JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 8
No. 2
Agustus
2022

Reduksi Dimensi Data menggunakan Metode *Wrapper Sequential Feature Selection* untuk Peningkatan Performa Algoritma *Naïve Bayes* terhadap Dataset Medis

Mochammad Yusa[#1], Funny Farady Coastera[#2], Muhammad Randa Yandika[#3]

[#]*Program Studi Informatika, Fakultas Teknik, Universitas Bengkulu*
Jl. W.R. Supratman Kandang Limun, Kec. Muara Bangka Hulu, Sumatera, Bengkulu 38371
[1]mochammad.yusa@unib.ac.id
[2]ffaradyc@unib.ac.id
[3]randayandika1@gmail.com

Abstrak— **Penggunaan Machine Learning sebagai alat bantu dalam penanganan medis saat ini berkembang dengan pesat. Salah satu penyakit medis yang dikembangkan menggukan algoritma komputasi adalah *Cardiovascullar Disease* (CVD). *Machine learning* model yang diterapkan didasarkan dataset rekam medis. Tujuan penelitian ini adalah menginvestigasi performa algoritma *naïve bayes* dengan menerpakan metode *Wrapper Sequential Feature Selection* (WSFS). Metode penelitian dimulai dari pengumpulan dataset, data preprocessing, penerapan model *Naïve Bayes*, dan atribut scoring menggunakan Wrapper SFS, dan validasi performa menggunakan uji validasi *10-Fold Cross-Validation*. Data history yang digunakan yaitu dataset *Heart Failure Clinical Records* yang terdiri dari 299 *instances* pada 13 *features*. Hasil penelitian menunjukkan bahwa metode Wrapper SFS dapat mengimprovisasi nilai performa Algoritma *Naïve Bayes* dari nilai akurasi, Precisi, dan Recall. Adapun kenaikan performa didapatkan dengan kombinasi 6 fitur ('anaemia', 'diabetes', 'ejection_fraction', 'serum_creatinine', 'gender', 'time') yang didapatkan dari seleksi fitur WSFS terhadap Algoritma tersebut yaitu nilai akurasi meningkat sebanyak 6,334%, skor *recall* meningkat 11,333%, dan nilai *precision* meningkat sebesar 20,07% dibandingkan dengan Algoritma *Naïve Bayes*.**

Kata kunci— **Penyakit Kardiovaskuler, *Machine Learning, Wrapper Sequential Feature Selection, Naïve Bayes***

I. PENDAHULUAN

Membangun sistem bantuan untuk kesehatan diperlukan suatu sistem cerdas tentunya membutuhkan keakuratan *machine learning* yang tinggi. Pendekatan Probabilitas *Naïve Bayes* merupakan salah satu model *machine learning* yang memiliki performa akurasi yang tinggi[1]. Namun dalam beberapa kasus penggalian data dengan dimensi yang besar, tingkat performa akurasi *Naïve Bayes* belum terlalu tinggi [2], [3]. Sehingga dibutuhkan improvisasi untuk meningkatkan akurasi dalam pendekatan *Naïve Bayes.*

Salah satu cara untuk meningkatkan akurasi algoritma *Naïve Bayes* adalah mereduksi dimensi data [4], [5]. *Dimensionality reduction* merupakan salah satu pendekatan dalam tahap *data preprocessing* yang dapat meningkatkan akurasi metode klasifikasi dalam data mining [6], [7]. Salah satu pendekatan reduksi dimensi adalah pendekatan dengan *feature selection*. *Feature selection* adalah proses pemilihan subset variabel input yang relevan untuk digunakan dalam konstruksi model dari kumpulan data besar [6]. Pemilihan fitur (*Feature Selector*) dapat mengurangi fitur yang tidak relevan dan redundan, sehingga membutuhkan lebih sedikit waktu untuk melatih model dan dapat membantu meningkatkan performa pengklasifikasi yang dihasilkan[8].

Salah satu cara melakukan seleksi fitur adalah dengan menggunakan metode *Wrapper*. Dalam metode *wrapper* evaluasi dilakukan pada subset variabel yang menjadi indicator dalam pemilihan feature sehingga tidak seperti penggunaan metode *filter* yang melihat dari segi kemungkinan komunikasi antara variabel[9]. Metode *wrapper* melakukan pemilihan fitur dengan cara memperhatikan algoritma pembelajaran yang akan digunakan. Keuntungan utama dibandingkan metode filter adalah ia menemukan fitur yang paling berpengaruh dan melakukan pemilihan fitur yang optimal untuk algoritma pembelajaran mesin[10].

