PENERAPAN FUZZY DECISION TREE DENGAN ALGORITMA C4.5 UNTUK KLASIFIKASI PENYAKIT JANTUNG

Nur Fadilahtul Muniroh
Program Studi Ilmu Komputer
Universitas Brawijaya Malang

Candra Dewi, S.Kom., M.Sc.
Program Studi Ilmu Komputer
Universitas Brawijaya Malang

Edy Santoso, S.Si., M.Kom.
Program Studi Ilmu Komputer
Universitas Brawijaya Malang

Abstrak : Metode *fuzzy decision tree* C4.5 dapat diimplementasikan untuk klasifikasi pada data penyakit jantung, yaitu berdasarkan faktor faktor umur, tekanan darah, kolesterol, denyut jantung dan *oldpeak*. Teknik pertama yang dilakukan adalah pembentukan himpunan *fuzzy* pada data latih, kemudian pembentukan *tree* dengan algoritma C4.5 yang menghasilkan aturan-aturan. Aturan yang telah terbentuk mengalami proses pengujian dengan menggunakan metode inferensi Mamdani. Hasil dari proses defuzzifikasi pada inferensi Mamdani inilah yang digunakan untuk menentukan kelas output klasifikasi

Fuzziness Control Threshold (FCT) dan *Leaf Decision Threshold* (LDT) sangat berpengaruh terhadap *rule* yang dihasilkan. Nilai FCT yang terlalu tinggi atau nilai LDT yang terlalu rendah dapat menyebabkan turunnya akurasi. Dari hasil uji coba dengan menggunakan data latih yang bervariasai maka didapatkan tingkat akurasi yang berbeda pula. Hasil akurasi tertinggi dicapai pada nilai FCT sebesar 50% dengan nilai LDT sebesar 5% yaitu 64,07%.

Kata kunci : Penyakit jantung, *Fuzzy DecisionTree*, Algoritma *Fuzzy* C4.5

1. PENDAHULUAN

Penyakit jantung adalah sebuah kondisi yang menyebabkan jantung tidak dapat melaksanakan tugasnya dengan baik. Penyakit ini digambarkan sebagai rasa sakit di bagian dada yang berkelanjutan selama setengah jam, menjalar kearah tangan kiri dan rahang disertai sukar bernafas dan perasaan takut yang berlebihan.

Pada saat ini penyakit jantung merupakan penyebab kematian nomor satu di dunia. Pada tahun 2005 sedikitnya 17,5 juta atau setara dengan 30,0 % kematian diseluruh dunia disebabkan oleh penyakit jantung. Di Indonesia, penyakit jantung juga cenderung meningkat sebagai penyebab kematian. Data survei kesehatan rumah tangga (SKRT) tahun 1996, menunjukkan bahwa proporsi penyakit ini meningkat dari tahun ke tahun sebagai penyebab kematian (Darmojo, 1999).

Penyakit jantung dapat memberikan perbedaan pengaruh pada pasien yang berbeda untuk tingkatan penyakit yang berbeda pula. Salah satu penelitian mengenai tingkatan penyakit jantung adalah data Haberman's Survival. Data ini disumbangkan oleh David W. Aha. Dataset ini berisi data yang berasal dari penelitian mengenai klasifikasi penyakit jantung di V.A. Medical Center, Longbeach dan Cleveland Clinic Foundation. Data Haberman's survival terdiri dari 303 kasus pasien penyakit jantung. Data ini telah digunakan dalam beberapa penelitian sebelumnya. Salah satu penelitian yang menggunakan data ini adalah menggunakan metode *fuzzy expert system* yang memperoleh tingkat akurasi sebesar 94% (Adeli, 2010).

Selain metode yang disebutkan di atas, metode lain yang dapat digunakan untuk mengklasifikasi penyakit jantung pada data Haberman's survival adalah dengan menggunakan penggabungan representasi *fuzzy* dengan *decision tree*.

Klasifikasi penyakit jantung sebenarnya dapat dilakukan dengan menggunakan logika tegas. Akan tetapi hal ini sangat kaku, karena dengan adanya perubahan yang kecil saja terhadap nilai dapat mengakibatkan perbedaan kategori. Logika *fuzzy* digunakan untuk mengantisipasi hal tersebut, karena dapat memberikan toleransi terhadap nilai, sehingga dengan adanya perubahan sedikit pada nilai tidak akan memberikan perbedaan yang signifikan. Dengan memanfaatkan kelebihan logika *fuzzy* dalam toleransi terhadap hal ambigu, diharapkan dapat menjadi pendukung keputusan dalam mengklasifikasikan data penyakit jantung berdasarkan pada faktor umur, tekanan darah, kolesterol, denyut jantung dan *oldpeak*.

