Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2023107467
Vol. 10, No. 6, Desember 2023, hlm. 1361-1368 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

KOMBINASI SELEKSI FITUR BERBASIS FILTER DAN WRAPPER MENGGUNAKAN NAIVE BAYES PADA KLASIFIKASI PENYAKIT JANTUNG

Siti Roziana Azizah[1], Rudy Herteno[*2], Andi Farmadi[3], Dwi Kartini[4], Irwan Budiman[5]

[1,2,3,4,5]Universitas Lambung Mangkurat, Banjarmasin
Email: [1]azizahroziana@gmail.com, [2]rudy.herteno@ulm.ac.id, [3]andifarmadi@gmail.com, [4]dwikartini@ulm.ac.id, [5]irwan.budiman@ulm.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 12 Juli 2023, diterima untuk diterbitkan: 27 November 2023)

Abstrak

Penyakit jantung menjadi salah satu penyebab utama kematian bersama dengan penyakit lainnya. Dalam bidang teknologi, data mining dapat digunakan untuk mendiagnosa suatu penyakit yang bersumber dari data rekam medis pasien. Pada klasifikasi dataset medis *Naive Bayes* merupakan salah satu metode terbaik yang digunakan, namun algoritma *Naive Bayes* memiliki kekurangan yaitu mengasumsikan bahwa tiap fitur dalam data tidak memiliki hubungan satu sama lain. Berdasarkan kekurangan dari algoritma Naive Bayes tersebut ditambahkan seleksi fitur untuk memilih fitur apa saja yang penting dalam data. Tujuan dari penelitian ini adalah untuk mengetahui perbandingan hasil akurasi dari *Naive Bayes* menggunakan beberapa seleksi fitur yaitu *Forward Selection, Backward Elimination,* kombinasi *union* hasil seleksi fitur *Forwad Selection* dan *Backward Elimination, Information Gain, Gain Ratio,* dan kombinasi *union* hasil seleksi fitur *Information Gain* dengan *Gain Ratio.* Data yang digunakan dalam penelitian ini adalah data penyakit jantung yang didapatkan dari *UCI Machine Learning Repository*. Dari implementasi pemodelan yang akan dilakukan menghasilkan nilai akurasi tertinggi sebesar 91.80% pada algoritma *Naive Bayes* dengan kombinasi *union* hasil seleksi fitur *Information Gain* dan *Gain Ratio* menggunakan perbandingan data latih dan data uji 80:20. Sedangkan akurasi *Naive Bayes* dengan kombinasi *union* hasil seleksi fitur *Forward Selection* dan *Backward Elimination* hanya memiliki nilai akurasi sebesar 83.61%.

Kata kunci: Penyakit Jantung, Klasifikasi, *Naive Bayes*, Seleksi Fitur, Kombinasi *Union*

COMBINATIONS OF FEATURE SELECTION BASED ON FILTER AND WRAPPER USING NAIVE BAYES IN HEART DISEASE CLASSIFICATION

Abstract

Heart disease is one of the leading causes of death along with other diseases. In the field of technology, data mining can be used to diagnose a disease sourced from patient medical record data. In the classification of medical datasets Naive Bayes is one of the best methods used, but the Naive Bayes algorithm has the disadvantage that it assumes that each feature in the data has no relationship with each other. Based on the shortcomings of the Naive Bayes algorithm, feature selection is added to select what features are important in the data. The purpose of this study is to compare the accuracy results of Naive Bayes using several feature selections, namely Forward Selection, Backward Elimination, a combination of union of Forwad Selection and Backward Elimination feature selection results, Information Gain, Gain Ratio, and a combination of union of Information Gain feature selection results with Gain Ratio. The data used in this research is heart disease data obtained from the UCI Machine Learning Repository. From the implementation of modeling that will be carried out, the highest accuracy value is 91.80% in the Naive Bayes algorithm with a combination of union of Information Gain and Gain Ratio feature selection results using a ratio of training data and test data of 80:20. While the accuracy of Naive Bayes with a combination of union selection results of Forward Selection and Backward Elimination features only has an accuracy value of 83.61%.

Keywords: *Heart Disease, Classification, Naive Bayes, Feature Selection, Union Combination*

1. PENDAHULUAN

Penyakit jantung menjadi salah satu penyebab utama kematian bersama dengan penyakit lainnya seperti stroke, kanker paru-paru, kanker payudara, dan AIDS (Nugroho, 2018). Pada survei *Sample Registration System* (SRS) tahun 2014 menjelaskan

bahwa 12,9% yang menjadi pelaku utama kematian pada semua umur di Indonesia adalah penyakit jantung (Ryfai, Hidayat and Santoso, 2022). Dalam melakukan diagnosa penyakit jantung secara medis harus melibatkan pihak yang ahli dalam bidangnya. Namun dalam bidang teknologi, data mining dapat digunakan untuk mendiagnosa suatu penyakit yang bersumber dari data rekam medis pasien, salah satunya penyakit jantung (Prasetyo and Prasetiyo, 2020)

Dalam bidang ilmu komputer, data mining dikenal sebagai metode untuk menggali data dengan tujuan menemukan suatu pola yang tersembunyi agar dapat menghasilkan suatu pengetahuan yang baru. Berdasarkan tujuan dan pemanfaatannya, data mining memiliki metodenya sendiri yaitu prediksi, estimasi, asosiasi, klastering, dan klasifikasi (Tarigan et al., 2022). Klasifikasi merupakan metode data mining dengan cara mengelompokkan data atas dasar keterikatan dengan data sampel (Oktanisa and Supianto, 2018).

