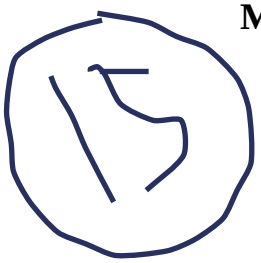
Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 9, September 2018, hlm. 2546-2554 http://j-ptiik.ub.ac.id

Seleksi Fitur *Information Gain* untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode *K-Nearest Neighbor* dan *Naïve Bayes*



Syafitri Hidayatul Annur Aini[1], Yuita Arum Sari[2], Achmad Arwan[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]syafitrihidayatul@gmail.com, [2]yuita@ub.ac.id, [3]arwan@ub.ac.id

Abstrak

Penyakit jantung merupakan salah satu penyakit tidak menular yang dapat menyebabkan kematian. Penyakit ini terjadi karena adanya penyempitan pada pembuluh darah sehingga menyebabkan fungsi jantung terganggu. Angka kematian akibat penyakit jantung diperkirakan terus meningkat oleh Kementrian Kesehatan Republik Indonesia pada tahun 2030 hingga mencapai 23,3 juta penduduk. Hal tersebut perlu diantisipasi karena jumlah dokter penyakit jantung di Indonesia masih sangat minim. Penelitian ini mengusulkan penerapan seleksi fitur *Information Gain* dengan kombinasi *K-Nearest Neighbor* (KNN) dan *Naïve Bayes* untuk mengatasi masalah efektifitas dan akurasi dalam klasifikasi penyakit jantung. Algoritme *Information Gain* digunakan untuk mengurangi dimensi atribut untuk mendapatkan atribut-atribut yang relevan. Setelah proses seleksi fitur *Information Gain* selesai, proses selanjutnya adalah melakukan klasifikasi menggunakan KNN untuk atribut-atribut numerik dan *Naïve Bayes* untuk atribut-atribut kategoris. Hasil penelitian ini menunjukkan nilai akurasi sebesar 92,31% pada saat pengujian sebaran kelas seimbang menggunakan 6 fitur dengan nilai *K*=25 dan pada saat pengujian sebaran kelas tidak seimbang menggunakan 4 fitur dengan nilai *K*=35. Berdasarkan hasil tersebut dapat disimpulkan bahwa algoritme seleksi fitur *Information Gain* dengan kombinasi KNN dan *Naïve Bayes* dapat digunakan untuk klasifikasi penyakit jantung.

Kata kunci: *penyakit jantung, seleksi fitur, Information Gain, klasifikasi, K-Nearest Neighbor, Naïve Bayes*

Abstract

Heart disease is one of the non contagious diseases that can lead to death. This disease occurs because of the narrowing of blood vessels that cause impairment of heart function. The death rate that caused by a heart disease is continuing increase and according by the Ministry of Health of the Republic of Indonesia research, in 2030 it reach 23.3 million peoples. It should be anticipated because the number of cardiologists in Indonesia is still very minimal. This research proposes framework Information Gain selection features with combination K-Nearest Neighbor and Naïve Bayes to overcome the problems on the effectiveness and accuracy in classification heart disease. Information Gain algorithm used for reduce variable dimention to get relevant variables. After Information Gain selection features process is completed, the next process is classify numeric atributes with KNN and categorical atributes with Naïve Bayes. The results of this research indicate an accuracy of 92.31% when the class distribution testing is balanced using 6 features with value of K=25 and when the class distribution testing is not balanced using 4 features with value of K=35. Based on these results, can be concluded that features selection Information Gain with combination KNN and Naïve Bayes algorithm can be used for classifying heart disease.

Keywords: *heart disease, feature selection, Information Gain, classification, K-Nearest Neighbor, Naïve Bayes*

1. PENDAHULUAN

Jantung merupakan organ penting dalam tubuh manusia yang memiliki fungsi utama untuk memompa darah ke seluruh bagian tubuh melalui pembuluh darah. (Susilawati, et al., 2014). Jika pembuluh darah mengalami penyempitan, maka fungsi jantung akan mengalami gangguan sehingga menyebabkan penyakit jantung. Penyakit ini adalah salah satu penyakit tidak menular yang dapat menyebabkan kematian. Kematian dini akibat penyakit jantung sekitar 4% terjadi di Negara berpenghasilan

tinggi dan 42% terjadi di Negara berpenghasilan rendah. Diperkirakan pada tahun 2030, angka kematian akibat penyakit jantung akan terus meningkat mencapai 23,3 juta penduduk. (Kementrian Kesehatan RI, 2014). Faktor resiko penyakit jantung antara lain, merokok, kolesterol tinggi, tekanan darah yang tinggi, diabetes, gaya hidup yang salah, pola makan yang tidak sehat dan stres. (Jabbar, Deekshatulu & Chandra, 2013).