Metode *decision tree* sangat terkenal daripada metode klasifikasi yang lainnya, karena metode ini tidak membutuhkan pengetahuan yang lebih atau pengaturan parameter (Han dan Khamber, 2001). Akan tetapi kebanyakan metode *decision tree* menghasilkan keputusan dari data kategorial untuk proses pelatihan dan klasifikasi data-data baru. Penggunaan data numerik pada *decision tree* memungkinkan dengan membentuk partisi pada data tersebut.

Penggabungan metode *decision tree* dengan fuzzy memungkinkan untuk menggunakan nilai-nilai numerik yang dihubungkan dengan atribut kuantitatif yang masing-masing memiliki nilai derajat keanggotaan. Menurut Liang, 2005, menggunakan teknik *fuzzy* dalam *decision tree* dapat meningkatkan kemampuan atribut-atribut kuantitatif

dalam melakukan penggolongan pada saat pelatihan. Selain itu, proses pengujian menggunakan metode fuzzy inferensi Mamdani adalah untuk mendapatkan output yang baik berdasarkan nilai derajat keanggotaan masing-masing atribut.

Pada skripsi ini, algoritma yang digunakan dalam *fuzzy decision tree* adalah C4.5. Algoritma C4.5 merupakan suksesor dari ID3 menggunakan *gain ratio* untuk memperbaiki *information gain*. Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan *split info* (Khairina, 2007). Metode *fuzzy decision tree* menggunakan algoritma C4.5 ini telah dilakukan pada beberapa riset. Rumusan metode ini dituliskan pada riset yang dilakukan Ahmad Saikhu tahun 2011 pada data diabetes Indian pima dan menghasilkan tingkat akurasi 78,91%.

Berdasarkan latar belakang yang telah dikemukakan, maka judul yang diambil dalam skripsi ini adalah "**Penerapan *Fuzzy Decision Tree* dengan Algoritma C4.5 Untuk Klasifikasi Penyakit Jantung".**

2. TINJAUAN PUSTAKA

2.1. Penyakit Jantung

Penyakit jantung adalah sebuah kondisi yang menyebabkan jantung tidak dapat melaksanakan tugasnya dengan baik. Penyakit yang mengenai jantung biasa disebut sebagai penyakit *kardiovaskular*. Masalah pada jantung dibagi menjadi dua bagian, yaitu penyakit jantung dan serangan jantung (*stroke*).

Penyakit jantung yang umum dikenal dan paling banyak diderita adalah penyakit jantung koroner. Penyakit ini paling sering menyebabkan serangan jantung pada seseorang yang bisa menyebabkan kematian. Gejala-gejala yang umumnya terjadi pada penderita penyakit jantung yaitu irama jantung tak beraturan, dada tertekan seperti ditimpa beban berat, rasa sakit, terjepit atau terbakar. Rasa sakit ini bisa menjalar ke seluruh dada, bahu kiri, lengan kiri, punggung, leher bawah dan rahang leher bawah.

Penyakit jantung dapat memberikan perbedaan pengaruh pada pasien yang berbeda untuk tingkatan penyakit yang berbeda pula. Sebuah gejala juga bisa mengindikasikan beberapa penyakit jantung yang berbeda. Menurut NYHA (New York Heart Assosiation), Penyakit jantung dibagi dalam 4 kelas yaitu 1, 2, 3, dan 4 (Schulman, 2004) yaitu :

1. Kelas pertama adalah penyakit jantung kategori ringan, dimana penderita tidak mengalami sesak napas atau jantung berdebar. Jadi seakan-akan penderita baik-baik saja, tanpa keterbatasan aktifitas fisik.
2. Kelas kedua adalah penyakit jantung kategori sedang, dimana penderita sehari-hari merasa sehat tetapi begitu beraktivitas sedikit berat, seperti berlari, maka jantung terasa sesak, berdebar atau cepat lelah.
3. Kelas ketiga sudah termasuk penyakit jantung kategori berat; saat istirahat penderita merasa nyaman, tetapi saat mengerjakan pekerjaan sehari-hari kendati aktivitas itu ringan, penderita akan mengalami sesak atau muncul gejala kelemahan jantung
4. Kelas keempat atau sudah masuk kategori sangat berat, tanpa mengerjakan apa-apa pun penderita sudah menderita sesak

2.2 Data Mining

Menurut Han dan Kamber (2001), data mining merupakan solusi yang mampu menemukan kandungan informasi yang tersembunyi berupa pola dan aturan dari sekumpulan data yang besar agar mudah dipahami. Informasi yang tersembunyi ini sangat menguntungkan dari sudut pandang penelitian, bisnis dan lainnya

Teknik-teknik data mining yang paling populer antara lain :

1. *Association Rule Mining*
   Teknik data mining untuk menemukan aturan asosiasi antara suatu kombinasi *item*.
2. *Classification*
   Proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memperkirakan kelas dari suatu obyek yang labelnya tidak diketahui.
3. *Clustering*
   Proses pengelompokkan data tanpa berdasarkan kelas data tertentu. *Clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui.

