SELEKSI FITUR MENGGUNAKAN PENAMBANGAN DATA BERBASIS *VARIABLE PRECISION ROUGH SET* (VPRS) UNTUK DIAGNOSIS PENYAKIT JANTUNG KORONER

Dwi Normawati, Sri Winiarti
Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
Kampus III, Jln. Prof. Dr. Soepomo, S.H. Umbulharjo, Yogyakarta 55161
e-mail: dwi.normawati@tif.uad.ac.id

Abstract

Coronary heart disease often causes death on human. This disease occurs when there is atherosclerosis (fat deposits) that blocks the flow of blood to the heart muscle in the coronary arteries. The gold standard method that doctors refer to diagnose coronary heart disease is the coronary angiography. However, this method is invasive, high risk and expensive.The purpose of this research is to perform a diagnosis of coronary heart disease based on computer using datamining by the feature selection method and by classifying the dataset of Cleveland heart disease.The feature selection method based on medical expert (MFS) and also the feature selection method based on computer, which is the feature selection method using the data mining method based on the Variable Precision Rough Set (VPRS) theory. Which is the development from the Rough Set Theory. This research begins by doing literature study about the feature selection method based on medical expert or motivated feature selection (MFS) and feature selection method based on computer that is the VPRS theory. To avoid missing features that the research considered by medical expert and by computer. In the end, the feature selection process based on VPRS and the combination of VPRS with MFS can improve the classification performance for diagnosing coronary heart disease significantly. This can be seen by the smaller number of rules generated and the values accuracy which is better than the classification with-out the feature selection.

Keywords: *coronary heart disease; Cleveland; feature selection*

Abstrak

Penyakit jantung koroner merupakan penyakit yang banyak menyebabkan kematian pada manusia. Penyakit ini terjadi ketika terdapat *atherosclerosis* (timbunan lemak) yang menghalangi aliran darah ke otot jantung pada arteri koronaria. Metode *gold standard* yang menjadi rujukan para dokter untuk mendiagnosis penyakit jantung koroner adalah *coronary angiography*. Namun metode ini *invasisve*, mempunyai resiko tinggi dan mahal. Tujuan penelitian ini adalah melakukan diagnosis penyakit jantung koroner berbasis komputer menggunakan *data mining* dengan melakukan seleksi fiture dan melakukan klasifikasi pada *dataset* penyakit jantung *Cleveland*. Pada penelitian ini, menggunakan metode seleksi fitur berbasis pakar medis (MFS) dan juga menggunakan metode seleksi fitur berbasis komputer yaitu metode seleksi fiture menggunakan metode *data mining* berbasis teori *Varible Precision Rough Set* (VPRS) yang merupakan pengembangan dari teori *Rough Set*. Pada penelitian ini, studi *literature* tentang metode seleksi fiture berbasis pakar medis atau *motivated feature selection* (MFS) dan metode seleksi fitur berbasis komputer yaitu berbasis teori VPRS dilakukan. Penggabungan metode seleksi fitur berbasis pakar medis dan komputer juga dilakukan agar dapat menghindari terhapusnya fitur-fitur yang dianggap penting oleh pakar medis. Pada akhirnya, proses seleksi fitur basis komputer yaitu VPRS dan penggabungan VPRS dengan MFS mampu meningkatkan performa klasifikasi secara signifikan untuk mendiagnosis penyakit jantung koroner, dilihat dari lebih sedikitnya jumlah *rule* yang dihasilkan dan nilai hasil akurasi yang lebih baik dibandingkan dengan klasifikasi tanpa seleksi fitur.