Jumlah dokter jantung di Indonesia masih sangat minim. Pada tahun 2012 terdapat 555 orang dokter spesialis penyakit jantung dan pembuluh darah (SpJP), dimana angka tersebut masih belum cukup ideal jika dibandingkan dengan jumlah penduduk di Indonesia yang mencapai 240 juta jiwa. (Noviardi, 2012). Perkumpulan Dokter Spesialis Kardiovaskular Indonesia (PERKI) menargetkan pada tahun 2019 akan ada 1500 dokter jantung tersebar di seluruh Indonesia. (Andy, 2016). Untuk mengantisipasi keterlambatan penanganan pasien, maka diperlukan suatu sistem yang dapat membantu dokter-dokter yang kurang berpengalaman. Pada umumnya pasien disarankan mengambil sejumlah tes untuk dapat diidentifikasi penyakitnya. Dalam beberapa kasus, tidak semua tes berkontribusi terhadap diagnosis yang efektif dari sebuah penyakit. Jika data medis terdiri dari fitur yang tidak relevan dan berlebihan, maka dapat menghasilkan klasifikasi yang kurang akurat. (Jabbar, Deekshatulu & Chandra, 2013). Menurut data pada UCI *Machine Learning Repository* terdapat 13 fitur yang digunakan dalam melakukan diagnosis penyakit jantung, yaitu umur, jenis kelamin, jenis nyeri dada, tekanan darah, kadar kolestrol, kadar gula darah, hasil *electrocardiography*, rata-rata detak jantung, *exercise induced angina*, *oldpeak*, *the slope of the peak exercise ST segment*, *number of major vessels (0-3) colored by flourosopy* dan *thal*. Jumlah fitur tersebut sangat banyak sehingga dibutuhkan sebuah sistem klasifikasi penyakit jantung dengan teknik seleksi fitur untuk menghasilkan diagnosis yang lebih efektif dan akurat.

Seleksi fitur merupakan teknik untuk mengurangi dimensi atribut. Pengurangan dimensi tersebut dilakukan untuk mendapatkan atribut-atribut yang relevan dan tidak berlebihan sehingga dapat mempercepat proses klasifikasi dan dapat meningkatkan akurasi dari algoritme klasifikasi. (Arifin, 2015). Metode seleksi fitur yang digunakan dalam penelitian ini adalah *Information gain*. Metode tersebut akan melakukan proses komputasi untuk mendapatkan atribut-atribut yang paling berpengaruh terhadap *dataset* penyakit jantung. Sedangkan untuk metode klasifikasi yang akan digunakan dalam penelitian ini adalah kombinasi *K-Nearest Neighbor* (KNN) dengan *Naïve Bayes*. Pada penelitian sebelumnya metode KNN maupun *Naïve Bayes* telah diusulkan untuk klasifikasi penyakit jantung. Lestari (2014) melakukan penelitian klasifikasi untuk mendeteksi penyakit jantung menggunakan KNN. Hasil akurasi yang diperoleh dari penelitian tersebut, yaitu sebesar 70%. Penelitian tentang penyakit jantung juga dilakukan oleh (Sharmila & Indragandhi, 2017) menggunakan metode *Naïve Bayes*. Hasil akurasi yang didapatkan, yaitu sebesar 83,7%. Penelitian selanjutnya dilakukan oleh (Arifin, 2015) menggunakan metode *Information Gain* dan KNN untuk memprediksi *customer churn* Telekomunikasi. Hasil penelitian tersebut menunjukkan bahwa penggunaan seleksi fitur *Information Gain* cukup akurat untuk metode klasifikasi KNN dan menghasilkan tingkat akurasi sebesar 89,8% pada K-11.

Naïve Bayes merupakan metode klasifikasi statistik yang mudah diimplementasikan. Tetapi ada satu permasalahan yang harus diselesaikan oleh *Naïve Bayes*, yaitu saat atribut-atribut bersifat numerik karena algoritme ini harus menentukan kondisi probabilitas dari setiap nilai yang memungkinkan pada semua atribut. Untuk memperbaiki masalah tersebut, perlu dilakukan diskritisasi atribut numerik ke dalam beberapa kelas dengan mengadopsi sebuah teknik diskritisasi dari berbagai pilihan yang tersedia. Jadi, teknik yang digunakan dari diskritisasi berperan penting terhadap akurasi. (Ferdousy, Islam & Matin, 2013). Sedangkan KNN juga memiliki suatu permasalahan, dimana permasalahan tersebut berlawanan dengan kasus *Naïve Bayes*. Persoalan yang dialami metode ini berhubungan dengan atribut yang bersifat kategoris. Sebagai algoritme yang melakukan pemilihan segmen dari data latih berdasarkan jarak, skema pengukuran sebuah jarak pada data kategoris harus diperoleh. (Ferdousy, Islam & Matin, 2013).