Naive Bayes merupakan metode prediksi berdasarkan probabilitas yang sederhana dengan menerapkan teorema Bayes serta mengasumsikan bahwa fitur dari suatu data tidak ada sangkut paut dengan fitur lain pada data yang sama (Suprianto, 2020). Metode Klasifikasi Naive Bayes memiliki kekurangan dalam memilih fitur yang terdapat dalam data, sehingga mempengaruhi nilai akurasinya (Arifin and Ariesta, 2019). Untuk menangani permasalahan tersebut, dapat dilakukan teknik seleksi fitur untuk mengetahui fitur yang penting dalam dalam data.

Penggunaan seleksi fitur bertujuan untuk mengurangi kompleksitas fitur pada data yang akan dilakukan *processing* serta analisa (Adnyana, 2019).Terdapat dua jenis teknik seleksi fitur, yaitu *filter* dan *wrapper.* Teknik *wrapper* merupakan teknik yang tergabung dalam algoritma pembelajaran dalam melakukan evaluasi subset fitur (Suchetha, Nikhil and Hrudya, 2019). Berdasarkan penelitian (Nurlia and Enri, 2021) membuktikan bahwa seleksi fitur *Forward Elimination* mampu meningkatkan hasil akurasi pada algoritma C4.5 dari 77,89% menjadi 84,29%. Selain itu, pada penelitian (Amilia, Oktavianto and Abdurrahman, 2021) menunjukkan hasil akurasi pada klasfikasi penyakit jantung meningkat saat menggunakan seleksi fitur *Backward Elimination* dari 94,56% menjadi 98,33%.

Seleksi fitur *filter* menggunakan korelasi, ukuran konsistensi, ukuran jarak, dan lainnya untuk menentukan peringkat fitur (Suchetha, Nikhil and Hrudya, 2019). Seperti pada penelitian (Aini, Sari and Arwan, 2018) melakukan pengujian dari kombinasi algoritma KNN dengan *Naive Bayes* menggunakan seleksi fitur *Information Gain* menghasilkan nilai akurasi tertinggi sebesar 92,31% pada data seimbang dengan 6 atribut menggunakan nilai K=35 dan pada saat pengujian sebaran kelas tidak seimbang menggunakan 4 fitur dengan nilai K=35.

Berdasarkan beberapa penelitian terdahulu, seleksi fitur berbasis *forward selection* dan *backward elimination* terbukti dapat meningkatkan nilai akurasi dari suatu algoritma. Selain menggunakan seleksi fitur yang sudah dipaparkan sebelumnya, dalam memilih fitur yang berpengaruh dari suatu data menggunakan seleksi fitur dapat dilakukan dengan mengkombinasikan dua atau lebih seleksi fitur dengan beberapa cara. Seperti penelitian yang dilakukan oleh (Putri, Nugroho and Herteno, 2021) menggunakan kombinasi dua seleksi fitur *filter* yaitu *Information Gain Ratio* dan *Correlation.* Pada penelitian tersebut, algoritma yang digunakan *K-Nearest Neighbor* dengan hasil akurasi tertinggi saat menggunakan dua kombinasi seleksi fitur yaitu 99,61% dibandingkan hasil akurasi tanpa seleksi fitur sebesar 99,59%.

Berdasarkan penjelasan latar belakang di atas, maka peneliti mengusulkan sebuah gagasan penelitian yaitu menggunakan kombinasi pada seleksi fitur *Information Gain* dan *Gain Ratio* untuk kategori *filter*, sedangkan pada kategori *wrapper* menggunakan kombinasi *Forward Selection* dan *Backward Elimination* untuk meningkatkan hasil akurasi *Naive Bayes* pada klasifikasi penyakit jantung.

2. METODE PENELITIAN

Tahapan penelitian secara singkat digambarkan pada gambar 1 dibawah ini.