Dalam skripsi ini, digunakan teknik *data mining*, yaitu klasifikasi.

2.3 Logika Fuzzy

Kata *fuzzy* merupakan kata sifat yang berarti kabur, tidak jelas. Logika *fuzzy* dikatakan sebagai logika baru yang lama, sebab ilmu tentang logika *fuzzy* modern dan metodis baru ditemukan beberapa tahun yang lalu, padahal sebenarnya konsep tentang logika *fuzzy* itu sendiri sudah ada sejak lama (Kusumadewi, 2002).

Konsep logika *fuzzy* dikembangkan oleh Prof. Lofti Zadeh pada tahun 1965. *Fuzzy* dinyatakan dalam derajat dari suatu keanggotaan dan derajat dari kebenaran. Oleh sebab itu sesuatu dapat dikatakan sebagian benar dan sebagian salah pada waktu yang sama (Kusumadewi, 2004). Logika *fuzzy* digunakan untuk menerjemahkan suatu besaran yang diekspresikan menggunakan bahasa (*linguistic*), misalkan besaran kecepatan laju kendaraan yang diekspresikan dengan pelan, agak cepat, cepat.

2.3.1 Himpunan Fuzzy

Himpunan *fuzzy* adalah himpunan elemen yang setiap elemennya memiliki derajat keanggotaan tertentu. Himpunan *fuzzy* digolongkan oleh suatu fungsi keanggotaan (*membership function*) yang memberikan nilai derajat keanggotaan tertentu kepada setiap elemen dalam himpunan tersebut. Nilai derajat keanggotaan tersebut berada pada interval [0,1]. Adapun logika crips atau logika klasik, derajat keanggotaan suatu elemen dalam suatu himpunan hanya ditentukan dengan dua nilai yaitu nol dan satu (Tangel, 2008).

Himpunan fuzzy dapat dilakukan 3 operasi, yaitu:

1. Intersection dari dua himpunan fuzzy A dan B dengan fungsi keanggotaan berturut-turut $\mu A(x)$ dan $\mu B(x)$ didefinisikan sebagai $\mu(A \cap B)(x) = min(\mu A(x), \mu B(x))$ ... (2.1)
2. Union dari dua himpunan fuzzy A dan B dengan fungsi keanggotaan berturut-turut $\mu A(x)$ dan $\mu B(x)$ didefinisikan sebagai $\mu(A \cup B)(x) = max(\mu A(x), \mu B(x))$ ... (2.2)
3. *Complement* suatu himpunan fuzzy dengan notasi $\bar{A}$ dan fungsi keanggotaan $\mu A(x)$ didefinisikan sebagai $\mu\bar{A}(x) = 1 - \mu A(x)$ ... (2.3)

2.3.2 Fungsi Keanggotaan

Fungsi keanggotaan (*membersip function*) adalah suatu kurva yang menunjukkan pemetaan titik-titik input data ke dalam nilai keanggotaannya yang memiliki interval antara 0 sampai 1. Salah satu cara yang dapat digunakan untuk mendapatkan nilai keanggotaan adalah dengan melalui pendekatan fungsi. Representasi dari fungsi keanggotaan ini dapat digambarkan dengan dua bentuk yaitu linear atau garis lurus dan kurva (Kusumadewi, 2004).

2.4 Sistem Inferensi Fuzzy Mamdani

Fuzzy mamdani merupakan salah satu metode yang sangat fleksibel dan memiliki toleransi pada data yang ada. Fuzzy Mamdani sering dikenal sebagai metode Min-Max. Metode ini diperkenalkan oleh Ebrahim Mamdani pada tahun 1975 (Lee, 2005). Fuzzy mamdani memiliki kelebihan yakni, lebih intuitif, diterima oleh banyak pihak. Berdasarkan logika fuzzy akan dihasilkan suatu model *fuzzy* mamdani yang mampu mengklasifikasi penyakit jantung. Dengan melakukan pendekatan fuzzy menghasilkan out put yang lebih dekat dengan keadaan sebenarnya.