Untuk mengatasi permasalahan dari metode KNN dan *Naïve Bayes*, terdapat penelitian yang telah dilakukan sebelumnya terkait dengan penggabungan metode KNN dan *Naïve Bayes*, yaitu penelitian yang dilakukan oleh (Ferdousy, Islam & Matin, 2013) yang menunjukkan bahwa

kombinasi antara KNN dan *Naïve Bayes* memberikan hasil yang lebih baik daripada menggunakan metode *Naïve Bayes* saja terutama dalam hal tingkat akurasi. Sebagai salah satu contoh data yang digunakan dalam penelitian tersebut adalah *dataset* tentang penyakit jantung menunjukkan tingkat akurasi sebesar 85,92%..

Pada penelitian ini kami mengusulkan metode *Information Gain* dengan kombinasi dua metode klasifikasi, yaitu KNN dan *Naïve Bayes* untuk mendapatkan hasil akurasi yang lebih tinggi pada klasifikasi penyakit jantung. Selain itu juga penggabungan dari metode KNN dan *Naïve Bayes* memiliki kelebihan, yaitu tidak perlunya melakukan diskritisasi lagi terhadap variabel yang bersifat kontinyu dan disaat yang sama juga tidak perlu lagi melakukan pengukuran jarak diantara atribut yang bersifat kategoris.

2. METODOLOGI PENELITIAN

Tahap awal seleksi fitur *Information Gain* untuk klasifikasi penyakit jantung menggunakan kombinasi metode KNN dan *Naive Bayes* adalah melakukan konversi data rekam medis yang bersifat numerik menjadi kategoris. Data yang sudah dikonversi akan diproses oleh *Information Gain* untuk mendapatkan atribut-atribut yang memiliki pengaruh yang tinggi terhadap klasifikasi penyakit jantung sehingga dapat dilakukan seleksi fitur atau pengurangan jumlah atribut yang akan dipakai dalam proses klasifikasi. Saat proses klasifikasi, data yang digunakan sebagai data latih adalah data rekam medis sebelum dikonversi. Proses yang pertama yaitu menghitung data yang bersifat numerik terlebih dahulu dengan metode KNN yang menggunakan konsep perhitungan jarak. Setelah mendapatkan hasil perhitungan KNN berupa data yang telah diurutkan dari jarak yang terkecil hingga terbesar sebanyak *K*, akan dilanjutkan dengan perhitungan data yang bersifat kategoris dengan metode *Naïve Bayes*. Alur penyelesaian masalah secara umum yang telah dijelaskan dapat dilihat pada Gambar 1.

2.1. Data

Data yang digunakan dalam penelitian ini adalah *dataset* statlog penyakit jantung yang didapat dari UCI *Machine Learning Repository*. *Dataset* statlog penyakit jantung terdiri dari 270 data dengan 13 atribut dan dua label kelas, yaitu Terkena Penyakit Jantung (TPJ) dan Tidak Terkena Penyakit Jantung (TTPJ). Atribut-atribut tersebut adalah sebagai berikut (Maspiyanti & Gatc, 2015):

1. Age: Umur
2. Sex: Jenis kelamin
3. Chest Pain Type: Jenis nyeri dada. Atribut ini memiliki empat nilai, yaitu *typical angina, atypical angina, non-anginal pain* dan *asymptomatic*
4. Resting Blood Pressure: Tekanan darah saat pasien beristirahat
5. Serum Cholestoral: Kadar kolesterol
6. Fasting Blood Sugar: Kadar gula darah. Atribut ini memiliki dua nilai, yaitu TRUE jika lebih dari 120 mg/dl dan FALSE jika kurang dari sama dengan 120 mg/dl
7. Resting Electrocardiographic Results: Kondisi *electrocardiography* pasien saat sedang beristirahat. Terdapat tiga nilai, yaitu 0 untuk kondisi normal, 1 untuk kondisi ST-T *wave abnormality* (kondisi saat gelombang *inversions* T dan atau ST meningkat ataupun menurun lebih dari 0,5 mV) dan 2 untuk kondisi saat ventricular kiri mengalami hipertropi
8. Maximum Heart Rate Achieved: Rata-rata detak jantung
9. Exercise Induced Angina: Kondisi saat pasien mengalami nyeri dada jika berolahraga
10. Oldpeak: Penurunan ST karena olahraga
11. The Slope of the Peak Exercise ST Segment: *slope* dari puncak ST setelah berolahraga. Memiliki tiga nilai antara lain, *upsloping, flat* dan *downsloping*
12. Number of Major Vessels (0-3) Colored by Flourosopy: Banyaknya pembuluh darah yang terdeteksi melalui proses pewarnaan *flourosopy*
13. Thal: Pemeriksaan thallium. Terdapat tiga nilai, yaitu normal, *fixed defect* dan *reversal defect*