Kata kunci: penyakit jantung coroner; *Cleveland*; seleksi fitur

1. Pendahuluan

Penyakit jantung koroner adalah penyakit pembunuh yang paling di seluruh dunia. Jumlah pasien dengan penyakit jantung koroner meningkat setiap tahun, data dari Organisasi Kesehatan Dunia (WHO) menyatakan bahwa 17,5 juta orang diperkirakan meninggal pada tahun 2012, yang mewakili 31% dari semua kematian global dan diperkirakan 7,4 juta disebabkan penyakit jantung koroner [1]. Gejala utama penyakit jantung koroner mendeteksi nyeri dada atau angina, tetapi cara ini sangat sulit dilakukan karena gejala pada pasien mungkin terlihat kabur dan nyeri dada juga bisa sering terjadi dalam kondisi yang mungkin tidak disertai dengan penyakit jantung koroner, ini kadang-kadang mengarah untuk diagnosis yang salah dengan penyakit lain [2]. Seiring dengan perkembangan teknologi informasi, banyak diagnosis penyakit jantung koroner dikembangkan menggunakan metode dibantu komputer [3]. Salah satu teknik untuk menangani sejumlah besar data adalah *data mining* [4]. *Data mining* di dunia medis memiliki potensi besar untuk menemukan pola tersembunyi dalam *data set* medis. Pola dapat digunakan untuk membantu dokter secara signifikan meningkatkan kualitas keputusan medis untuk mengungkapkan ada tidaknya penyakit [5]. Berbagai penelitian-penelitian lain berkaitan dengan diagnosis penyakit jantung koroner melalui algoritme *data mining* telah dilakukan. Peneliti menyajikan algoritme *data mining* berbasis *rules* dalam mendiagnosis penyakit jantung koroner. Algoritme berbasis *rules* dipilih karena menghasilkan aturan/*rules* yang sederhana tetapi memiliki akurasi yang cukup tinggi [6]. Metode seperti RIPPER [7][8], *Artificial Neural Network* (ANN) [9], *Decision Tree* [10] diusulkan untuk mendiagnosis penyakit jantung koroner. Pada diagnosis medis, reduksi data merupakan masalah yang penting. Data medis sering mengandung sejumlah fitur yang tidak relevan, *redundant*, dan sejumlah kasus yang *relative* sedikit sehingga mempengaruhi kualitas diagnosis penyakit [11]. Oleh karena itu, proses seleksi fitur dilakukan untuk menyeleksi fitur-fitur yang relevan pada data medis. Proses seleksi fitur diusulkan dalam banyak penelitian untuk meningkatkan akurasi pada proses diagnosis penyakit jantung koroner. Jumlah data riwayat medis pasien yang berkaitan dengan penyakit jantung koroner didunia ini semakin hari semakin bertambah banyak, namun disisi lain dengan jumlah data yang besar terdapat data yang kurang relevan. Untuk mengatasi permasalah tersebut teori *rough set* diusulkan oleh Pawlak [12]. Teori *rough set* mampu menangani dan menganalisa data dan menemukan pengetahuan dari data yang tidak lengkap, tidak tepat dan ambigu. Dengan kata lain, pendekatan *rough set* mampu menemukan pola data dari data yang tidak sempurna. Namun dalam perkembangannya teori *rough set* memiliki banyak kekurangan ketika digunakan dalam *data mining* yaitu masalah analisis klasifikasi karena terkadang informasi yang tersedia hanya dapat memenuhi untuk klasifikasi parsial [13], maka Ziarko memperkenalkan suatu model baru yaitu *Variable Precision Rough Set* (VPRS).

Penelitian menggunakan metode VPRS untuk kasus diagnosis penyakit jantung koroner telah dilakukan untuk menemukan pola data berupa *rules* berbasis klasifikasi [14], menghasilkan aturan/*rules* yang jumlahnya lebih sedikit dibandingkan dengan metode *rough set* dan *rules* yang dihasilkan oleh VPRS lebih mudah dipahami dan apabila dilakukan reduksi *rules* maka nilai akurasi justru menurun [15]. Diagnosis Penyakit Jantung Koroner dengan klasifikasi VPRS mengahasilkan nilai akurasi 75,22% [12] ujicoba dilakukan dengan melakukan pengacakan data sebanyak 30 kali, akan tetapi performa akurasi setiap *rule* tidak diketahui.

Oleh karena itu, penelitian tentang menggunakan *data mining* berbasis VPRS dilakukan untuk solusi seleksi fitur meningkatkan performa keakurasian diagnosis penyakit jantung koroner. Pada penelitian ini diteliti metode seleksi fitur pada *data set* penyakit jantung koroner dari *Repository Machine Learning* UCI *AZZAASW* untuk performa diagnosis. Diharapkan hasil diagnosis dari penelitian ini mampu memberikan masukan kepada dokter sebelum melakukan uji *coronary angiography* sehingga dapat menghindari prosedur *invasive*, berisiko dan mahal.

2. Data

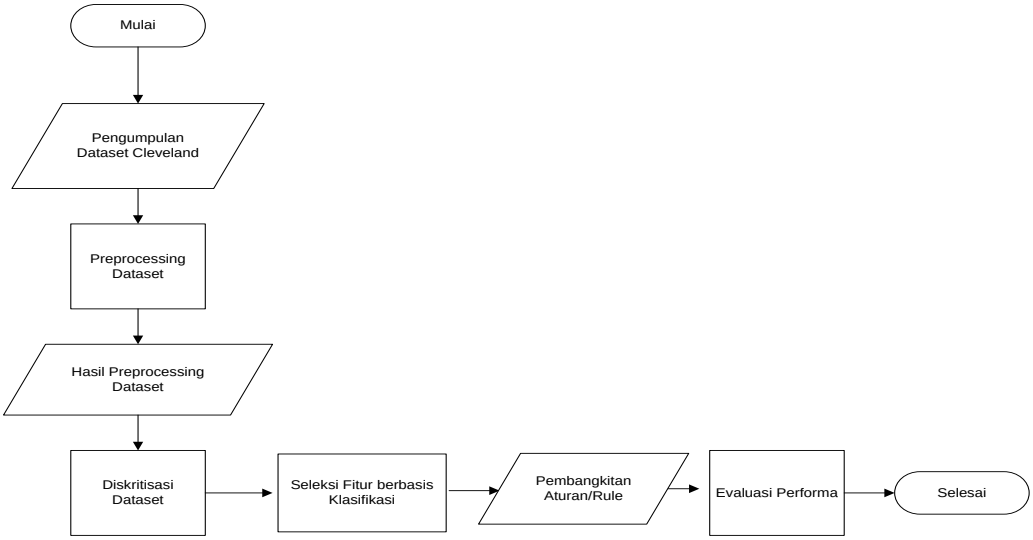
Dalam penelitian ini, 303 data *data set Cleveland* Penyakit Jantung dari UCI *repository* pembelajaran mesin memiliki 7 data nilai yang hilang. Data dari nilai-nilai yang hilang dihapus agar tidak mempengaruhi hasil klasifikasi. *Data set Cleveland* Penyakit Jantung terdiri dari 14 atribut. Dua kelas *data set* ini adalah tidak adanya dan adanya penyakit jantung. Tabel 1 menggambarkan atribut di *Cleveland* penyakit jantung *data set* [16].