2.1 Pengumpulan *Dataset*

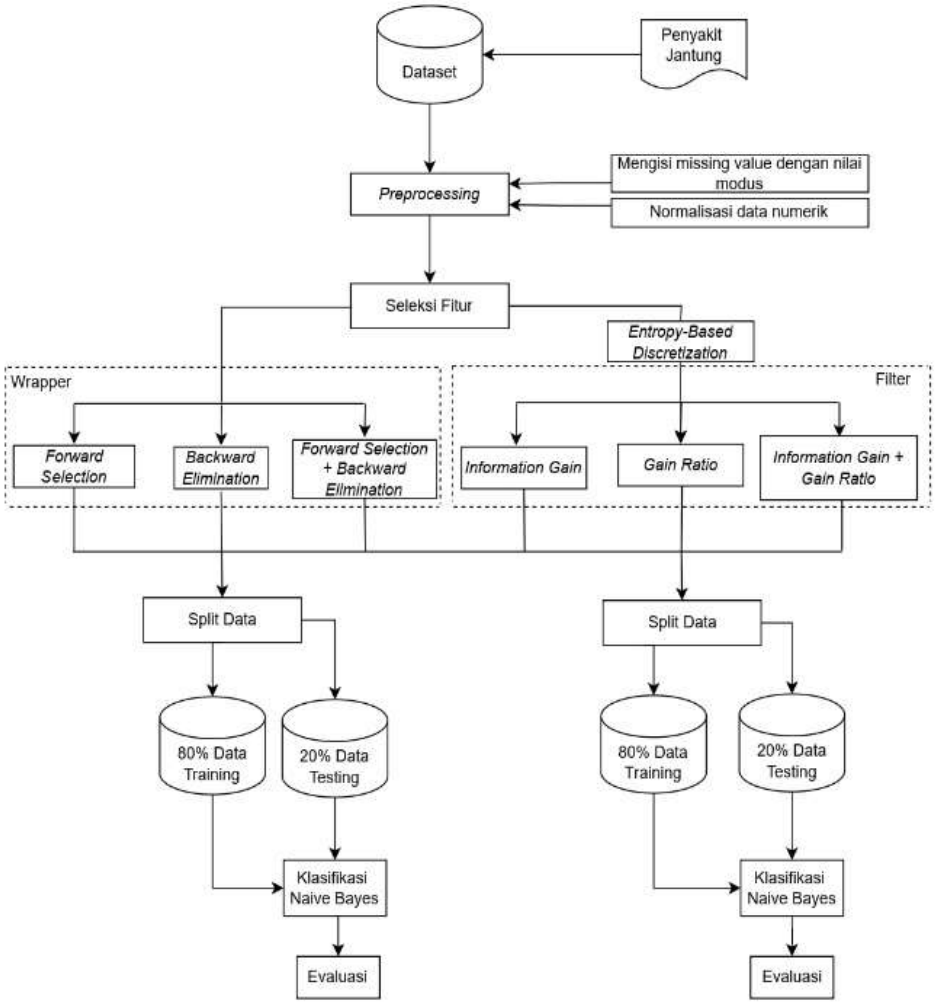
Pada penelitian ini data yang digunakan adalah data penyakit jantung dari UCI *Machine Learning Repository* yang terdiri dari 303 *record* dengan 13 atribut independen dan 1 atribut dependen. Untuk atribut dependen memiliki dua nilai yaitu 1 untuk terdiagnosis penyakit jantung, dan 0 tidak terdiagnosis penyakit jantung.

2.2 *Pre-processing* Data

Terdapat 2 tahap *preprocessing* pada penelitian ini, yaitu *missing value handling* dan normalisasi data.

2.2.1 *Missing Value*

Missing value dalam data adalah kondisi tidak tersedianya suatu informasi dalam suatu data. *Missing value* dalam data dapat mempengaruhi hasil penelitian karena *missing value* dalam data dapat mengurangi tingkat akurasi dari hasil penelitian (Martha & Sulistianingsih, 2018). *Missing value handling* pada penelitian ini dengan cara mengisi nilai kosong dengan nilai modus dari atributnya.



Gambar 1. Alur Penelitian

2.2.2 *Min-Max Scaler*

Normalisasi data merupakan salah satu bagian dari *preprocessing* dalam mesin pembelajaran. Alasan perlu dilakukannya normalisasi data pada penelitian ini adalah rentang nilai dari atribut-atribut yang terdapat dalam data berbeda-beda, sehingga perlu dilakukannya normalisasi data. Teknik normalisasi data mengasumsikan bahwa semua atribut berapa pada rentang yang sama, biasanya antara 0 hingga 1. Terdapat beberapa teknik normalisasi data, salah satunya adalah teknik *Min-Max Scaler*. Normalisasi data menggunakan *min-max scaler* memiliki kelebihan, yaitu membutuhkan waktu yang lebih singkat dalam mencapai konvergensinya daripada teknik normalisasi yang lain. Seperti pada penelitian yang dilakukan oleh (Suryanegara, Adiwijaya and Purbolaksono, 2021) normalisasi menggunakan *min-max scaler* memiliki nilai akurasi tertinggi dibandingkan penggunaan teknik normalisasi data lainnya. Rumus dari *Min-Max scaler* terdapat pada formula 1.

$$X_i' = \frac{X_i - \min(X)}{\max(X) - \min(X)} \quad (1)$$

2.3 Seleksi Fitur

Seleksi fitur merupakan sebuah istilah yang biasa digunakan dalam *machine learning* dan statistik. Seleksi fitur merupakan metode untuk melakukan pemilihan subset fitur yang relevan sebelum pembangunan model. Pada data yang memiliki dimensi tinggi, fitur yang tidak relevan atau berlebihan akan dihapus (Hameed et al., 2018).

2.3.1 Filter

Pada tahap selanjutnya dilakukan pemodelan Naïve Bayes menggunakan seleksi fitur *filter.* Seleksi fitur filter yang digunakan adalah *Information Gain* (IG), *Gain Ratio* (GR), dan kombinasi keduanya.

Konsep kerja dari seleksi fitur *filter* adalah menggunakan metode statistika untuk menilai atributnya, kemudian setiap atribut akan diurutkan sesuai tingkatannya serta dibuat matriks guna mengetahui atribut yang relevan, sedangkan atribut yang tidak relevan akan dieliminasi (Sasongko and Arifin, 2019).