2.2. *Information Gain*

Information Gain merupakan metode seleksi fitur paling sederhana dengan melakukan perangkingan atribut dan banyak digunakan dalam aplikasi kategorisasi teks, analisis data microarray dan analisis data citra. (Chormunge & Jena, 2016). *Information Gain* dapat membantu mengurangi *noise* yang disebabkan oleh fitur-fitur yang tidak relevan. *Information Gain* mendeteksi fitur-fitur yang paling banyak memiliki informasi berdasarkan kelas tertentu. Penentuan atribut terbaik dilakukan dengan

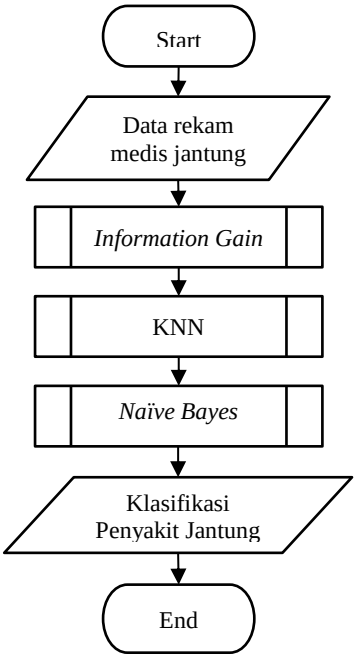
menghitung nilai *entropy* terlebih dahulu. *Entropy* merupakan ukuran ketidakpastian kelas dengan menggunakan *probabilitas* kejadian atau atribut tertentu. (Shaltout, et al., 2014). Rumus untuk menghitung *entropy* ditunjukkan pada persamaan (1). Setelah mendapatkan nilai *entropy*, maka perhitungan *Information Gain* dapat dilakukan dengan menggunakan persamaan (2). (Firmahsyah & Gantini, 2016).

$$Entropy(S) = \sum_{i}^{c} -P_i log_2 P_i \quad (1)$$

Dengan c adalah jumlah nilai yang ada pada kelas klasifikasi dan P_i merupakan jumlah sampel untuk kelas *i*.

$$Gain(S, A) = Entropy(S) - \sum_{Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

Dengan *A* merupakan atribut, *v* adalah nilai yang mungkin untuk atribut *A*, *Values(A)* adalah himpunan nilai-nilai yang mungkin untuk *A*, $|S_v|$ adalah jumlah sampel untuk nilai *v*, |S| merupakan jumlah seluruh sampel data dan *Entropy*(S_v) adalah *entropy* untuk sampel-sampel yang memiliki nilai *v*.



Gambar 1 Diagram Alir Usulan Metode

2.3. *K-Nearest Neighbor*

K-Nearest Neighbor disebut juga *lazy learner* karena berbasis pembelajaran. *K-Nearest Neighbor* menunda proses pemodelan data pelatihan sampai dibutuhkan untuk mengklasifikasikan sampel data uji. Sampel data latih dijelaskan oleh atribut-atribut numerik pada n-dimensi dan disimpan dalam ruang n-dimensi. Ketika sampel data uji (label kelas tidak diketahui) diberikan, *K-Nearest Neighbor* mencari sampel k pelatihan yang paling dekat dengan sampel data uji. (Karegowda, et al., 2012). "Kedekatan" biasanya didefinisikan dalam hal jarak metrik. Dalam penelitian ini, pengukuran jarak akan dilakukan menggunakan *euclidean distance*. Rumus *euclidean distance* direpresentasikan pada persamaan (3). (Lestari, 2014).

$$d_{(x_i,x_j)} = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2} \quad (3)$$

Keterangan:
d(xi,xj) = Jarak *euclidean*
n = Dimensi data
x_i = Data uji/testing
x_j = Data latih

Secara umum langkah-langkah untuk perhitungan KNN pada penelitian ini, yaitu:
1. Menentukan nilai *K*
2. Menghitung jarak antara data uji dengan data latih yang bersifat numerik
3. Mengurutkan jarak dari yang terkecil hingga terbesar
4. Mengambil data sebanyak *K* terdekat
5. Memilih kelas mayor

2.4. *Naïve Bayes*

Klasifikasi *Bayesian* merupakan klasifikasi statistik yang dapat memprediksi probabilitas keanggotaan kelas. Klasifikasi *Bayesian* didasarkan pada teorema *Bayes*. Klasifikasi *Bayesian* lebih dikenal sebagai klasifikasi *Naïve Bayes*. Naïve bayes berasumsi bahwa pengaruh dari nilai atribut pada kelas yang diberikan adalah saling lepas dengan nilai-nilai atribut lainnya. Hal ini dilakukan untuk menyederhanakan perhitungan yang terlibat dan dalam pengertian ini dianggap "naive". (Han, 2012). Teorema *Bayes* menyediakan cara menghitung probabilitas posterior *P*(*c*|*e*) dari *P*(*c*), *P*(*e*) dan *P*(*e*|*c*) yang ditunjukkan pada persamaan (4) dan (5). (Rahangdale, et al., 2016).