Tabel 1. Keterangan *data set* penyakit jantung *cleveland*

Nama Atribut	Deskripsi Atribut	Keterangan
Age	*Age*	*Numerik*
Sex	*Sex*	0: Perempuan, 1: Laki-laki
Cp	*Chest pain type*	1: *typical angina*; 2: *atypical angina*; 3: *non-anginal pain*; 4: *asymptomatic*
Trestbps	*Resting blood pressure*	*Numerik*
Chol	Serum kolesterol	*Numerik*
Fbs	*Fasting blood sugar* >120 mg/dl	0: *false*; 1: *true*
Restecg	Hasil ECG selama istirahat	0: normal; 1: abnormal (memiliki kelainan gelombang ST-T); 2: *Hipertrofil ventrikel*
Thalac	Detak jantung maksimal yang dicapai	*Numeric*
Exang	Ukuran *boolean* yang menunjukan apakah latihan angina industri terjadi	0: *No*, *and* 1: *Yes*
Oldpeak	*Segment* ST yg diperoleh dari latihan relatif terhadap istirahat	*Numerik*
Slope	Kemiringan segmen ST untuk latihan maksimal (puncak).	1: *upsloping*; 2: *flat*; 3: *downsloping*
Ca	Jumlah *vessel* utama yang diwarnai oleh *fluoroskopi*	0, 1, 2 *and* 3
Thal	*Thal*	3: normal; 6: cacat tetap; 7: cacat *reversible*

3. Metodologi

Secara umum ada 6 tahapan yang dilakukan pada penelitian ini yaitu pengumpulan *data set Cleveland, preprocessing data set*, diskritisasi *data set*, pembangkitan aturan/*rule* dari *data set*, seleksi fitur berbasis klasifikasi *data set*, dan evaluasi performa. Tujuan penelitian ini dilakukan untuk menganalisa performa seleksi fitur berbasis klasifikasi untuk memperoleh fitur-fitur tebaik dalam mendiagnosis penyakit jantung koroner. Jalan penelitan terdiri dari beberapa langkah seperti yang ditunjukan pada diagram alir penelitian di Gambar 1.



Gambar 1. Diagram alir penelitian

3.1. Prapengolahan Data

Prapengolahan data adalah langkah pertama dalam mendiagnosis penyakit jantung koroner. Proses data prapengolahan pada penelitian ini akan terbagi menjadi dua langkah yaitu

data *cleaning* dan konversi data *multiclass* menjadi *binary class*. Pada proses data *cleaning* dilakukan penghapusan data yang *missing values*. *Missing Value* merupakan permasalahan yang sering ditemukan dalam menangani data medis. Pada *data set Cleveland* penyakit jantung koroner terdiri dari 303 *instances*, terdapat 7 *instances* yang memiliki *missing values*. Kemudian proses konversi data *multiclass* menjadi *binary class* dilakukan dengan mengasumsikan bahwa satu kelas positif yaitu sehat (0) dan empat kelas lainnya menjadi satu kelas negatif yaitu sakit (1). Konversi data *multiclass* menjadi *binary class* dilakukan karena metode klasifikasi yang digunakan merupakan metode klasifikasi yang hanya menggunakan data biner.

3.2. Diskritisasi Data

Diskretisasi merupakan suatu proses konversi data tipe numerik menjadi tipe diskrit [11]. Nilai-nilai diskrit memiliki sejumlah batas *interval* dalam sebuah *spectrum* numerik, sedangkan nilai-nilai numerik tak terhingga banyaknya. Dengan proses diskretisasi data dapat dikurangi, disederhakan untuk pengguna dan para ahli, serta fitur diskrit lebih mudah dipahami, digunakan dan dijelaskan pada algoritme *machine learning* [17]. Pada penelitian ini proses diskritisasi data menggunakan *Entropy/MDL* (*Minimum Description Length*) yaitu metode diskretisasi yang diperkenalkan oleh J. Dougherty, R. Kohavi, and M. Saham. Metode ini didasarkan pada partisi secara rekursif pada nilai-nilai setiap atribut sehingga hasil pengukuran *entropy* dapat dioptimalkan. MDL atau *Minimum Description Length* mendefinisikan kapan berhentinya proses partisi. Nilai-nilai yang hilang (*missing values*) diabaikan dalam pencarian pemotongan. Jika tidak ada pemotongan yang ditemukan untuk sebuah atribut, atribut yang tersisa tidak diproses [17].

3.3. Seleksi Fitur

Seleksi fitur berbasis komputer dilakukan unutk mereduksi data *Cleveland* dan juga memilih fitur-fitur yang relevan terhadap hasil keputusan diagnosis penyakit jantung koroner. Pada penelitian ini digunakan dua jenis seleksi fitur yaitu berbasis komputer dan berbasis pakar medis. Seleksi fitur dilakukan dengan cara mereduksi fitur-fitur dengan metode seleksi fitur VPRS dan menggabungkan metode seleksi fitur berbasis pakar medis (MFS) agar dapat menghindari terhapusnya fitur-fitur yang dianggap penting oleh pakar medis untuk diagnosis penyakit jantung koroner. Pada penelitian ini untuk seleksi fitur dengan metode VPRS menggunakan *software* ROSE2.