Dalam studi ini dataset penyakit kardiovaskular atau *cardiovascular disease* (CVD) menjadi focus data dalam proses reduksi dimensi. CVD adalah gangguan pada jantung dan pembuluh darah termasuk, penyakit jantung koroner (serangan jantung), penyakit serebrovaskular (stroke), gagal jantung, dan jenis patologi lainnya [11]. Penyakit kardiovaskular menyebabkan kematian sekitar 18,6 juta orang di seluruh dunia, dengan tingkat resiko 1 dari 5 orang berusia <70 tahun meninggal, serta persentase kematian akibat penyakit kardiovaskular mencapai 32%.

Submitted 21-04-2022; Revised 01-08-2022; Accepted 22-08-2022

Gagal jantung (*cardiovascular*) terjadi ketika jantung tidak dapat memompa cukup darah ke tubuh, dan biasanya disebabkan oleh diabetes, tekanan darah tinggi, atau kondisi atau penyakit jantung lainnya. Mengingat pentingnya organ vital seperti jantung, memprediksi gagal jantung telah menjadi prioritas bagi dokter dan dokter medis, tetapi hingga saat ini peramalan kejadian terkait gagal jantung dalam praktik klinis biasanya gagal mencapai akurasi yang tinggi. Selain itu, dataset pada penyakit ini memiliki atribut yang banyak. Berdasarkan penelitian [12] terdapat 13 atribut yang dipertimbangkan menjadi parameter untuk mengukur tingkat keberlangsungan hidup pasien pengidap gagal jantung yaitu Umur, Anemia, Tekanan darah tinggi, Enzim CPK, Diabetes, Fraksi ejeksi , Jenis Kelamin, Trombosit, Serum kreatinin, Serum Natrium, Habit Merokok, Waktu tindak lanjut. Berdasarkan hal tersebut, penelitian ini bertujuan untuk menerapkan *wrapper feature selection* sebagai parameter eliminasi atribut yang akan digunakan dan diuji menggunakan algoritma *Naïve Bayes*.

II. PENELITIAN TERKAIT

Pada bidang medis, klasifikasi berbasis *machine learning* telah banyak digunakan untuk membantu dokter dan ahli kesehatan dalam diagnosis penyakit maupun penentuan tindakan perawatan dan pengobatan, konsep umum klasifikasi adalah mengelompokkan atau mengkategorikan sesuatu berdasarkan atribut atau fitur yang ada.

Chicco & Jurman (2020) [11] melakukan evaluasi terhadap penggunaan beberapa algoritma, salah satunya yaitu *naïve bayes*. Penelitian tersebut hanya menghasilkan akurasi sebesar 69,6% menggunakan metode *naïve bayes*. Namun dengan akurasi tersebut lebih rendah dibandingkan penelitian dengan menggunakan metode *naïve bayes* lainnya seperti penelitian yang dilakukan oleh Nugraha et al. (2017) [13] melakukan klasifikasi terhadap data pasien dengan penyakit stoke pada RSUD Undata Palu sebanyak 203 data dengan menggunakan algoritma *naïve bayes*, menghasilkan tingkat akurasi sebesar 89,65% dalam proses pengklasifikasian data.

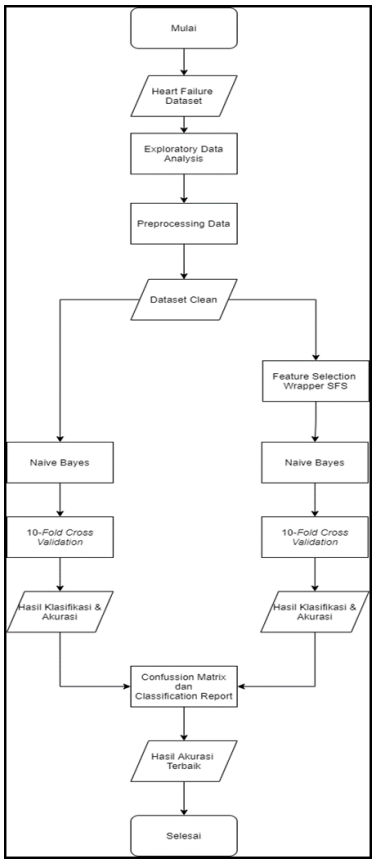
Penelitian lain yang dilakukan oleh Derisma (2020) [14] melakukan perbandingan kinerja algoritma *machine learning* terhadap *Heart Disease Dataset* yaitu dengan menggunakan algoritma *naïve bayes*, random forest dan neural network untuk memprediksi orang dengan penyakit jantung. Perbandingan yang dihasilkan menunjukkan algoritma *naïve bayes* akurat dan presisi dalam memprediksi orang yang punya penyakit jantung dengan akurasi dihasilkan sebesar 83%. Dengan demikian, diperlukan sebuah proses untuk meningkatkan nilai akurasi dari klasifikasi tingkat keselamatan pasien dengan gagal jantung yang dilakukan. Dengan melakukan proses mengurangi jumlah fitur yang ada pada dataset dengan metode *feature selection* berbasis *wrapper*, diharapkan dalam melakukan klasifikasi dengan metode *naïve bayes* dapat menghasilkan proses yang lebih efektif dan hasil akurasi yang lebih tinggi dari pada sebelumnya.