IG adalah suatu fitur filter yang populer. Seleksi fitur melakukan prosesnya dengan cara meranking atribut yang dianggap berpengaruh besar pada kelasnya. Penggunaan seleksi fitur ini akan

membantu mendapatkan atribut yang relevan pada kelas target (Nur, Ahsan and Harianto, 2022).

GR adalah metode seleksi fitur modifikasi dari metode *information gain*. Seleksi fitur *gain ratio* memodifikasi dengan cara memperbaiki *information gain* dengan pengambilan informasi intrinsik pada atribut (Nur, Ahsan and Harianto, 2022).

Selain dua seleksi fitur IG dan GR, penelitian ini juga menggunakan kombinasi dari dua seleksi fitur tersebut. Teknik kombinasi yang akan digunakan pada penelitian ini adalah kombinasi *union*.

2.3.2 Wrapper

Seleksi fitur *Wrapper* yang digunakan pada penelitian ini adalah *Forward Selection* (FS), *Backward Elimination (BE)*, serta kombinasi keduanya.

Seleksi fitur *Wrapper* menjalankan prosesnya dengan melakukan pencarian dari kumpulan atribut yang ada, kemudian dibandingkan satu sama lainnya. Dalam seleksi fitur *Wrapper*, digunakan model algoritma yang bertujuan melakukan evaluasi terhadap kumpulan kombinasi atribut yang ada (Sasongko and Arifin, 2019).

FS adalah seleksi fitur dengan proses memasukan atribut independen yang memiliki korelasi paling besar dengan atribut dependennya (atribut yang paling potensial memiliki hubungan secara linier dengan Y). Setelah itu dilanjutkan dengan memasukkan atribut independen yang potensial berikutnya hingga proses berhenti sampai tidak ada atribut independen yang potensial lagi (Sidik, 2019).

BE merupakan sebuah metode seleksi fitur yang menggunakan teknik rekursif untuk mencari kombinasi fitur terbaik dari kumpulan kombinasi fitur. Cara kerja dari *backward elimination* dengan melakukan pengujian pada seluruh atribut atau fitur terlebih dahulu, lalu mengurangi fitur yang tidak signifikan berdasarkan perbandingan evaluasi hasil uji dari setiap kombinasi fitur secara bertahap (Raihan et al., 2021).

Selain dua seleksi fitur *Forward Selection* dan *Backward Elimination,* penelitian ini juga menggunakan kombinasi dari hasil seleksi fitur tersebut. Teknik kombinasi yang akan digunakan pada penelitian ini adalah kombinasi *union*.

2.4 *Split* Data

Metode *split data* adalah pemisahan data menjadi 2 bagian yaitu data uji (*testing*) dan data latih (*training*). Pada penelitian ini membagi data dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Berdasarkan penelitian yang dilakukan (Nurkholifah, Am and Oktorina, 2023) pemodelan algoritma NB menggunakan pembagian data 80:20 memiliki nilai akurasi terbaik dibandingkan pemodelan NB menggunakan pembagian data 70:30.

2.5 Klasifikasi Naive Bayes (NB)

Algoritma ini adalah salah satu metode klasifikasi suatu variabel tertentu yang menggunakan teknik probabilitas atau peluang dengan cara mencari frekuensi peluang terbesar dari kemungkinan klasifikasi. Salah satu keuntungan dari *Naive Bayes* adalah metode ini tidak memerlukan jumlah data pelatihan yang besar untuk dapat menentukan parameter yang akan digunakan untuk proses klasifikasi (Irawan, Windarto and Wanto, 2019). Sedangkan kekurangan metode ini adalah terdapatnya asumsi bahwa atribut-atribut yang ada pada data saling bebas atau independen. Padahal dalam prakteknya, tiap-tiap variabel yang ada memiliki hubungan satu sama lainnya (Siregar, Siregar and Sudirman, 2020). Dari penjelasan mengenai kekurangan algoritma NB, dalam penelitian ini menambahkan teknik seleksi fitur untuk memilih atribut apa saja yang penting dalam data.

2.6 Evalulasi

Setelah melakukan pemodelan *Naive Bayes* dengan masing-masing seleksi fitur, tahap selanjutnya melakukan evaluasi untuk mengetahui nilai akurasi dari masing-masing meode. Rumus perhitunganya terdapat pada formula 2.

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

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini *tools* yang digunakan untuk membantu dalam melakukan penelitian adalah *jupyter lab* dan *microsoft office excel.*

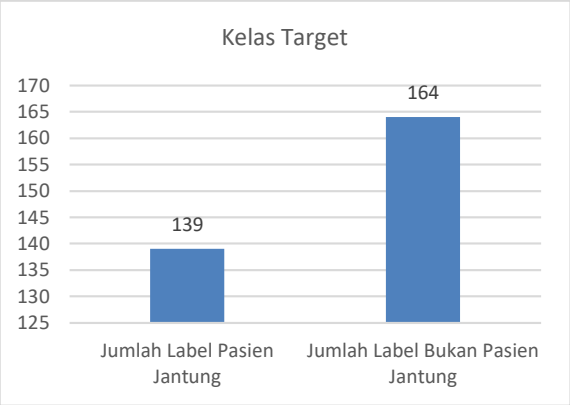
3.1 Pengumpulan Data

Data yang digunakan untuk penelitian ini diperoleh dari UCI Mechine Learning Repasitory yang memuat data medis mengenai penyakit jantung. Data penyakit jantung ini berjumlah 303 baris dengan 13 atribut independen dan 1 atribut dependen sebagai atribut target seperti pada tabel 1 dibawah ini.