3.3.1. Seleksi Fitur berbasis Pakar Medis

Seleksi fitur berbasis pakar medis atau *motivated feature selection* (MFS) didasarkan pada pengetahuan yang dimiliki oleh pakar medis. Pada kasus penyakit jantung koroner, pakar medis menentukan delapan faktor signifikansi medis yang berpengaruh dalam proses diagnosis. Delapan faktor tersebut adalah *age, chest pain type(angina, abnang, notang, asympt), resting blood pressure, cholesterol, fasting blood sugar, resting heart rate (normal, abnormal, ventricular hypertrophy), maximum heart rate* dan *exercise induced angina* [18]. Delapan faktor tersebut digunakan sebagai seleksi fitur berbasis pakar medis.

3.3.2. Seleksi Fitur berbasis Variable *Precision Rough Set* (VPRS)

Model VPRS merupakan kelanjutan dari model *rough set* klasik, yang diusulkan untuk menganalisis dan mengidentifikasi pola data yang mewakili *trend statistic* yang lebih fungsional [19]. VPRS berkaitan dengan klasifikasi parsial dengan memperkenalkan parameter presisi β. Ziarko mendefinisikan nilai β sebagai kesalahan klasifikasi dan berkisar dalam nilai $0 \le \beta < 0{,}5$. *Procedures* Prosedur model VPRS memiliki 4 (empat) langkah sebagai berikut yaitu langkah ke-1: Memilih nilai parameter presisi β, ke-2: Mencari himpunan penuh dari *β-reduct,* ke-3: Menghapus duplikat objek, ke-4: *Rule extraction* [13]. VPRS adalah pendekatan untuk analisis data yang bergantung pada 2 (dua) konsep dasar yaitu β-*lower* dan β-*upper approximations* yang dapat disajikan dalam persamaan sebagai berikut.

$$\underline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)\le\beta} \{x_i \in E(P)\} \quad (1)$$

$$\overline{C}_{\beta}(D) = \bigcup_{1-P_r(Z|x_i)<1-\beta} \{x_i \in E(P)\} \quad (2)$$

dimana E(P) menunjukan sebuah himpunan kelas-kelas ekuivalen, dan kondisi kelas berdasarkan *subset* atribut P, sedangkan $Z \subset E(D)$,

$$P_r(Z \mid x_i) = \frac{Card(Z \cap x_i)}{Card(x_{i)}} \quad (3)$$

Menurut [20], ukuran *quality of classification* untuk model VPRS dapat didefinisikan dengan persamaan sebagai berikut :

$$\gamma(P, D, \beta) = \frac{Card(\bigcup_{1-P_r(Z|x_i) \le \beta} \{x_i \in E(P)\}}{Card(U)} \quad (4)$$

dimana $Z \subset E(D)$ dan $P \subseteq C$, untuk nilai β tertentu. Nilai persamaan (2-5) mengukur proporsi obyek pada himpunan semesta (U) untuk klasifikasi berdasarkan atribut keputusan D, dan memungkinkan untuk nilai β tertentu.

Prosedur untuk menghasilkan aturan keputusan dari suatu sistem informasi dilakukan dengan 2 (dua) langkah utama sebagai berikut: Langkah ke-1: Pemilihan himpunan terkecil yang terbaik dari atribut-atribut (contoh: pemilihan nilai *β-reduct*) and Langkah ke-2: Penyederhanaan sistem informasi dapat dicapai dengan menjatuhkan nilai-nilai tertentu dari atribut yang tidak perlu untuk sistem informasi. Ziarko [21] mengindikasikan bahwa setiap himpunan terkecil atribut dianggap sebagai kelompok atribut alternatif yang dapat digunakan sebagai pengganti semua atribut yang tersedia di pengambilan keputusan berbasis kasus. Kesulitan utama adalah bagaimana memilih nilai *β-reduct* yang optimal. Untuk kasus tersebut ada dua pendekatan yang dapat digunakan. Pendekatan pertama, *β-reduct* dengan jumlah atribut paling sedikit yang dipilih, sedangkan pendekatan kedua, *β-reduct* yang memiliki jumlah terkecil kombinasi dari nilai-nilai atributnya yang dipilih.

3.4. Pembangkitan Aturan/*Rule IF-THEN*

Langkah selanjutnya adalah melakukan proses pembangkitan aturan-aturan (*rule generation*). Sebelum dilakukan pembangkitan aturan (*rule generation*), *data set* hasil diskritisasi dibagi menjadi dua yaitu *data set* latih (2/3 bagian) dan *data set* uji (1/3 bagian) dengan metode stratifikasi. Pada penelitian ini menggunakan metode pembangkitan aturan VPRS dengan menggunakan *toolkits* ROSE2.

Diagnosis penyakit jantung koroner berbasis komputer dilakukan dengan cara mengklasifikasikan aturan-aturan (*Rules*) yang dihasilkan pada proses pembangkitan aturan *IF-THEN* terhadap objek *data set* uji, kemudian menghitung nilai akurasi dengan menggunakan *confusion matrix*.