1. Fuzzifikasi

   Fuzzifikasi adalah proses pengubahan data keanggotaan dari himpunan suatu bobot skor biasa (konvensional) ke dalam keanggotaan himpunan bilangan fuzzy. Proses fuzzifikasi memerlukan suatu fungsi keanggotaan (*membership function*) untuk mendapatkan derajat keanggotaan suatu bobot skor ke dalam suatu himpunan (kelas). Fungsi keanggotaan dibuat berdasarkan pendekatan fungsi keanggotaan (Kainz, 2003).
2. Fungsi Implikasi

   Fungsi implikasi yang digunakan pada pengambilan keputusan dengan metode Mamdani ada 2 yaitu fungsi implikasi untuk premis OR adalah Max dan fungsi implikasi untuk premis AND adalah Min.
3. Komposisi Aturan

   Metode komposisi aturan pada FIS mamdani memakai metode Max yaitu metode yang mengambil nilai Max aturan kemudian menggunakannya untuk modifikasi daerah *fuzzy* dan mengaplikasikannya ke output dengan menggunakan operator OR (union). Secara umum metode ini dapat dituliskan sebagai berikut:

   $$\mu sf[x_i] = max(\mu sf[x_i], \mu kf[x_i]) \text{ .... (2.4)}$$

   Dengan:
   $\mu sf[x_i]$ = nilai keanggotaan solusi *fuzzy* aturan ke-i
   $\mu kf[x_i]$ = nilai keanggotaan konsekuen *fuzzy* aturan ke-i
4. Defuzzifikasi

   Input dari proses defuzzifikasi adalah suatu himpunan *fuzzy* yang diperoleh dari komposisi aturan-aturan *fuzzy*, sedangkan *output* yang dihasilkan merupakan suatu bilangan pada domain himpunan *fuzzy* tersebut, sehingga jika diberikan suatu himpunan fuzzy dalam *range* tertentu, maka harus dapat diambil suatu nilai *crisp* tertentu sebagai keluarannya (Sutikno, 2000). Metode defuzzifikasi pada sistem inferensi *fuzzy* Mamdani yang digunakan adalah metode centroid.

   Metode Centroid adalah solusi *crisp* diperoleh dengan cara mengambil titik pusat daerah fuzzy, secara umum dirumuskan pada persamaan 2.5 untuk variabel kontinyu dan persamaan 2.6 untuk variabel diskrit.

   $$\mu(x) = \frac{\int_a^b x\mu(x)dx}{\int_a^b \mu(x)dx} \quad \text{... (2.5)}$$

   $$\mu(x) = \frac{\sum_{i=1}^{n} x_i \mu(x_i)}{\sum_{i=1}^{n} \mu(x_i)} \quad \text{... (2.6)}$$

   Dengan:
   x_i= nilai tiap titik sampel
   $\mu(x_i)$ = derajat keanggotaan titik sampel x_i

2.5 Fuzzy Decision Tree dengan C4.5

Fuzzy Decision tree merupakan suatu pendekatan yang sangat populer dan praktis dalam *machine learning* untuk menyelesaikan permasalahan klasifikasi yang mengalami ketidakpastian. *Fuzzy decision tree* memungkinkan

untuk menggunakan nilai-nilai *numeric-symbolic* selama konstruksi atau saat mengklasifikasikan kasus-kasus baru. Manfaat dari teori himpunan *fuzzy* dalam *decision tree* ialah meningkatkan kemampuan atribut-atribut kuantitatif. bahkan dengan menggunakan teknik *fuzzy* dapat meningkatkan ketahanan saat melakukan klasifikasi kasus-kasus baru (Romansyah,dkk, 2009).

Pada himpunan data *fuzzy* terdapat penyesuaian rumus untuk menghitung nilai *entropy* untuk atribut, *information gain, split info* dan *gain ratio* karena adanya ekspresi data *fuzzy*. Persamaan 2.7 berikut adalah persamaan untuk mencari nilai *fuzzy entropy* dari keseluruhan data.

$$Info(S) = -\sum_{i=1}^{k}\left\{\frac{freq(C_i,S)}{|S|}\times log_2\frac{freq(C_i,S)}{|S|}\right\} \quad \text{... (2.7)}$$

Dimana $info(S)$ adalah *entropy* seluruh data dengan $S = \{s_1, s_2, \ldots, s_x\}$ dan $freq(C_i, S)$ adalah frekuensi data sampel yang masuk kedalam kelas C_i. |S| adalah jumlah data sampel yang termasuk dalam S. k adalah jumlah kategori yang membagi data kedalam beberapa kelas.