Tabel 1. Dataset Penyakit Jantung

No	Sex	Cp	Fbs	...	Oldpeak	Target
1	1	2	0	...	2,3	0
2	1	3	1	...	1,5	1
3	1	1	1	...	2,6	1
...	...	...	...	...	...	...
303	0	4	0	...	0	0

Pada dataset ini, keluaran dari atribut dependen sebagai atribut targetnya adalah 1 sebagai pasien penyakit jantung dan 0 sebagai bukan pasien penyakit jantung yang masing-masing memiliki jumlah 139 dan 164 data seperti pada gambar 2.



Gambar 2. Bar Chart Distribusi Kelas Target

3.2 *Pre-Processing* Data

Pada penelitian ini, *preprocessing* data yang dilakukan ada dua proses yaitu *missing value handling* dan normalisasi data.

3.2.1 *Missing Value Handling*

Proses m*issing value handling* pada dataset bertujuan untuk mengetahui adanya data kosong atau data hilang kemudian dilakukan penangan terhadap data tersebut. Pada dataset penyakit jantung yang diperoleh untuk penelitian ini terdapat dua atribut yang *missing value* yaitu ca dengan jumlah *missing value* 4 dan Thal dengan jumlah *missing value* 2 seperti pada tabel 2.

Tabel 2. Atribut *Missing Value*

No	Ca	Thal
88	0	-
167	-	3
193	-	7
267	0	-
288	-	7
303	-	3

Penangan dari *missing value* tersebut adalah mengisi data kosong nilai modus dari atribut yang terdapat *missing value*. Nilai modus merupakan salah satu metode penanganan *missing value* dengan teknik imputasi. Selain itu, mengisi *missing* value menggunakan nilai modus cocok digunakan pada data yang berbentuk kategorik (Hendrawati, 2015). Atribut yang memiliki *missing value* pada penelitian ini yaitu atribut ca dan Thal. Selanjutnya kedua atribut tersebut akan diisi dengan nilai 0 untuk atribut ca dan nilai modus dari atribut Thal yaitu 3 seperti pada tabel 3.

Tabel 3. Hasil *Missing Value Handling*

No	Ca	Thal
88	0	**3**
167	**0**	3
193	**0**	7
267	0	**3**
288	**0**	7
303	**0**	3

3.2.2 Normalisasi Data

Pada dataset penyakit jantung yang diperoleh untuk penelitian ini memiliki rentang nilai yang berbeda-beda. Sehingga dilakukan normalisasi dengan teknik min-maks terhadap atribut numerik yaitu Oldpeak, Thalach, Chol, Trestbps, dan Age seperti pada tabel 4.

Tabel 4. Data Sebelum Normalisasi

No	Age	Trestbps	Chol	Thalach	Oldpeak
1	63	145	233	150	2.3
2	67	160	286	108	1.5
3	67	120	229	129	2.6
...	...	...	...	...	...
303	38	138	175	173	0

Hasil dari normalisasi data menggunakan min-maks *scaler* dapat dilihat pada tabel 5.

Tabel 5. Data Hasil Normalisasi

No	Age	Trestbps	Chol	Thalach	Oldpeak
1	0.70833	0.48113	0.24429	0.60305	0.37097
2	0.79167	0.62264	0.36530	0.28244	0.24194
3	0.79167	0.24528	0.23516	0.44275	0.41935
...	...	...	...	...	...
303	0.18750	0.41509	0.11187	0.77863	0

3.3 Seleksi Fitur

3.3.1 *Forward Selection*

Tahap pertama dari seleksi fitur FS adalah melatih model dengan semua fitur independennya, kemudian memilih 1 fitur sebagai fitur awal dan menghapus fitur yang yang tidak memiliki performa yang baik. Proses tersebut terus berlanjut dengan menambahkan 1 fitur untuk tiap proses dan akan berhenti pada saat parameter sudah terpenuhi. Setiap proses dari FS akan menghasilkan kombinasi-kombinasi fitur seperti pada tabel 6.

Tabel 6, Hasil Seleksi Fitur FS

Iterasi	Fitur Index	Skor
1	(6)	0.762131
2	(6, 7)	0.788525
3	(6, 7, 11)	0.808306
4	(4, 6, 7, 11)	0.828306
5	(4, 5, 6, 7, 11)	0.828361
6	(1, 4, 5, 6, 7, 11)	0.848087
7	(0, 1, 4, 5, 6, 7, 11)	0.851475
8	(0, 1, 4, 5, 6, 7, 8, 11)	0.848142
9	(0, 1, 4, 5, 6, 7, 8, 10, 11)	0.844918
10	(0, 1, 4, 5, 6, 7, 8, 9, 10, 11)	0.841585
11	**(0, 1, 4, 5, 6, 7, 8, 9, 10, 11, 12)**	**0.85153**
12	(0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12)	0.841639
13	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)	0.831639

Tabel tersebut menunjukan hasil seleksi fitur FS. Tiap iterasi terdapat kombinasi-kombinasi fitur dari proses yang sudah dilakukan dan memiliki skor terhadap atribut target. Pada tabel tersebut, dipilih kombinasi fitur pada iterasi ke-11, yaitu Sex, Cp, Exang, Slope, Thal, ca, Age, Trestbps, Chol, Thalach, dan Oldpeak. Kombinasi fitur tersebut dipilih karena memiliki skor tertinggi, kemudian kombinasi fitur

tersebut akan digunakan dalam implementasi algoritma *Naive Bayes* dengan pembagian data 80:20.