3.5. Evaluasi Performa

Evaluasi performa dilakukan dengan menganalisis performa klasifikasi. Klasifikasi merupakan pemetaan dari *instance* ke *class* prediksi [22]. *Confusion matrix* adalah alat visualisasi yang digunakan untuk menyajikan performa klasifikasi yang berisi informasi tentang hasil aktual dan hasil prediksi yang dilakukan oleh sistem. Pada penelitian ini evaluasi performa dilakukan dengan metode VPRS dengan menggunakan tiga performa yaitu akurasi, sensitivitas dan spesifisitas, dengan mengguanakan *confusion matrix* hasil klasifikasi, sehingga diperoleh hasil seleksi fitur terbaik.

Perbandingan performa metode klasifikasi dengan menggunakan tiga performa yaitu akurasi, sensitivitas dan spesifisitas. Dengan *confusion matrix* dapat diketahui perbandingan 4 kategori yang ditunjukan pada Tabel 2.

Tabel 2. *Confusion matrix* hasil klasifikasi

		Predicted	
		Negative	*Positive*
Actual	*Negative*	*TN*	*FP*
	Positive	*FN*	*TP*

Persamaan (3-1) memberikan formula untuk menghitung nilai akurasi.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

Dimana:
TP: *True positive* (banyaknya data positif yang terprediksi sehat/positif)
TN: *True Negative* (banyaknya data negatif yang terprediksi sakit/negatif)
FP: *False Positive* (banyaknya data positif yang terprediksi sakit/negatif)
FN: *False Negative* (banyaknya data negatif yang terprediksi sehat/positif)

4. Hasil dan Pembahasan

Dalam penelitian, 296 data yang digunakan dari *data set Cleveland* penyakit jantung. Data di diskritisasi dengan menggunakan *software* ROSETTA, seleksi fitur dengan metode VPRS menggunakan *software* ROSE2, pembangkitan aturan/*rule IF-THEN* menghasilkan *rule* menggunakan *software* VPRS, dan proses klasifikasi dihitung secara manual dengan menggunakan Microsoft Excel.

4.1. Data Preprocessing

Langkah pertama dalam data prapengolahan adalah proses data *cleaning*. Proses data *cleaning* yaitu menghapus data yang ber-*missing value* dilakukan secara manual. *Data set* penyakit jantung *Cleveland* terdiri dari 303 *instances* dan terdapat 7 *intances* yang ber-*missing values* akan dihapus, sehingga *data set* yang digunakan pada penelitian ini sebanyak 296 *instances*.

Tabel 3. Karakteristik *data set cleveland* sesudah *cleaning* data

Dataset	Kelas Positif	Jumlah *instance* kelas positif	Jumlah *instance* kelas negative	Jumlah total *instance*
H-0	Sehat	160	136	296
Sick-1	Sakit1	53	244	296
Sick-2	Sakit2	35	261	296
H-0	Sehat	160	136	296
Sick-1	Sakit1	53	244	296

Langlah kedua adalah mengkonversi *data set multiclass* menjadi *binary class*. *Data set Cleveland* memiliki lima kelas yaitu H-0, *Sick-1, Sick-2, Sick-3*, dan *Sick-4* (0,1,2,3,4), proses konversi data *multiclass* menjadi *binary class* dilakukan dengan mengasumsikan bahwa satu kelas positif yaitu sehat (0) dan empat kelas lainnya menjadi satu kelas negatif yaitu sakit (1).

Tabel 4. Karakteristik *data set cleveland* sesudah konversi data

Dataset	Kelas Positif	Jumlah *instance* kelas positif	Jumlah *instance* kelas negatif	Jumlah total *instance*
Health	Sehat	160	136	296
Sick	Sakit	136	160	296

4.2. Diskritisasi Data dan Nilai Parameter Presisi

Langkah berikutnya adalah data diskritisasi yaitu mengubah tipe data dari atribut dari numerik ke diskrit. Beberapa atribut dengan jenis numerik yang memiliki *Cleveland data set* adalah usia, *trestbps, chol, thalach, oldpeak* dan *ca*, mereka berubah menjadi tipe diskrit menggunakan algoritma Entropi / MDL. Tabel 5 menunjukkan hasil data diskritisasi.

Tabel 5. Hasil diskretisasi *data set*

	Atribut-atribut bertipe numerik					
	Age	*Trestbps*	*Chol*	*Thalach*	*Oldpeak*	*Ca*
Nilai Diskritisasi	[*, 71) [71, 77) [77, *)	[*, 186) [186, *)	[*, 276) [276, 277) [277, 280)	[*, 148) [148, 151) [151, 162)	[*, 2.5) [2.5, 2.7) [2.7, 3.1)	[*,3) [3, *)

[280, 295)	[162, 170)	[3.1, 3.5)
[295, 299)	[170, 172)	[3.5, 3.6)
[299, 301)	[172, 175)	[3.6, 4.3)
[301, 319)	[175, 176)	[4.3, *)
[319, 320)	[176, 178)	
[320, 322)	[178, 183)	
[322, 324)	[183, 195)	
[324, 326)	[195, 199)	
[326, 338)	[199, *)	
[338, 341)		
[341, 342)		
[342, 348)		
[348, 354)		
[354, 401)		
[401, 413)		
[413, *)		

Setelah dilakukan diskritisasi data, langkah selanjutnya adalah menentukan nilai parameter presisi diperoleh dengan menghitung rata-rata nilai akurasi dari *validasi basic minimal covering* dengan menggunakan *toolkit* ROSE2. Setiap prosesnya didapat nilai rata-rata akurasi yang selalu berubah, maka pada penelitian ini dilakukan perhitungan nilai sebanyak 30 kali untuk setiap nilai β. Nilai β yang diuji yaitu nilai yang didefinisikan oleh Ziarko yaitu $0 \leq \beta < 0,5$ dengan selisih masing-masing setiap nilai β-nya adalah 0,05.