Hairani & Innuddin (2020) [15] melakukan pengujian klasifikasi data kesehatan dengan menggunakan Kombinasi Metode *Correlated Naive Bayes* dan Metode Seleksi Fitur *Wrapper*. Dengan menggunakan kombinasi kedua metode tersebut menghasilkan kenaikan akurasi pada dataset Pima Indan Diabetes sebesar 4,1%, dari 67,3% menjadi 71,4% dan kenaikan akurasi pada dataset Thyroid sebesar 0,48% dari 78,9% menjadi 79,38%. Penelitian yang dilakukan oleh [16] melakukan reduksi atribut Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0. Hasil yang didapatkan dengan tingkat akurasi tertinggi ketika dilakukan klasifikasi dengan 11 kombinasi atribut dimana terdapat 1 atribut yang direduksi, kenaikan akurasi yang didapatkan sebesar 5,61% dari 83,5% menjadi 89,11%.

Berdasarkan hal tersebut maka penelitian yang akan dilakukan ini bertujuan untuk mengetahui peningkatan hasil performa akurasi menggunakan kombinasi antara *Wrapper Sequential Feature selection (WSFS)* dan *Naïve Bayes* pada dataset medis khususnya dataset penyakit Cardio Vascular.

III. METODE PENELITIAN

Penelitian ini merupakan jenis penelitian eksperimen. Metode ini digunakan atas dasar pertimbangan bahwa sifat penelitian eksperimental yaitu mencobakan sesuatu untuk mengetahui pengaruh atau akibat dari suatu perlakuan atau treatment. Gambar 1 merupakan tahapan dari desain penelitian, yaitu:



Gambar 1 Desain penelitian

A. *Data Collecting*

Jenis data yang akan digunakan pada penelitian ini adalah data sekunder. Dataset yang digunakan merupakan *public dataset* yang akan diambil dari *UCI Machine Learning Repository.*

B. *Exploratory Data Analysis*

EDA digunakan dalam tujuan pengurangan dimensi data atau memperkaya pemahaman atas analisis data melalui visualisasi data [17]. EDA juga digunakan diantaranya untuk mengoptimalkan pengetahuan mengenai data, menghasilkan variabel yang penting, mendeteksi outlier dan anomali pada data, dan menguji asumsi awal [18].

Eksplorasi data yang dilakukan pada tahapan ini seperti melihat deskripsi terhadap dataset, melihat tipe data pada setiap fitur, menampilkan index value, menampilkan informasi terhadap dataset, serta menampilkan korelasi antar data apakah sebuah data dapat memberikan pengaruh terhadap data features lainnya.

C. *Preprocessing Data*

Dalam tahap ini akan dilakukan beberapa proses untuk mendapatkan dataset yang bersih (clean dataset). Adapun langkah-langkah tersebut adalah *handling missing value, konversi tipe data, duplicated data analysis*, dan *data transformation*.

Handling Missing Value

Tahapan preprocessing data dengan melakukan analisis data hilang atau missing value. Sehingga didapat bahwa setiap fitur pada dataset memiliki jumlah missing value 0 atau tidak terdapat data hilang.

1) Konversi tipe data: Tahapan preprocessing data untuk mengubah tipe data sesuai dengan kondisi aslinya.

2) Cek data duplikat: Tahapan untuk mengecek apakah terdapat data ganda atau duplikat pada dataset yang digunakan. Fungsi yang digunakan adalah *duplicated()* dan fungsi *sum()* untuk menampilkan jumlah data duplikat jika ada.

3) Data transformation: Tahapan untuk melakukan perubahan nama pada fitur dilakukan agar sesuai dengan pembahasan paper sebelumnya dan sesuai dengan kondisi asli yaitu pada fitur *platelets.* Setelah semua sub proses dalam *data preprocessing* dilakukan, *outcome* yang diharapkan adalah dataset yang bersih (*clean dataset*) yang nantinya akan digunakan untuk uji validasi model.

D. *Naïve Bayes Tanpa Proses Feature Selection*

Pada eksperimen pertama yang dilakukan yaitu mengimplementasikan metode *naïve bayes* untuk dataset clean yang telah didapatkan, tanpa melalui proses *feature selection.*

Kemudian pada tahapan ini dilakukan pengujian dengan menggunakan metode *10-fold cross validation* dengan membagi dataset menjadi 10 bagian dan akan membentuk 10 iterasi dengan *data test* dan *data train* yang berbeda. Lalu *output* dari pengujian yang dilakukan adalah hasil akurasi atau scoring lain yang ditentukan dengan rata-rata dari 10 iterasi yang dilakukan.