3.3.2 *Backward Elimination*

Pada iterasi pertama proses BE melakukan pelatihan model yang digunakan dengan semua fitur independen sebagai fitur awal. Pada iterasi selanjutnya, setelah melatih model dengan semua fitur independennya, kemudian mengeliminasi 1 fitur tidak signifikan. Proses tersebut terus berlanjut dengan mengeliminasi 1 atribut dari kombinasi atribut iterasi sebelumnya dan akan berhenti saat parameter sudah terpenuhi. Hasil kombinasi-kombinasi atribut dari proses BE dapat dilihat pada tabel 7.

Tabel 7. Hasil Seleksi Fitur BE

Iterasi	Fitur Index	Skor
1	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)	0.831639
2	(0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12)	0.841639
3	(0, 1, 2, 4, 5, 6, 7, 8, 9, 11, 12)	0.85153
4	(0, 1, 2, 4, 5, 6, 7, 9, 11, 12)	0.854699
5	**(0, 1, 4, 5, 6, 7, 9, 11, 12)**	**0.857978**
6	(0, 1, 5, 6, 7, 9, 11, 12)	0.844809
7	(0, 1, 5, 6, 7, 11, 12)	0.844754
8	(1, 5, 6, 7, 11, 12)	0.841421
9	(1, 5, 6, 7, 12)	0.828087
10	(1, 5, 6, 7)	0.824754
11	(1, 6, 7)	0.798361
12	(6, 7)	0.788525
13	(6)	0.762131

Pada tabel 7 dipilih kombinasi fitur pada iterasi ke-5. Kombinasi fitur tersebut dipilih karena memiliki skor tertinggi, kemudian kombinasi fitur tersebut akan digunakan dalam implementasi algoritma *Naive Bayes* dengan pembagian data 80:20 antara data latih dan data uji.

3.3.3 Kombinasi FS dan BE

Pada penelitian ini menggunakan kombinasi FS dan BE sebagai salah satu cara untuk memilih atribut yang akan digunakan dalam pemodelan. Teknik kombinasi pada penelitian ini adalah dengan menggabungkan hasil seleksi fitur FS dan BE secara *union* seperti pada tabel 8.

Tabel 8. Hasil Kombinasi FS dan BE

Seleksi Fitur	Atribut Index
FS	(0, 1, 4, 5, 6, 7, 8, 9, 10, 11, 12)
BE	(0, 1, 4, 5, 6, 7, 9, 11, 12)
FS + BE	(0, 1, 4, 5, 6, 7, 8, 9, 10, 11, 12)

Dapat dilihat dari tabel 6 bahwa kombinasi *union* seleksi fitur FS dan BE menghasilkan 11 kombinasi atribut, yaitu Sex, Cp, Exang, Slope, Thal, ca, Age, Trestbps, Chol, Thalach, dan Oldpeak. Kombinasi atribut tersebut akan digunakan dalam implementasi *Naive Bayes* dengan pembagian 80:20

3.3.4 *Information Gain*

Proses seleksi fitur IG menghasilkan nilai bobot yang disebut dengan nilai gain dari tiap atribut independen. Atribut-atribut yang sudah memiliki nilai gain akan dirangking dari nilai gain yang tertinggi seperti tabel 9.

Tabel 9. Nilai IG

No.	Atribut	Nilai IG
1	Thal	0.2063
2	Cp	0.2050
3	Ca	0.1815
4	Exang	0.1391
5	Thalach	0.1260
6	Oldpeak	0.1217
7	Slope	0.1124
8	Age	0.0602
9	Sex	0.0573
10	Restecg	0.0241
11	Chol	0.0177
12	Trestbps	0.0159
13	Fbs	0.0005

Setelah melakukan perangkingan atribut berdasarkan atribut tertinggi, selanjutnya memilih atribut yang akan dipilih berdasarkan perhitungan berikut.

$$log_2(13) = 3.70$$

Berdasarkan perhitungan diatas, dari 13 atribut independen pada dataset penyakit jantung dipilih 4 atribut dengan nilai gain tertinggi, yaitu Thal, Cp, ca, dan Exang. Atribut-atribut tersebut yang akan digunakan dalam implementasi model *Naive Bayes*.

3.3.5 *Gain Ratio*

Proses seleksi fitur GR dilakukan dengan menghitung nilai bobot gain ratio yang didapatkan dengan cara membagi nilai IG dengan Split Information. Selanjutnya nilai bobot GR tiap atribut dirangking dari yang tertinggi hingga terendah seperti pada tabel 10.