Tabel 6. Rata-rata nilai *validasi basic minimal covering* untuk mencari nilai β

Percobaan	Nilai β									
Ke-	0,05	0,1	0,15	0,2	0,25	0,3	0,33	0,35	0,4	0,45
1	81,74	82,44	80,41	77,64	77,07	79,68	79,08	80,80	82,45	78,34
2	79,74	79,40	78,37	79,37	80,08	79,77	80,02	80,13	79,70	80,36
3	77,63	78,10	78,09	82,78	78,72	80,75	82,16	81,79	80,43	77,45
4	80,76	81,71	75,70	77,99	77,74	79,00	78,69	80,78	81,75	78,01
5	79,40	82,08	80,44	78,72	81,09	82,05	79,47	77,67	79,33	79,08
6	80,38	77,36	80,79	78,46	80,44	80,71	80,44	81,71	78,06	78,37
7	78,02	80,43	81,11	81,46	80,77	80,06	75,34	82,75	83,10	79,74
8	79,07	80,05	81,44	82,45	80,06	78,37	79,83	77,03	78,33	78,74
9	82,38	78,75	81,79	79,68	77,01	80,74	81,45	77,02	79,07	81,45
10	80,36	76,44	80,76	81,76	78,37	78,39	80,15	80,76	77,70	77,32
11	77,74	82,74	79,46	78,47	80,09	79,38	78,69	79,69	80,45	80,37
12	79,07	79,06	82,83	79,06	79,87	80,03	78,08	81,78	81,10	79,08
13	79,36	78,66	82,16	79,07	80,78	81,40	80,39	78,36	81,83	80,01
14	80,72	81,13	79,46	78,70	80,07	79,34	78,03	82,46	77,71	81,08
15	80,74	81,38	79,71	83,46	77,38	79,45	79,40	78,69	79,39	77,06
16	80,07	79,74	79,69	80,08	80,08	76,61	79,37	79,72	78,69	79,44
17	79,79	78,68	79,67	79,45	80,43	79,75	78,70	79,70	79,08	79,67
18	82,15	77,70	79,75	79,72	79,10	82,49	77,07	80,10	80,79	78,66
19	80,07	78,36	78,36	80,74	81,79	80,76	81,45	78,68	81,07	80,72
20	81,07	80,05	80,46	82,40	79,79	76,70	80,44	79,77	79,67	81,10
21	77,79	79,84	80,72	78,40	79,05	79,34	82,45	80,10	80,83	77,70
22	80,78	78,38	80,78	81,75	78,34	79,11	77,37	80,75	80,00	78,15
23	78,13	78,70	81,75	79,08	80,05	79,45	82,41	78,39	79,36	81,13
24	78,37	81,06	80,80	77,70	79,38	80,44	81,47	78,67	80,02	78,70
25	79,40	79,06	81,07	78,75	80,09	80,69	80,38	80,70	80,36	80,10
26	77,74	80,44	80,72	79,79	79,41	80,01	77,07	80,06	81,06	79,07
27	79,41	80,08	79,37	77,70	78,77	79,45	80,10	78,37	77,76	79,38
28	80,40	78,43	80,46	79,77	79,43	78,37	79,74	81,75	77,02	78,38
29	79,68	79,00	81,41	76,74	81,76	80,72	78,40	80,11	77,71	77,36
30	80,40	80,43	81,08	81,01	80,11	78,68	79,06	80,45	75,33	79,74
Rata-Rata	**79,75**	**79,66**	**80,29**	**79,74**	**79,57**	**79,72**	**79,56**	**79,96**	**79,64**	**79,19**

Dari hasi pengujian yang ditunjukan pada table 4, nilai rata-rata akurasi tertinggi adalah 80.29, sehingga nilai β yang digunakan pada metode VPRS adalah 0.15.

4.3. Seleksi Fitur

Langkah awal pada proses seleksi fitur ini adalah membagi *data set* menjadi dua bagian atau dikenal dengan *splitting data set*. *Data set* dibagi menjadi dua bagian dengan jumlah yang sama yaitu 2/3 data digunakan sebagai data latih dan 1/3 data lainnya digunakan sebagai data uji. Proses seleksi fitur dilakukan pada *data set* latih, kemudian dilakukan reduksi data pada *data set* latih dan uji sesuai dengan fitur-fitur yang dipilih. Tabel 7 menunjukan fitur-fitur yang dipilih oleh MFS dan VPRS.