E. *Naïve Bayes + Feature selection Wrapper*

Pada eksperimen kedua, proses yang dilakukan adalah melakukan *feature selection wrapper* terhadap *dataset clean* yang telah didapatkan. Untuk proses mencari fitur-fitur dalam *wrapper* sendiri akan menggunakan *search method Sequential Feature Selector (SFS).*

Untuk metode SFS, dimulai dengan mencari sebuah fitur dengan performa terbaik. Tahap selanjutnya yaitu mengkombinasikan fitur yang didapat sebelumnya dengan fitur lain yang memiliki performa terbaik dibandingkan dengan fitur lain. Proses itu terus diulangi hingga membentuk kombinasi-kombinasi fitur.

Hasil setiap kombinasi dari fitur dari proses *W*SFS selanjutnya akan dilakukan tahapan klasifikasi menggunakan metode *Naïve Bayes* lalu dilakukan pengujian model menggunakan metode *K-Fold Cross Validation* dengan k yang ditentukan yaitu nilai k=10.

Semua kemungkinan kombinasi dari fitur-fitur yang digunakan akan melalui tahapan-tahapan yang disebutkan. Setiap kemungkinan tersebut nantinya akan menghasilkan nilai akurasi yang berbeda-beda menggunakan metode *Naïve Bayes* yang diuji dengan *K-Fold Cross Validation*.

F. *Evaluasi*

Selanjutnya hasil dari akurasi tersebut dilakukan proses evaluasi yang dilakukan untuk mengukur kinerja dari algoritma yang kita gunakan dengan menggunakan *Confusion Matrix* dan menampilkan nilai-nilai pada *Classification Report* seperti nilai *Precission, Recall & F1-Score*.

Setelah melakukan semua tahapan eksperimen, dapat dilihat mana hasil pengujian yang memiliki akurasi terbaik. Eksperimen *Naïve bayes* tanpa menggunakan *Feature selection Wrapper* atau dengan menggunakan *Wrapper Sequential Feature Selector (WSFS).*

Dalam penelitian ini juga akan dilakukan analisis fitur-fitur mana saja yang memiliki korelasi tinggi dan berpengaruh terhadap label yang ada pada dataset serta menghasilkan akurasi tertinggi dari klasifikasi *naïve bayes* yang divalidasi.

IV. HASIL DAN PEMBAHASAN

A. *Data Collecting*

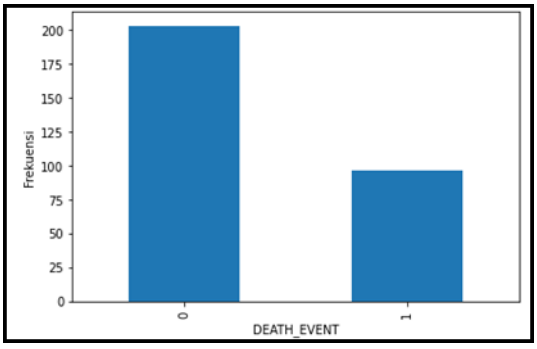
Penelitian ini menggunakan dataset yang berasal dari *UCI Repository Machine learning*. Dataset berisi 13 fitur, yang melaporkan klinis, informasi tubuh, dan gaya hidup seperti anemia, tinggi tekanan darah, diabetes, jenis kelamin, dan merokok. Penjelasan lengkap mengenai seperti fitur, deskripsi, ukuran dan *range* data yang ada pada dataset dapat dilihat pada Tabel 1.

TABEL I
DESKRIPSI DATASET

Atribut Index	Fitur	Penjelasan	Ukuran	*Range*
1	**Umur**	Umur Pasien	Tahun	(40, ..., 95)
2	**Anemia**	Penurunan sel darah merah	Boolean	0, 1 0=False, 1=True
3	**Tekanan darah tinggi**	Jika pasien menderita hipertensi	Boolean	0, 1 0=False, 1=True
4	**Enzim CPK**	Tingkat enzim CPK dalam darah	mcg/L	(23, ..., 7861)
5	**Diabetes**	Jika pasien menderita diabetes	Boolean	0, 1 0=False, 1=True
6	**Fraksi ejeksi**	Persentase keluarnya darah jantung di setiap kontraksi	Persentase	(14, ..., 80)
7	**Jenis Kelamin**	Pria/Wanita	Binary	0= F, 1=M
8	**Trombosit**	Trombosit di darah	kiloplatelets / mL	(25.01, ..., 850.00)
9	**Serum kreatinin**	Tingkat kreatinin dalam darah	mg/dL	(0.50, ..., 9.40)
10	**Serum Natrium**	Tingkat natrium dalam darah	mEq/L	114, ..., 148
11	**Merokok**	Jika Pasien Merokok	Boolean	0, 1 0=False, 1=True
12	**Waktu**	Periode tindak lanjut	Hari	(4, ..., 285)
Target	**(TARGET) Bertahannya pasien Survive / Not Survive**	Jika pasien tidak bertahan selama masa tindak lanjut	Boolean	0, 1 0=Survived, 1=Not Survived