$$P(c|e) = \frac{P(e|c).P(c)}{P(e)} \quad (4)$$

$$P(c|e) = P(e_1|c) * P(e_2|c) * \ldots * P(e_n|c) * P(c) \quad (5)$$

Di mana:

$P(c|e)$ = Probabilitas posterior (c merupakan kelas dan e merupakan atribut atau event)
$P(c)$ = Probabilitas prior dari kelas
$P(e|c)$ = Probabilitas likelihood
P(e) = Probabilitas prior dari *predictor* (*event*)

Pada penelitian ini pehitungan dengan *Naïve Bayes* dilakukan berdasarkan hasil dari pehitungan KNN. Data latih yang digunakan merupakan data kategoris yang diambil berdasarkan pengurutan jarak KNN dan jumlah pemilihan tetangga terdekat dari KNN akan menjadi jumlah data latih dari *Naïve Bayes*. Secara umum langkah-langkah perhitungan *Naïve Bayes* pada penelitian ini adalah sebagai berikut:

1. Menghitung prior masing-masing kelas, yaitu dengan cara menghitung total masing-masing label kelas pada data latih dan membaginya dengan total data latih.
2. Menghitung likelihood, yaitu menghitung probabilitas masing-masing atribut
3. Menghitung posterior
4. Menentukan label kelas dengan melakukan perbandingan antar nilai posterior. Label kelas dengan nilai posterior terbesar akan menjadi label kelas data yang diuji

3. PENGUJIAN DAN ANALISIS

3.1. Pengujian

Pengujian yang dilakukan pada penelitian ini adalah pengujian akurasi dengan menggunakan jumlah fitur 3 sampai dengan 13 dan nilai K dengan kelipatan 10 yang dimulai dari 5 sampai dengan 95. Terdapat dua skenario dalam pengujian, yaitu pengujian dengan sebaran kelas seimbang dan pengujian dengan sebaran kelas tidak seimbang pada data latih. Jumlah data uji yang digunakan dalam penelitian ini adalah 26 data yang bersifat tetap, sedangkan untuk data latih berjumlah 236. Jumlah data latih tersebut dipilih karena saat melakukan perhitungan dengan metode *Information Gain* skenario dengan sebaran kelas seimbang dan sebaran kelas tidak seimbang memiliki hasil urutan atribut yang sama untuk masuk ke proses klasifikasi menggunakan KNN dan *Naïve Bayes*.

3.1.1. Pengujian Sebaran Kelas Seimbang pada Data Latih

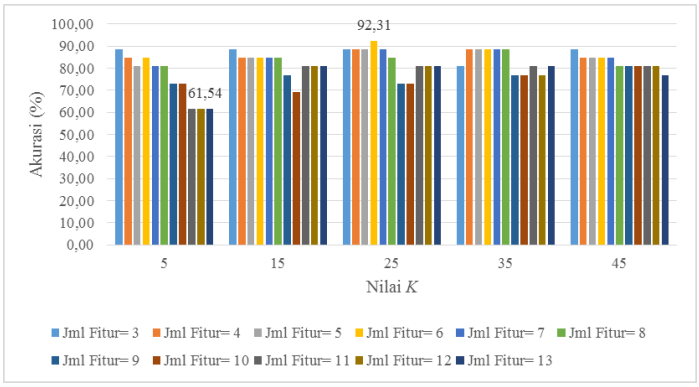
Proses pengujian sebaran kelas seimbang ini menggunakan data latih dengan label kelas TPJ berjumlah 118 dan label kelas TTPJ berjumlah 118. Sedangkan untuk data uji yang digunakan adalah 26 data dengan sebaran kelas 2 data berlabel TPJ dan 24 data berlabel TTPJ. Hasil pengujian terhadap sebaran kelas tidak seimbang pada data latih dapat dilihat pada Gambar 2 dan Gambar 3. Gambar 2 menunjukkan grafik hasil pengujian saat nilai K=5 sampai K=45 dan Gambar 3 menunjukkan grafik hasil pengujian saat nilai K=55 sampai K=95.

Berdasarkan hasil pengujian yang dilakukan pada sebaran kelas seimbang, diperoleh hasil akurasi terendah dan tertinggi yang ditunjukkan pada Gambar 2. Nilai akurasi terendah sebesar 61,54% diperoleh pada saat jumlah fitur yang digunakan adalah 11, 12 dan 13 dengan nilai K= 5 dan nilai akurasi tertinggi sebesar 92,31% diperoleh saat jumlah fitur yang digunakan adalah 6 dengan nilai K=25.