Tabel 7. Hasil seleksi fitur

Metode Seleksi Fitur	Hasil Seleksi Fitur
MTF	*age, cp, trestbps, chol, fbs, restecg, thalach, exang*
VPRS	*cp, chol, restecg, thalach, exang, oldpeak, slope, thal*
MTF+VPRS	*age, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, thal*

4.4. Pembangkitan Aturan/*Rule IF-THEN*

Pada tahap ini, langkah yang dilakukan adalah melakukan pembangkitan aturan/*rule IF-THEN* dengan menggunakan metode VPRS pada *data set* latih hasil seleksi fitur dengan MTF, VPRS dan penggabungan MTF & VPRS. *Spliting data set* dilakukan untuk membagi data menjadi *data set* latih dan *data set* uji, *data set* latih digunakan untuk mencari pengetahuan dalam data, sedangkan *data set* uji digunaka untuk menguji data dengan cara mencocokan hasil pengetahuan dengan data.

Pada penelitian ini aturan/*rules IF-THEN* didapatkan menggunakan *toolkit* ROSE2 untuk metode VPRS dengan data yang digunakan adalah *data set* latih yang telah didiskrit. Untuk mendapatkan aturan/*rules IF-THEN* atau aturan keputusan untuk metode VPRS menggunakan *toolkit* ROSE2 dengan nilai β = 0.15. Aturan/*Rules IF-THEN* yang dihasilkan ditunjukan pada Tabel 8, Tabel 9 dan Tabel 10.

Tabel 8. Aturan/*rules data set* metode seleksi fitur berbasis pakar (MTF)

Rule	Aturan/*Rules*
1	(cp = 3) & (*chol* = 0) & (*exang* = 0) => (*class* = 0)
2	(*chol* = 0) & (*thalach* = 2) & (*exang* = 0) => (*class* = 0)
…	……
…	……
59	(cp = 2) & (*chol* = 3) & (fbs = 0) & (*thalach* = 2) => (*class* = 0) OR (*class* = 1)
60	(cp = 4) & (fbs = 0) & (*restecg* = 0) & (*thalach* = 3) & (*exang* = 0) => (*class* = 0) *OR* (*class* = 1)

Tabel 9. Aturan/*Rules data set* metode seleksi fitur VPRS

Rule	Aturan/*Rules*
1	(cp = 2) & (*chol* = 0) & (*thal* = 3) => (*class* = 0)
2	(*thalach* = 1) & (*thal* = 3) => (*class* = 0)
…	……
…	……
54	(cp = 2) & (*thalach* = 2) & (*slope* = 2) => (*class* = 0) *OR* (*class* = 1)
55	(*chol* = 0) & (*restecg* = 2) & (*thalach* = 0) & (*oldpeak* = 0) & (*slope* = 1) & (*thal* = 7) => (*class* = 0) OR (*class* = 1)

Tabel 10. Aturan/*Rules data set* metode seleksi fitur VPRS dan MTF

Rule	Aturan/*Rules*
1	(cp = 2) & (fbs = 0) & (*thal* = 3) => (*class* = 0)
2	(cp = 3) & (*thalach* = 1) & (*slope* = 1) => (*class* = 0)
…	……
…	……

56	(cp = 2) & (*chol* = 3) & (fbs = 0) & (*thalach* = 2) => (*class* = 0) *OR* (*class* = 1)
57	(fbs = 0) & (*restecg* = 2) & (*thalach* = 0) & (*oldpeak* = 0) & (*slope* = 1) & (*thal* = 7) => (*class* = 0) *OR* (*class* = 1)

4.5. Evaluasi Performa

Setelah didapat aturan-aturan/*rules*, pada penelitian ini aturan-aturan dengan menggunakan metode VPRS tersebut akan dilakukan pengujian terhadap *data set* uji menggunakan perangkat lunak *Microsoft Excel* untuk mendapatkan nilai akurasi pengujian aturan keputusan. Pengujian dilakukan terhadap 3 *data set* hasil seleksi fitur dengan MTF, VPRS dan MTF & VPRS. Dari hasil pengujian *rules* maka didapatkan *confusion* matriks untuk masing-masing *data set*.

Tabel 11. Confusion matriks dari data set-data set hasil seleksi fitur

Data set hasil seleksi fitur	Jumlah *Rules*	TP	TN	FP	FN
VPRS	55	51	33	15	0
MFS	60	51	35	13	0
MFS+VPRS	57	51	33	15	0

Tabel 11 menunjukan komposisi nilai confusion matriks menggunakan penggabungan metode VPRS dengan RIPPER berupa nilai TP (*True positive*), TN (*True Negative*), FP (*False Positive*) dan FN (*False Negative*) dari 30 *data set* pengacakan. Nilai TP, TN, FP dan FN masing-masing *data set* tersebut digunakan untuk menghitung nilai akurasi, sensitivitas dan spesitivitas yang digunakan untuk menentukan diagnosis penyakit jantung koroner menggunakan metode VPRS.

Dari tabel *confusion* matriks dapat dihitung nilai akurasi setiap metode yang telah diterapkan berdasarkan *data set – data set* hasil seleksi fitur.

Tabel 12. Perbandingan nilai akurasi

Dataset hasil seleksi fitur	Nilai Akurasi (%)	Nilai Sensitivitas (%)	Nilai Spesifisitas (%)
VPRS	84,84	100	68,75
MFS	86,86	100	72,91
MFS+VPRS	84,84	100	68,75

Dalam konteks medis hanya terdapat dua kelas yaitu “sakit” atau “sehat”, dan kelas “sakit” lebih penting dibandingkan kelas “sehat”. Tujuan dari diagnosis medis adalah fokus pada peningkatan akurasi kelas “sakit” atau sensitivitas dengan tetap menjaga akurasi pada kelas “sehat” atau spesifisitas. Dalam kasus diagnosis peyakit jantung koroner sensitivitas merupakan seberapa banyak pasien yang memiliki penyakit jantung koroner diprediksi benar memiliki penyakit jantung koroner.