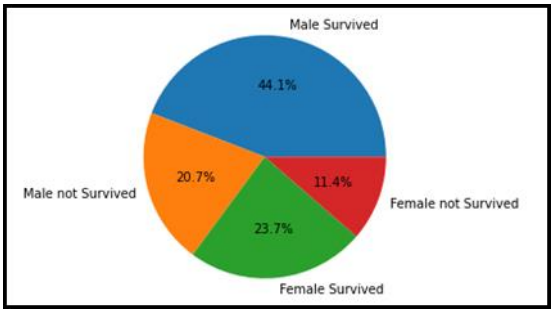
A. Exploratory Data Analysis

Pada eksplorasi data yang dilakukan, akan melihat deskripsi detail dari dataset yang digunakan seperti, melihat tipe data dari masing-masing fitur, melihat *index value*, melihat distribusi untuk setiap fitur, serta melihat korelasi data pada fitur yang ada pada dataset.



Gambar 2. *Plot bar* jumlah data

Gambar 2 menampilkan plot bar dari jumlah frekuensi pasien DEATH_EVENT yang statusnya 0=*Survived* dan 1=*Not Survived*. Dari data yang ada, total ada 203 pasien yang Survived dan ada 96 pasien yang *Not Survived*.



Gambar 3. *Pie char explanatory data*

Gambar 3 merepresentasikan persentase dari feature DEATH_EVENT berdasarkan jenis kelamin dimana laki-laki *survived* memiliki persentase sebesar 44,1%, laki-laki *not survived* memiliki persentase sebesar 20,7%, perempuan *survived* memiliki persentase sebesar 23,7%, perempuan *not survived* memiliki persentase sebesar 11,4%.

B. Preprocessing Data

1) Handling Missing Value: Tahapan *Handling Missing Value* yang dilakukan, didapatkan bahwa setiap fitur pada dataset memiliki jumlah missing value 0 atau tidak terdapat data hilang. Pada proses ini, tidak terdapat *missing value* didalam dataset.

2) Konversi Tipe Data: Konversi tipe data yang dilakukan pada tahapan ini adalah untuk fitur-fitur seperti anemia, diabetes, high_blood_ pleasure, smoking, gender dari tipe data integer menjadi boolean.

3) Cek Data Duplikat: Tahapan pengecekan data duplikasi yang dilakukan, tidak terdapat data duplikat. Pada proses ini, tidak terdapat data duplikat didalam dataset.

4) Mengubah Value (data transformation) Pada Fitur Platelets dan merubah nama pada fitur Sex: Tahapan untuk melakukan perubahan nama pada fitur sex menjadi gender dengan menggunakan fungsi rename dan merubah value pada fitur platelets menjadi kiloplatelets/mL dengan membagi nilai pada fitur tersebut dengan 1000.

C. Wrapper Sequential Feature selection (WSFS)

Sequential Feature Selecton (Forward) dimulai dengan melatih model *machine learning* untuk setiap fitur dalam kumpulan data dan memilih sebagai fitur awal, fitur yang mengembalikan model berperforma terbaik, menurut kriteria evaluasi tertentu akan dipilih. Selanjutnya

membuat model *machine learning* untuk semua kombinasi fitur yang dipilih pada langkah sebelumnya dan fitur kedua. Ini memilih kombinasi fitur yang menghasilkan algoritma berkinerja terbaik. Proses berlanjut dengan menambahkan 1 fitur untuk setiap iterasi/proses dari fitur yang telah dipilih sebelumnya pada langkah sebelumnya, hingga proses berhenti pada parameter yang telah ditentukan sebelumnya.

Setiap iterasi pada proses *Wrapper Sequential Feature Selection* (WSFS), akan menghasilkan kombinasi fitur-fitur terbaik pada setiap iterasinya. Hasil dari setiap iterasi yang dilakukan dapat dilihat pada tabel 2.