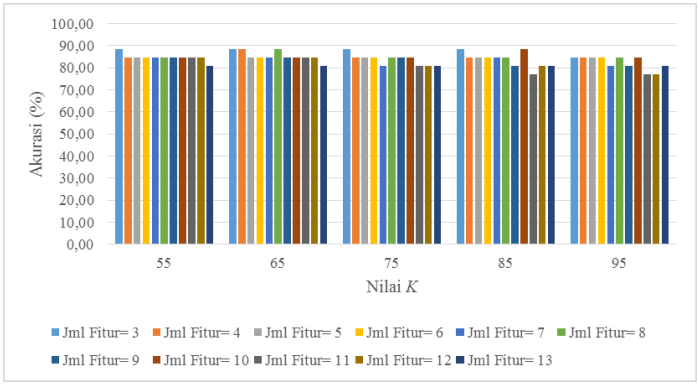
3.1.2. Pengujian Sebaran Kelas Tidak Seimbang pada Data Latih

Proses pengujian sebaran kelas tidak seimbang ini menggunakan data latih dengan label kelas TPJ berjumlah 110 dan label kelas TTPJ berjumlah 126. Sedangkan untuk data uji yang digunakan adalah 26 data dengan sebaran kelas 2 data berlabel TPJ dan 24 data berlabel TTPJ. Hasil pengujian terhadap sebaran kelas tidak seimbang pada data latih dapat dilihat pada Gambar 4 dan Gambar 5. Gambar 4 menunjukkan grafik hasil pengujian saat nilai K=5 sampai K=45 dan Gambar 5 menunjukkan grafik hasil pengujian saat nilai K=55 sampai K=95.

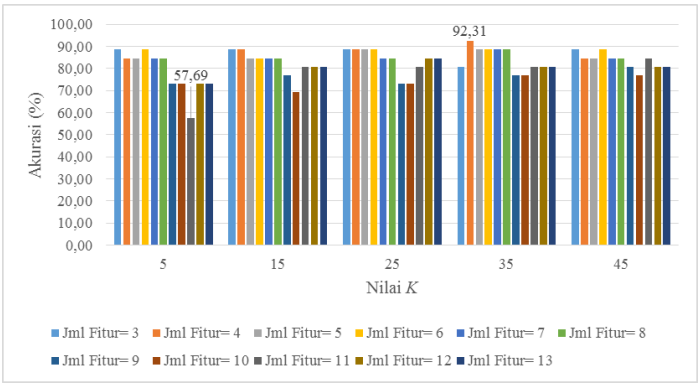
Berdasarkan hasil pengujian yang dilakukan pada sebaran kelas tidak seimbang, diperoleh hasil akurasi terendah dan tertinggi yang ditunjukkan pada Gambar 4. Nilai akurasi terendah sebesar 57,69% diperoleh pada saat jumlah fitur yang digunakan adalah 11 dengan nilai K= 5 dan nilai akurasi tertinggi sebesar 92,31% diperoleh saat jumlah fitur yang digunakan adalah 4 dengan nilai K=25.



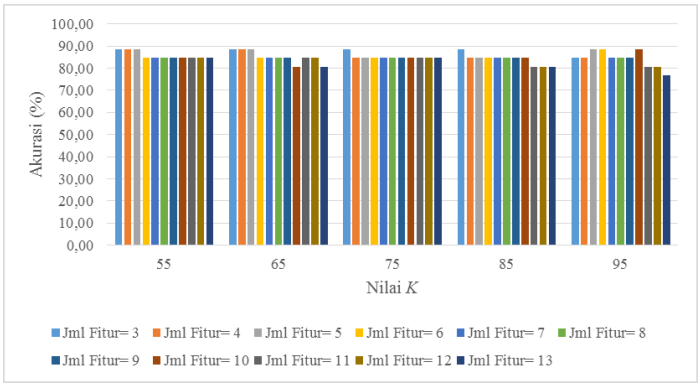
Gambar 2 Grafik Hasil Pengujian Sebaran Kelas Seimbang (*K*=5 sampai *K*=45)



Gambar 3 Grafik Hasil Pengujian Sebaran Kelas Seimbang (*K*=55 sampai *K*=95)



Gambar 4 Grafik Hasil Pengujian Sebaran Kelas Tidak Seimbang (*K*=5 sampai *K*=45)



Gambar 5 Grafik Hasil Pengujian Sebaran Kelas Tidak Seimbang (*K*=55 sampai *K*=95)

3.2. Analisis

Dalam penelitian ini nilai *K* sangat berpengaruh dalam menentukan hasil klasifikasi, karena saat prosess KNN selesai, maka nilai *K* tersebut akan digunakan untuk pembentukan model pada metode selanjutnya, yaitu *Naïve Bayes*. Nilai *K* akan menentukan jumlah data latih yang akan digunakan *Naïve Bayes* untuk melakukan perhitungan *probabilitas*, sehingga dapat menentukan hasil klasifikasi.