5. Kesimpulan

Penelitian ini memberikan skema *data mining* untuk diagnosis penyakit jantung koroner dengan melakukan seleksi fitur dan klasifikasi menggunakan *data set cleveland*.

Dari penelitian ini beberapa poin dapat disimpulkan sebagai berikut:

1. Hasil diagnosis penyakit jantung koroner dengan seleksi fitur VPRS menghasilkan peningkatan nilai akurasi dibandingkan dengan diagnosis tanpa seleksi fitur yang telah dilakukan pada penelitian sebelumnya [12].
2. Metode seleksi fitur kombinasi VPRS dan MFS, menghasilkan *Rules* lebih sedikit dibandingkan dengan MFS, sedangkan untuk nilai akurasi untuk VPRS dengan kombinasi VPRS dan MFS mempunyai nilai akurasi yang sama yaitu 84,84%.

Jadi dari hasil perbandingan proses pengujian menunjukan bahwa proses diagnosis penyakit jantung koroner dengan menggunakan metode seleksi fitur kombinasi VPRS dan MFS menunjukan nilai akurasi lebih baik dibandingkan tanpa menggunakan seleksi fitur [12] dengan jumlah *rules* dan atribut-atribut yang telah terseleksi sehingga atribut yang digunakan untuk mendiagnosis penyakit jantung koroner lebih sedikit yaitu *atribut age, cp, trestbps, chol, fbs,*

restecg, thalach, exang, oldpeak, slope, thal, dan tanpa menghilangkan atribut yang signifikan menurut pakar medis.

Referensi

[1] WHO, "Cardiovascular diseases (CVDs)," 2015. [Online]. Available: http://www.cdc.gov/heartdisease/. [Accessed: 05-Feb-2015].
[2] B. L. Zaret, M. Moser, and E. K. Hunt, *Yale University School of Medicine Heart Book*. New York, 1992.
[3] R. A. Pramunendar, I. N. Dewi, and H. Asari, "Penentuan Prediksi Awal Penyakit Jantung Menggunakan Algoritma Back Propagation Neural Network dengan Metode Adaboost," vol. 2013, no. November, pp. 298–304, 2013.
[4] T. J. Peter and K. Somasundaram, "Study and development of novel feature selection framework for heart disease prediction," *Int. J. Sci. Res. Publ.*, vol. 2, no. 10, pp. 1–7, 2012.
[5] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.
[6] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," no. 1988, pp. 63–91, 1993.
[7] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction," *Ijcst*, vol. 4333, no. 2229, pp. 304–308, 2011.
[8] W. W. Cohen, "Fast effective rule induction," *Proc. Twelfth Int. Conf. Mach. Learn.*, pp. 115–123, 1995.
[9] a H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, "HDPS: Heart disease prediction system," *2011 Comput. Cardiol.*, pp. 557–560, 2011.
[10] A. Rachman, A. B. Nurulniza, and C. P. Utomo, "Diagnosa Penyakit Jantung Menggunakan Teknik Automatic Post Pruning Decision Tree," *J. Sist. Inf.*, vol. 5, no. 2, pp. 132–137, 2014.
[11] Dwi Wahyu Prabowo, "Seleksi Fitur Berbasis Komputer Untuk Diagnosis Penyakit Jantung Koroner," University of Gadjah Mada, 2014.
[12] D. Normawati, "Diagnosis penyakit jantung koroner menggunakan penambangan data berbasis variable precision rough set (vprs) dan repeated incremental pruning to produce error reduction (ripper)," university of gajah mada, 2015.
[13] C. T. Su and J. H. Hsu, "Precision parameter in the variable precision rough sets model: An application," *Omega*, vol. 34, no. 2, pp. 149–157, 2006.
[14] R. P. Sanjaya, "Deteksi Penyakit Jantung Koroner Menggunakan Model Variable Precision Rough Set dan Logika Fuzzy," University of Gadjah Mada, 2014.
[15] B. . Tripathy, D. . Acharjya, and V. Cynthya, "A Framework for Intelligent Medical Diagnosis Using Rough Set with Formal Concept Analysis," *Int. J. Artif. Intell. Appl.*, vol. 2, no. 2, pp. 45–66, 2011.
[16] UCI, "Heart Disease Dataset," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. [Accessed: 24-Mar-2017].
[17] Fathul Ihsan and Noor Akhmad Setiawan, "Perbandingan Metode Diskretisasi Untuk Berbagai Macam Algoritma Machine Learning," University of Gadjah Mada, 2013.
[18] T. Herawan, W. Maseri, W. Mohd, and A. Noraziah, "Applying Variable Precision Rough Set for Clustering Diabetics Dataset."
[19] W. Ziarko, "Probabilistic Decision Tables in the Variable Precision Rough Set Model," *Comput. Intell.*, vol. 17, no. 3, pp. 593–603, 2001.
[20] W. Ziarko, "Variable Precision Rough Set," 1993.
[21] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, no. 1, pp. 39–59, 1993.
[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.