TABEL II
HASIL WRAPPER SFS

Iterasi Ke-	Atribut Index	Atribut *Scoring* terhadap target
1	(11,)	0,820402
2	(4, 11)	0,828448
3	(4, 7, 11)	0,858046
4	(3, 4, 7, 11)	0,862213
5	(3, 4, 7, 9, 11)	0,849569
6	(1, 3, 4, 7, 9, 11)	0,849713
7	(1, 3, 4, 7, 8, 9, 11)	0,849713
8	(1, 3, 4, 6, 7, 8, 9, 11)	0,845546
9	(0, 1, 3, 4, 6, 7, 8, 9, 11)	0,832902
10	(0, 1, 2, 3, 4, 6, 7, 8, 9, 11)	0,829023
11	(0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11)	0,816236
12	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	0,832902

Tabel 2 menunjukkan hasil dari proses *Feature Selection Wrapper* dengan menggunakan *Sequential Feature Selector (SFS)* untuk tiap iterasi. Tiap iterasi akan menghasilkan sebuah fitur terbaik, fitur yang didapatkan akan dikombinasikan dengan fitur yang didapatkan pada tahapan sebelumnya. Kombinasi yang didapatkan untuk tiap iterasi merupakan hasil terbaik antara 1 fitur dengan fitur yang lain dibandingkan fitur lainnya di dalam dataset.

Kemudian setiap kombinasi dari fitur yang ddapatkan akan dilakukan implementasi dengan menggunakan metode *naïve bayes* dan pengujian *10-Fold Cross Validation* untuk melihat hasil klasifikasi dan peforma dengan menggunakan model tersebut.

D. Hasil Pengujian Model

Model *naïve bayes* tanpa *feature selection*. Pada proses ini pengujian dataset dengan melakukan *feature selection wrapper* dan melakukan klasifikasi *naïve bayes* dengan kombinasi atribut yang berbeda–beda akan menghasilkan tingkat akurasi yang berbeda pula. Tingkat akurasi pada dataset tanpa melakukan proses *Feature Selection Wrapper* sebesar 75,241%. Hasil dari proses klasifikasi *Naïve Bayes* tanpa menggunakan *feature selection wrapper* dapat dilihat pada Tabel 3.

TABEL IIII
HASIL PERFORMA DARI KLASIFIKASI *NAÏVE BAYES*

Naïve Bayes Tanpa Feature selection Wrapper			
Jumlah Fitur	***Accuracy***	***Recall***	***Precision***
12 Fitur	75,241	43,556	67,662

Proses klasifikasi dengan metode *naïve bayes* berhasil mengklasifikasikan label pasien dengan gagal jantung dengan perbandingan yaitu untuk kelas *surivived* sebesar 203 pasien dari data asli dan 235 untuk hasil prediksinya. Sedangkan untuk kelas *not survived* sebesar 96 pasien dari data asli dan 64 pasien untuk hasil prediksinya. Dari perbandingan tersebut antara data asli dan hasil prediksi, didapatlah nilai *scoring* dari klasifikasi yang dilakukan. Hasil yang ditampilkan pada Tabel 4 menunjukkan bahwa proses klasifikasi menggunakan *naïve bayes* dan pengujian dengan *10-fold cross validation* menghasilkan akurasi sebesar 75,241%, recall sebesar 43,556%, dan presisi sebesar 67,662%.

TABEL IIIV
HASIL DARI PROSES *FEATURE SELECTION WRAPPER & NAÏVE BAYES*

Jumlah Fitur	***WSFS***		
	Accuracy	***Recall***	***Precision***
3 Fitur	81,241	53,889	**87,787**
4 Fitur	80.908	52.778	87.827
5 Fitur	80,575	51,889	87,454
6 Fitur	**81.575**	**54.889**	87.727
7 Fitur	77.230	46.778	77.475
8 Fitur	77,563	47,889	77,216
9 Fitur	78,908	52	83,209
10 Fitur	75,575	43,667	67,643
11 Fitur	75,575	42,667	68,337

Tabel 4 mengindikasikan hasil akhir dari proses *Feature selection Wrapper* dan *Naïve Bayes.* Hasil eksperimen yang dilakukan menunjukkan bahwa, dari setiap fitur-fitur yang digunakan akan mempengaruhi hasil akhir dari proses klasifikasi keselamatan pasien dengan gagal jantung dan juga dari setiap fitur akan menghasilkan nilai-nilai *scoring* yang berbeda. Pengujian dilakukan dari model dengan 3 fitur dimana hal tersebut merujuk pada penelitian sebelumnya yang dilakukan oleh Chicco & Jurman (2020) [11] dimana dapat melakukan prediksi hanya dengan menggunakan fitur *ejection fraction* dan *serum creatinine.* Pada penelitian ini kedua fitur tersebut tidak boleh dihapus, kombinasi yang terdapat kedua metode tersebut yaitu kombinasi 3 sampai 11 fitur.