Tabel 10. Nilai GR

No.	Atribut	Nilai IG
1	Oldpeak	0.1685
2	Thal	0.1659
3	Exang	0.1526
4	Thalach	0.1284
5	Cp	0.1180
6	Ca	0.1165
7	Slope	0.0869
8	Sex	0.0633
9	Age	0.0604
10	Trestbps	0.0490
11	Restecg	0.0222
12	Chol	0.0207
13	Fbs	0.0008

Setelah melakukan perangkingan atribut berdasarkan atribut tertinggi, selanjutnya memilih atribut yang akan dipilih berdasarkan perhitungan berikut.

$$log_2(13) = 3.70$$

Berdasarkan perhitungan diatas, dari 13 atribut independen pada dataset penyakit jantung dipilih 4 atribut dengan nilai gain tertinggi, yaitu Oldpeak, Thal, Exang, Thalach. Atribut-atribut tersebut yang akan digunakan dalam implementasi model *Naive Bayes.*

3.3.6 Kombinasi IG dan GR

Pada penelitian ini menggunakan kombinasi *Information Gain* dan *Gain Ratio* sebagai salah satu cara untuk memilih atribut yang akan digunakan dalam pemodelan. Teknik kombinasi dilakukan dengan menggabungkan hasil hasil dari seleksi fitur IG dan GR secara *union* seperti pada tabel 11.

Tabel 11. Hasil Kombinasi IG dan GR

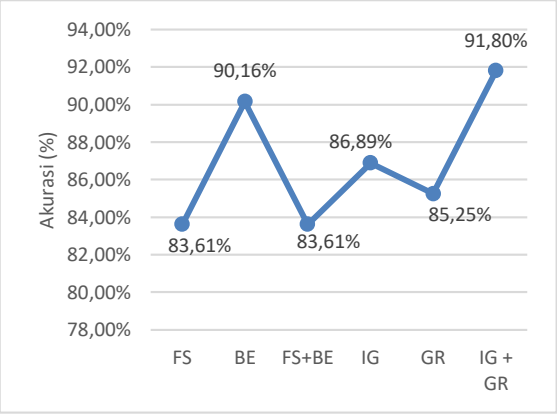
Seleksi Fitur	Atribut Index
IG	Thal, Cp, ca, Exang
GR	Thal. Thalach, Oldpeak, Exang
IG + GR	Thal, Exang, Cp, ca, Thalach, Oldpeak

3.4 Hasil Evaluasi Model

Evaluasi dilakukan terhadap model klasifikasi yang digunakan dengan kombinasi-kombinasi seleksi fiturnya. Evaluasi dilakukan guna mengetahui nilai akurasi model klasifikasi yang dibangun. Metode evaluasi pada penelitian ini yakni metode confusion matrix. Nilai akurasinya dapat dilihat pada tabel 12 dan gambar 3 berikut.

Tabel 12. Hasil Akurasi Setiap Model

Model	Akurasi
NB + FS	83.61%
NB + BE	90.16%
NB + FS + BE	83.16%
NB + IG	86.89%
NB + GR	85.24%
NB + IG + GR	91.80%



Gambar 3. Perbandingan Nilai Akurasi Setiap Model

Pada pemodelan NB menggunakan seleksi fitur FS dan kombinasinya dengan BE memiliki nilai akurasi yaitu sebesar 83.61%. Sedangkan pemodelan NB dengan BE saja mendapat nilai akurasi sebesar 90.16%.

Pemodelan NB dengan seleksi fitur IG memiliki nilai akurasi sebesar 86,89%. Selanjutnya pemodelan NB dengan seleksi fitur GR memiliki nilai akurasi sebesar 85.24%. Pemodelan yang terakhir adalah NB menggunakan kombinasi seleksi fitur IG dan GR yang memiliki hasil akurasi tertinggi pada penelitian ini yaitu 91.80% yang artinya mengalami peningkatan sebesar 8.19% dari pemodelan NB menggunakan kombinasi seleksi fitur FS dan BE.

4. KESIMPULAN

Berdasarkan paparan hasil dan pembahasan yang sudah dipaparkan sebelumnya, klasifikasi penyakit jantung menggunakan *Naïve Bayes* dengan perbandingan data latih dan uji sebesar 80:20 dengan seleksi fitur FS, BE, dan kombinasi keduanya serta seleksi fitur IG, GR, dan kombinasi keduanya dapat disimpulkan bahwa kombinasi seleksi fitur IG dan GR dengan teknik *union* memiliki nilai akurasi tertinggi yaitu sebesar 91.80% lebih tinggi dari pemodelan NB menggunakan kombinasi seleksi fitur IG dan GR dengan teknik *union* yaitu sebesar 83.61% saja.

DAFTAR PUSTAKA

ADNYANA, I.M.B., 2019. Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa. *Jurnal Sistem dan Informatika*, 13(2), pp.72–76.

AINI, S.H.A., SARI, Y.A. AND ARWAN, A., 2018. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, [online] 2(9), pp.2546–2554. Available at: <http://j-ptiik.ub.ac.id>.

AMILIA, I.R., OKTAVIANTO, H. AND ABDURRAHMAN, G., 2021. Penerapan Backward Elimination Untuk Seleksi Fitur Pada Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Gagal Jantung. *Jurnal Smart Teknologi*, 1(1), pp.100–102.

ARIFIN, T. AND ARIESTA, D., 2019. Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization. *Jurnal Tekno Insentif*, 13(1), pp.26–30. https://doi.org/10.36787/jti.v13i1.97.