Tabel 1 Perbandingan Akurasi Kelas Tidak Seimbang dengan 13 fitur dan 6 fitur

Nilai *K*	Sebaran Kelas Tidak Seimbang	
	Akurasi Tanpa Menggunakan *Information Gain* (13 fitur)	Akurasi Menggunakan *Information Gain* (6 fitur)
5	73,08%	88,46%
15	80,77%	84,62%
25	84,62%	88,46%
35	80,77%	88,46%
45	80,77%	88,46%
55	84,62%	84,62%
65	80,77%	84,62%
75	84,62%	84,62%
85	80,77%	84,62%
95	76,92%	88,46%

Tabel 2 Perbandingan Akurasi Kelas Seimbang dengan 13 fitur dan 6 fitur

Nilai *K*	Sebaran Kelas Seimbang	
	Akurasi Tanpa Menggunakan *Information Gain* (13 fitur)	Akurasi Menggunakan *Information Gain* (6 fitur)
5	61,54%	84,62%
15	80,77%	84,62%
25	80,77%	92,31%
35	80,77%	88,46%
45	76,92%	84,62%
55	80,77%	84,62%
65	80,77%	84,62%
75	80,77%	84,62%
85	80,77%	84,62%
95	80,77%	84,62%

Berdasarkan pengujian yang telah dilakukan, nilai akurasi terendah yang diperoleh saat pengujian sebaran kelas seimbang dan tidak seimbang berada saat nilai *K*=5. Hal tersebut terjadi karena saat nilai *K*=5 maka secara otomatis data latih yang digunakan untuk proses *Naïve Bayes* hanya berjumlah 5 dan data latih tersebut diambil secara acak berdasarkan hasil pengurutan KNN. Nilai *K*=5 terlalu sedikit untuk dijadikan data latih *Naïve Bayes* dengan jumlah fitur yang cukup banyak, karena saat melakukan perhitungan peluang kemunculan data, terlalu banyak data yang bernilai 0 sehingga mengakibatkan kesalahan dalam hasil klasifikasi. Dalam penelitian ini, saat nilai akhir pada *Naïve Bayes* bernilai 0 maka keputusan hasil klasifikasi diambil dari kelas mayor pada KNN.

Tabel 3 Perbandingan Akurasi Kelas Seimbang dengan 13 fitur dan 4 fitur

Nilai *K*	Sebaran Kelas Seimbang	
	Akurasi Tanpa Menggunakan *Information Gain* (13 fitur)	Akurasi Menggunakan *Information Gain* (4 fitur)
5	73,08%	84,62%
15	80,77%	88.,46%
25	84,62%	88,46%
35	80,77%	92,31%
45	80,77%	84,62%
55	84,62%	88,46%
65	80,77%	88,46%
75	84,62%	84,62%
85	80,77%	84,62%
95	76,92%	84,62%

Tabel 4 Perbandingan Akurasi Seimbang dengan 13 fitur dan 4 fitur

Nilai *K*	Sebaran Kelas Seimbang	
	Akurasi Tanpa Menggunakan *Information Gain* (13 fitur)	Akurasi Menggunakan *Information Gain* (4 fitur)
5	61,54%	84,62%
15	80,77%	84,62%
25	80,77%	88,46%
35	80,77%	88,46%
45	76,92%	84,62%
55	80,77%	84,62%
65	80,77%	88,46%
75	80,77%	84,62%
85	80,77%	84,62%
95	80,77%	84,62%

Tabel 1 sampai dengan Tabel 4 menunjukkan bahwa penggunaan seleksi fitur *Information Gain* menghasilkan nilai akurasi yang lebih baik dibandingkan tanpa menggunakan *Information Gain*. Saat nilai *K*=5 akurasi yang dihasilkan sistem tanpa menggunakan *Information Gain* menunjukkan hasil yang kurang baik pada sebaran kelas seimbang maupun tak seimbang yaitu 61,54% pada sebaran kelas seimbang dan 73,08% pada sebaran kelas tidak seimbang, sedangkan saat dilakukan pengurangan fitur dengan *Information Gain* hasil akurasi sistem yang diperoleh cukup baik yaitu 84,62% bahkan mencapai 88,46% saat fitur yang digunakan berjumlah 6 dengan sebaran kelas tidak seimbang. Pengurangan fitur yang dilakukan dengan *Information Gain* juga menunjukkan nilai akurasi sistem yang cukup

stabil dengan nilai *K* yang bervariasi baik menggunakan 6 fitur maupun 4 fitur.