Setelah pengujian model didapat dengan menggunakan kombinasi 3 fitur sampai kombinasi 11 fitur yang telah dilakukan implementasi menggunakan metode *naïve bayes*. Hasil klasifikasi dengan metode *naïve bayes* mengalami perubahan pada setiap nilai *Scoring* setelah dilakukan proses *Feature Selection Wrapper*. Dimana dengan menggunakan kombinasi metode tersebut lalu dilakukan pengujian dengan menggunakan *10-fold cross validation,* menghasilkan nilai akurasi tertinggi yang didapat sebesar 81.575% dari scoring 6 atribut. Nilai *recall* tertinggi sebesar yang didapat sebesar 54.889 sedangkan performa *precision* tertinggi yang didapat sebesar 87,787% yang diapatkan dari kombinasi 3 feature. Dari hasil evaluasi performa tersebut didapatlah model dengan menggunakan kombinasi 6 fitur yang menghasilkan nilai akurasi tertinggi

dibandingkan dengan kombinasinya. Fitur-fitur yang didapatkan dengan menggunakan *Wrapper Sequential Feature Selection* (WSFS), yaitu *('anaemia', 'diabetes', 'ejection_fraction', 'serum_creatinine', 'gender', 'time'),* Model dengan menggunakan kombinasi 6 fitur ini, menghasilkan performa dengan nilai akurasi sebesar 81.575%. Nilai *Recall* sebesar 54,889%. Nilai *Precision* sebesar 87,787%.

E. Pengaruh metode Feature selection Wrapper terhadap performa Naïve Bayes

Berdasarkan hasil yang didapat pada proses yang telah dijalankan, metode *Feature Selection Wrapper* dapat meningkatkan mempengaruhi dari hasil pengujian yang dilakukan terhadap model seperti pada Tabel 3 dan 4 yang menampilkan perbandingan hasil akurasi. *Feature selection Wrapper* memilih fitur-fitur mana saja yang penting terhadap target/label pada dataset dan dapat memilih fitur-fitur yang tidak perlu digunakan dalam proses klasifikasi menggunakan *naïve bayes*.

Dari 12 fitur yang ada, didapatlah 9 kombinasi dari proses SFS. Kombinasi dengan menggunakan 6 fitur dapat memberikan hasil akurasi paling tinggi dibandingkan dengan kombinasi lainnya. Kombinasi dengan 6 fitur tersebut yaitu fitur *('anaemia', 'diabetes', 'ejection_fraction', 'serum_creatinine', 'gender', 'time').* Hasil tersebut sesuai dengan paper penelitian rujukan sebelumnya yang dilakukan oleh Chicco & Jurman (2020) [11] bahwa *machine learning* dapat memprediksi keselamatan pasien dengan gagal jantung dari *ejection_fraction* dan *serum_creatinine*. Dimana dari 6 fitur yang didapatkan pada proses *Feature Selection Wrapper*, fitur *ejection_fraction* dan *serum_creatinine* terdapat didalam 6 fitur tersebut. Itu membuktikan hasil penelitian ini sesuai dengan penelitian sebelumnya.

Jika dibandingkan dengan kasus pada dataset asli dan hasil prediksi dari model yang dibuat, untuk kelas Survived terdapat 203 pasien dan untuk prediksi dari model dengan kombinasi 6 fitur yang memiliki akurasi tertinggi terdapat 232 pasien. Sedangkan untuk kelas not survived, pada kasus dataset asli terdapat 96 pasien dan untuk prediksi dengan kombinasi 6 fitur yang memiliki akurasi tertinggi terdapat 67 pasien.

Dengan menggunakan *WSFS*, hasil akurasi dari klasifikasi *Naïve Bayes* dengan menggunakan kombinasi 6 fitur meningkat menjadi 81,575 % dari nilai akurasi menggunakan algoritma *Naïve Bayes* yaitu 75,241% atau terjadi peningkatan akurasi sebesar 6,334%. Peningkatan tersebut diikuti dengan nilai peroforma recal dan precision. Untuk performa *recall*, menggunakan WSFS dan *Naïve Bayes* terjadi peningkatan sebesar 11,333% yaitu dari 43,556% menjadi 54,889% sedangkan nilai *precision*-nya meningkat sebanyak 20,07% dari 67,662% menjadi 87,727%. Berdasarkan hasil *experiment* tersebut, kombinasi metode *Wrapper Sequential Feature Selection* (WSFS) dan *naïve bayes* membuktikan terjadinya peningkatan performa baik akurasi, *recall*, dan *precision* dalam dataset medis khususnya penyakit gagal jantung atau cardiovascular disease (CDC).

