut penelitian yang dilakukan oleh (Newaz et al., 2021) hasil prediksi akurasi menggunakan metode RF adalah sebesar 76.25%. Sedangkan hasil prediksi akurasi yang dilakukan oleh peneliti menggunakan metode SVM dan RF adalah sebesar 97%. Sehingga dapat disimpulkan juga bahwa penelitian yang dilakukan oleh peneliti ini merupakan penelitian dengan akurasi lebih baik.

DAFTAR PUSTAKA

Rozie, F., Hadary, F., & Wigyarianto, F. T. P. (2016). Rancang Bangun Alat Monitoring Jumlah Denyut Nadi / Jantung Berbasis Android. Jurnal Teknik Elektro Universitas Tanjungpura, 1(1), 1–10.

Andiani, L., Sukemi, S., & Rini, D. P. (2020, February). Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest. In Annual Research Seminar (ARS) (Vol. 5, No. 1, pp. 165-169).

Putra, P. D., & Rini, D. P. (2020, February). Prediksi Penyakit Jantung dengan Algoritma Klasifikasi. In Annual Research Seminar (ARS) (Vol. 5, No. 1, pp. 95-99).

Trianifa, N., Arifin, A., & Rini Novitasari, D. (2019). KLASIFIKASI PENYAKIT JANTUNG MENGGUNAKAN METODE SUPPORT VECTOR MACHINE BERDASARKAN PERBANDINGAN ALGORITMA PEMBACAAN WAKTU DENGAN TEKSTUR SINYAL SEBAGAI METODE EKSTRAKSI SINYAL EKG. MathVisioN, 2(1), 7-11. Retrieved from http://journal.unirow.ac.id/index.php/mv/article/view/136

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access, 9, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., Sun, R., & Garciá-Magarinõ, I. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018. https://doi.org/10.1155/2018/3860146

Annisa Hapsari (2021). Penyakit Jantung Koroner: Gejala, Penyebab, dan Pengobatan. Diakses pada 03 Maret 2022. https://hellosehat.com/jantung/jantung-koroner/pengertian-jantung-koroner/

dr. Gabriella Florencia (2019). Ketahui Fungsi Jantung Ini pada Tubuh Manusia. Diakses pada 04 Matet 2022. https://www.halodoc.com/artikel/ketahui-fungsi-jantung-ini-pada-tubuh-manusia

Sarang Narkhede (2018). Understanding Confusion Matrix. Diakses pada 04 Maret 2022. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Chicco, D., Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16. https://doi.org/10.1186/s12911-020-1023-5

Shaheen, H., Agarwal, S., & Ranjan, P. (2020). MinMaxScaler binary PSO for feature selection. In First international conference on sustainable technologies for computational intelligence (pp. 705-716). Springer, Singapore.

S. Rahayu, J. J. Purnama, A. B. Pohan, F. S. Nugraha, S. Nurdiani, and S. Hadianti. (2020). "Prediction of survival of heart failure patients using random forest," J. Pilar Nusa Mandiri, vol. 16, no. 2, pp. 255–260.

N. M. Lutimath, C. Chethan and B. S. Pol. (2019). "Prediction of heart disease using machine learning," International Journal of Recent Technology and Engineering, vol. 8, no. 2 Special Issue 10, pp. 474-477, 1 9.

A. Rohman and J. Banjarsari Barat No. (2016). "Komporasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Jantung,".

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. Proceedings -IEEE Symposium on Computers and Communications, Iscc, 204–207. https://doi.org/10.1109/ISCC.2017.8024530

A. B. Wibisono and A. Fahrurozi. (2019). " Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," Jurnal Ilmiah Teknologi dan Rekayasa, vol. 24, no. 3, pp. 161-170.

Kramar, Vadym & Alchakov, Vasiliy & Dushko, Veronika & Kramar, Tatiana. (2018). Application of support vector machine for prediction and classification. Journal of Physics: Conference Series. 1015. 032070. 10.1088/1742-6596/1015/3/032070.

A. Kaur. (2018). " Heart Disease Prediction Using Data Mining Techniques: A SURVEY,"

International Journal of Advanced Research in Computer Science, vol. 9, no. 2, pp. 569-572, 20 2.

A. Riani, Y. Susianto and N. Rahman. (2018). "Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes," Journal of Innovation Information Technology and Application (JINITA), vol. 1, no. 01, pp. 25- 34, 26 12.

Purbianto, P., & Agustanti, D. (2017). Analisis Faktor Risiko Gagal Jantung di RSUD dr. H. Abdul Moeloek Provinsi Lampung. Jurnal Ilmiah Keperawatan Sai Betik, 11(2), 194-203.

Agustina, A., Afiyanti, Y., & Ilmi, B. (2017). Pengalaman pasien gagal jantung kongestif dalam melaksanakan perawatan mandiri. Healthy-Mu Journal, 1(1), 6-14.

Azis, H., Purnawansyah, P., Fattah, F., Putri, I. P., Ilmiah, I. J., Bustami, B., & Mengklasifikasi, P. A. N. B. U. A. Somasundaram and US Reddy. (2016).—Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data, Proc. 1st Int. Conf. Res. Eng. Comput. Technol. (ICRECT 2016), no. November, pp. 28–34. Alfisahrin, SNN, & Mantoro, T. (2014). Datamining techniques for. Science, 1(2), 29-33.

Nugroho, K. S. (2020). Validasi Model Klasifikasi Machine Learning pada RapidMiner. Diakses pada 17 Maret 2022. https://ksnugroho.medium.com/validasi-model-machine-learning-pada-rapidminer-50be0080df14

Afifah, L. (2020). Evaluasi Model Machine Learning: Train/Test Split. Diakses pada 17 Maret 2022. https://ilmudatapy.com/evaluasi-model-machine-learning-dengan-train-test-split/

Asif, N., Nadim, A., Farhan, S. H. (2021). Survival prediction of heart failure patients using machine learning technique. ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100772. (https://www.sciencedirect.com/science/article/pii/S2352914821002458)