4. KESIMPULAN

Kesimpulan dari hasil penelitian tentang klasifikasi penyakit jantung menggunakan seleksi fitur *Information Gain* dengan kombinasi KNN dan *Naïve Bayes*, yaitu nilai akurasi tertinggi yang diperoleh saat menggunakan data latih dengan label kelas seimbang, yaitu sebesar 92,31% saat menggunakan enam fitur dengan nilai *K*=25. Enam fitur yang digunakan antara lain, *thal*, jenis nyeri dada, *flourosopy*, rata-rata detak jantung, *oldpeak* dan *exercise induced angina*. Sedangkan nilai akurasi tertinggi yang diperoleh saat menggunakan data latih dengan label kelas tidak seimbang, yaitu sebesar 92,31% saat menggunakan empat fitur dengan nilai *K*=35. Empat fitur yang digunakan antara lain, *thal*, jenis nyeri dada, *flourosopy* dan rata-rata detak jantung. Hasil tersebut menunjukkan bahwa algoritme *Information Gain* dengan kombinasi KNN dan *Naïve Bayes* dapat digunakan untuk klasifikasi penyakit jantung.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah sistem dapat dikembangkan dengan menggunakan teknik seleksi fitur lain, yaitu analisis korelasi dan melakukan *weighting* pada kelas untuk mengatasi keadaan saat posterior pada *Naïve Bayes* bernilai sama.

5. DAFTAR PUSTAKA

Andy, 2016. *Indonesia Butuh 1.500 Dokter Jantung pada 2019, Ini Alasannya.* [Online] Available at: http://liputan8.com/2016/04/16/indonesia-butuh-1-500-dokter-jantung-pada-2019-ini-alasannya/ [Accessed 05 Maret 2017].

Arifin, M., 2015. IG-KNN untuk Prediksi Customer Churn Telekomunikasi. *Jurnal SIMETRIS,* Volume 6, pp. 1-10.

Chormunge, S. & Jena, S., 2016. Efficient Feature Subset Selection Algorithm for High Dimensional Data. *International Journal of Electrical and Computer Engineering (IJECE)* , Volume 6, pp. 1880-1888.

Ferdousy, E. Z., Islam, M. M. & Matin, M. A., 2013. Combination of Naive Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. *Computer and Information Science,* Volume 6, pp. 48-56.

Han, J., 2012. *Data Mining Concepts and Techniques.* third ed. Amerika: s.n.

Jabbar, M. A., Deekshatulu, B. & Chandra, P., 2013. Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. *Global Journal Of Computer Science And Technology Neural & Artificial Intelligence,* 13(3), pp. 5-14.

Kementrian Kesehatan RI, 2014. PUSAT DATA DAN INFORMASI KEMENTRIAN KESEHATAN RI. In: *Situasi Kesehatan Jantung*. Jakarta: Kementrian Kesehatan RI, pp. 1-8.

Lestari, M., 2014. Penerapan Algoritma Klasifikasi Nearest Neighbor (K-Nn) Untuk Mendeteksi Penyakit Jantung. *Faktor Exacta*, pp. 366-371.

Lichman, M., 2013. *UCI Machine Learning Repository.* [Online] Available at: http://archive.ics.uci.edu/ml [Accessed 09 Januari 2017].

Maspiyanti, F. & Gatc, J., 2015. Diagnosa Penyakit Jantung Pada Ponsel Menggunakan Pohon Keputusan. *Teknologi Terpadu,* Volume 1, pp. 13-20.

Noviardi, A., 2012. *JUMLAH DOKTER: Indonesia butuh tambahan spesialis jantung.* [Online] Available at: http://industri.bisnis.com/read/20120423/12/73793/jumlah-dokter-indonesia-butuh-tambahan-spesialis-jantung [Accessed 05 Maret 2017].

Rahangdale, G., Ahirwar, M. M. & Motwani, D. M., 2016. Application of k-NN and Naive Bayes Algorithm in Banking and Insurance Domain. *International Journal of Computer Science,* 13(5), pp. 69-75.

Shaltout, N. A., El-Hefnawi, M., Rafea, A. & Moustafa, A., 2014. *Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts.* London, U.K, WCE.

Sharmila, S. & Indragandhi, M. P., 2017. Improved Heart Disease Prediction used Data Mining Techniques. *International Journal of Information Technology (IJIT),* 3(2), pp. 38-40.

Susilawati, Rachman, A., Nurulniza, A. B. & Utomo, C. P., 2014. Diagnosa Penyakit Jantung Menggunakan Teknik

Automatic Post Pruning Decision Tree. *Jurnal Sistem Informasi*, Volume 5, pp. 132-137.