V. KESIMPULAN

Penelitian ini telah menguji kefektivitasan *Wrapper Sequential Feature Selection* (WSFS) yang dikombinasikan dengan algoritma klasifikasi *Naïve Bayes*. Berdasarkan dari hasil eksperimen pengujian dan analisa dari penelitian yang telah dilakukan, metode WSFS dapat memberikan rekomendasi fitur yang dapat digunakan untuk meningkatkan performa algoritma klasifikasi setelah diuji dengan menggunakan metode validasi *10-Fold Cross validation*. Dari dataset medis yang dipilih sebagai data uji didapatkan bahwa kombinasi 6 (enam) fitur yaitu *'anemia', 'diabetes', 'ejection_fraction', 'serum_creatinine', 'gender', 'time'* dari total 11 (sebelas) fitur dapat meningkatkan performa algoritma klasifikasi *Naïve Bayes* dengan nilai akurasi sebesar 81.575%. Nilai *Recall* sebesar 54,889%. Nilai *Precision* sebesar 87,787%. Jika dibandingkan dengan hanya menerapkan algoritma klasifikasi *Naïve Bayes*, nilai performa yang dihasilkan hanya 75,241% untuk performa akurasi, recall hanya 43,556% dan Precision yaitu 67,662%. Dari hasil tersebut dapat diketahui bahwa terjadi peningkatan sebesar 6,334%. untuk akurasi, 11,333% untuk nilai recall dan 20,07% untuk nilai performa precision-nya. Hal tersebut membuktikan bahwa metode seleksi fitur WSFS dapat secara signifikan meningkatkan performa algoritma klasifikasi khususnya *Naïve Bayes* dalam kasus yang diangkat ini.

REFERENSI

[1] S. M. Gorade, A. Deo, and P. Purohit, "A Study Some Data Mining Classification Techniques," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 1, pp. 210–215, 2017, doi: 10.21884/ijmter.2017.4031.zt9tv.

[2] P. Golpour *et al.*, "Comparison of support vector machine, naïve bayes and logistic regression for assessing the necessity for coronary angiography," *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, pp. 1–9, 2020, doi: 10.3390/ijerph17186449.

[3] M. Yusa and E. Utami, "Classifiers evaluation: Comparison of performance classifiers based on tuples amount," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, vol. 4, doi: 10.11591/eecsi.4.1086.

[4] P. R. Anukrishna and V. Paul, "A review on feature selection for high dimensional data," *Proc. Int. Conf. Inven. Syst. Control. ICISC 2017*, vol. 5, no. 6, pp. 395–402, 2017, doi: 10.1109/ICISC.2017.8068746.

[5] O. Saini and S. Sharma, "A Review on Dimension Reduction Techniques in Data Mining," *Comput. Eng. Intell. Syst.*, vol. 9, no. 1, pp. 7–14, 2018.

[6] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: a review," *AIMS Bioeng.*, vol. 4, no. 2, pp. 179–197, 2017, doi: 10.3934/bioeng.2017.2.179.

[7] G. Chao, Y. Luo, and W. Ding, "Recent Advances in Supervised Dimension Reduction: A Survey," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 1, pp. 341–358, 2019, doi: 10.3390/make1010020.

[8] G. Kicska and A. Kiss, "Comparing swarm intelligence algorithms for dimension reduction in machine learning," *Big Data Cogn. Comput.*, vol. 5, no. 3, 2021, doi: 10.3390/bdcc5030036.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic

data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013, doi: 10.1007/s10115-012-0487-8.

[10] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[11] 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, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[12] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 345–351, 2020, doi: 10.1109/ICDABI53623.2021.9655783.

[13] D. W. Nugraha, A. Y. E. Dodu, and N. Chandra, "Klasifikasi Penyakit Stroke Menggunakan Metode Naive Bayes Classifier (Studi Kasus Pada Rumah Sakit Umum Daerah Undata Palu)," *semanTIK*, vol. 3, no. 2, pp. 13–22, 2017.

[14] D. Derisma, "Perbandingan Kinerja Algoritma untuk Prediksi Penyakit Jantung dengan Teknik Data Mining," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 84–88, 2020, doi: 10.30871/jaic.v4i1.2152.

[15] H. Hairani and M. Innuddin, "Kombinasi Metode Correlated Naive Bayes dan Metode Seleksi Fitur Wrapper untuk Klasifikasi Data Kesehatan," *J. Tek. Elektro*, vol. 11, no. 2, pp. 50–55, 2020, doi: 10.15294/jte.v11i2.23693.

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

[17] M. Abukmeil, S. Ferrari, A. Genovese, V. Piuri, and F. Scotti, "A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning," *ACM Comput. Surv.*, vol. 54, no. 5, 2021, doi: 10.1145/3450963.

[18] S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," *J. Med. Virol.*, vol. 92, no. 6, pp. 632–638, 2020, doi: 10.1002/jmv.25743.