HAMEED, S.S., PETINRIN, O.O., HASHI, A.O. AND SAEED, F., 2018. Filter-wrapper combination and embedded feature selection for gene expression data. *International Journal of Advances in Soft Computing and its Applications*, 10(1), pp.90–105.

HENDRAWATI, T., 2015. Kajian Metode Imputasi dalam Menangani Missing Data. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS*, [online] pp.637–642. Available at: <http://hdl.handle.net/11617/5804>.

IRAWAN, E., WINDARTO, A.P. AND WANTO, A., 2019. Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, (September), pp.750–761. https://doi.org/10.30645/senaris.v1i0.81.

MARTHA, S. AND SULISTIANINGSIH, E., 2018. K Nearest Neighbor Dalam Imputasi Missing Data. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, [online] 07(1), pp.9–14. Available at: <http://archive.ics.uci.edu/ml/datas/Iris.>.

NUGROHO, F.A., 2018. Perancangan Sistem Pakar Diagnosa Penyakit Jantung dengan Metode Forward Chaining. *Jurnal Informatika Universitas Pamulang*, 3(2), pp.75–79. https://doi.org/10.32493/informatika.v3i2.1431.

NUR, F., AHSAN, M. AND HARIANTO, W., 2022. Komparasi Tingkat Akurasi Information Gain Dan Gain Ratio Pada Metode K-Nearest Neighbor. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 6(1), pp.386–391. https://doi.org/10.36040/jati.v6i1.4694.

NURKHOLIFAH, M., AM, A.N. AND OKTORINA, F.K., 2023. Analisa Penyakit Jantung Menggunakan Algoritma Naïve Bayes. *Journal of System and Computer Engineering (JSCE)*, 4(1), pp.26–36. https://doi.org/10.47650/jsce.v4i1.671.

NURLIA, E. AND ENRI, U., 2021. Penerapan Fitur Seleksi Forward Selection Untuk Menentukan Kematian Akibat Gagal Jantung Menggunakan Algoritma C4.5. *Jurnal Teknik Informatika Musirawas) Elin Nurlia*, 6(1), pp.42–50.

OKTANISA, I. AND SUPIANTO, A.A., 2018. Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 5(5), pp.567–576. https://doi.org/10.25126/jtiik.201855958.

PRASETYO, E. AND PRASETIYO, B., 2020. Increased Classification Accuracy C4.5 Algorithm Using Bagging Techniques in Diagnosing Heart Disease. 7(5), pp.1035–1040. https://doi.org/10.25126/jtiik.202072379.

PUTRI, N.L., NUGROHO, R.A. AND HERTENO, R., 2021. Intrusion Detection System Berbasis Seleksi Fitur Dengan Kombinasi Filter Information Gain Ratio Dan Correlation. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(3), p.457. https://doi.org/10.25126/jtiik.0813154.

RAIHAN, M.A., HARYANDI, P., SUBAGJA, R.A., Purnaminyan, R. and Chamidah, N., 2021. Implementasi Seleksi Fitur dengan Backward Elimination untuk Klasifikasi Prediksi Perceraian. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, (April), pp.644–654.

RYFAI, D.A., HIDAYAT, N. AND SANTOSO, E., 2022. Klasifikasi Tingkat Resiko Serangan Penyakit Jantung menggunakan Metode K-Nearest Neighbor. 6(10).

SASONGKO, T.B. AND ARIFIN, O., 2019. Implementasi Metode Forward Selection pada Algoritma Support Vector Machine (SVM) dan Naive Bayes Classifier Kernel Density (Studi Kasus Klasifikasi Jalur Minat SMA). *Jurnal Teknologi Informasi dan Ilmu Komputer*, [online] 6(4), pp.383–388. https://doi.org/10.25126/jtiik.201961000.

SIDIK, Z., 2019. Klasifikasi Kelancaran Kredit Furniture Menggunakan Algoritma K-Nearest Neighbor Berbasis Forward Selection.

SIREGAR, N.C., SIREGAR, R.R.A. AND SUDIRMAN, M.Y.D., 2020. Jurnal Teknologia Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ) Jurnal Teknologia. *Jurnal Teknologia*, 3(1), pp.102–110.

SUCHETHA, N.K., NIKHIL, A. AND HRUDYA, P., 2019. Comparing the wrapper feature selection evaluators on twitter sentiment classification. *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*, (October). https://doi.org/10.1109/ICCIDS.2019.8862033.

SUPRIANTO, S., 2020. Implementasi Algoritma Naive Bayes Untuk Menentukan Lokasi Strategis Dalam Membuka Usaha Menengah Ke Bawah di Kota Medan (Studi Kasus: Disperindag Kota Medan). *Jurnal Sistem Komputer dan Informatika (JSON)*, 1(2), pp.125–130. https://doi.org/10.30865/json.v1i2.1939.

SURYANEGARA, G.A.B., ADIWIJAYA AND PURBOLAKSONO, M.D., 2021. Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(1), pp.114–122. https://doi.org/10.29207/resti.v5i1.2880.

TARIGAN, P.M.S., HARDINATA, J.T., QURNIAWAN, H., M.Safii and Winanjaya, R., 2022. Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus: Toko Swapen Jaya Manokwari). *Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, 12(2), pp.51–61. https://doi.org/10.33379/gtech.v7i1.1938.