$Info_{Xp}$ adalah *entropy* untuk atribut X_p dimana sampel yang masuk kedalam T yang membagi kedalam beberapa subset $T_j(j: 1-n)$ dengan atribut X_p yang dijelaskan dalam persamaan 2.8. Perhitungan $info(T_j)$ dijelaskan pada persamaan 2.9.

$$Info_{Xp}(T) = \sum_{j=1}^{n}\frac{|T_j|}{|T|}\times info(T_j) \quad \text{... (2.8)}$$

$$info(T_j) = \sum_{i=1}^{k}\left\{\frac{freq(C_i,T_j)}{|T_j|}\times log_2\frac{freq(C_i,T_j)}{|T_j|}\right\} \quad \text{... (2.9)}$$

Atribut X_p membagi data kedalam *fuzzy set* $T_j(j: 1-n)$ dan diberikan derajat kemungkinan $\mu(T_j, S_h)(h: 1-x)$. $freq(C_i, T_j)$ adalah frekuensi data sampel yang masuk kedalam kelas C_i dan termasuk kedalam subset T_j seperti pada persamaan 2.10.

$$freq(C_i,T_j) = \sum_{h=1}^{x}\{\mu(C_i,S_h)\times\mu(T_j,S_h)\} \quad \text{... (2.10)}$$

Information gain pada atribut X_p dilihat pada persamaan 2.11 berikut

$$Gain(X_p) = Info(S) - Info_{Xp}(T) \quad \text{... (2.11)}$$

Selanjutnya, untuk menghitung nilai *split info atau* rasio perolehan yang perlu diketahui pada suatu term baru dapat dilihat pada persamaan 2.12

$$Split\ Info(X_p) = -\sum_{j=1}^{n}\frac{|T_j|}{|T|}\times log_2\frac{|T_j|}{|T|} \quad \text{... (2.12)}$$

Dimana T_j adalah jumlah *membeship function* dari pemecahan *S* pada atribut X_p dan T adalah jumlah data sampel dalam S. Setelah mendapatkan nilai *Split Info*, maka *gain ratio* dapat dicari dengan persamaan 2.13 berikut.

$$Gain\ Ratio(X_p) = \frac{Gain(X_p)}{Split\ Info(X_p)} \quad \text{... (2.13)}$$

Untuk menangani masalah *missing value* pada *fuzzy decision tree* adalah menggunakan metode yang sama seperti yang dikemukakan oleh Quinlan pada pembangunan decision tree. Caranya adalah membagi rata ke semua contoh anak-anak, jika fitur yang akan diuji *missing value*. Misalkan e_k adalah distribusi merata pada semua anak cabang jika nilai dari u_k^i pada atribut X_p *missing value* (Wang, 2003).

$$\mu_{a_p^i}(u_k^i) = \frac{1}{|D_i|} \quad \text{... (2.14)}$$

Dimana, jika u_k^i *missing value* dan $|D_i|$ adalah jumlah fuzzy set pada atribut A_i.

Information gain pada data uji dengan *missing value* jelas tidak dapat memberikan informasi tentang kelas keanggotaan. Oleh karena itu penilaian kandidat atribut harus diubah, sehingga atribut yang *missing value* dihilangkan. Misalkan diberi satu set referensi E memiliki *missing value* untuk atribut X_p. Kemudian perhitungan *Information gain* untuk atribut X_p dari fuzzy diubah sebagai berikut dengan perhitungan *entropy* pada atribut yang ada nilainya.

$$Gain(X_p) = \frac{|Tr|}{|T|}\times(Info(S) - Info_{Xp}(T)) \quad \text{... (2.15)}$$

$$\alpha = \frac{|Tr|}{|T|} \quad \text{... (2.16)}$$

Dimana T_r adalah jumlah data pada subset X_p yang diketahui nilainya dan T adalah jumlah data sampel pada subset X_p.

2.6 Data Haberman's Survival

Data Haberman's Survival merupakan data hasil dari penelitian mengenai penyakit jantung di V.A. Medical Center, Long Beach and Cleveland Clinic Foundation www.archive.ics.uci.edu/ml/datasets/Heart+Disease. Data ini disumbangkan oleh David W Aha. Data Haberman's survival terdiri dari 303 kasus pasien penderita penyakit jantung. Masing-masing data terdiri dari 5 atribut yaitu umur, tekanan darah, kolesterol, denyut jantung, *oldpeak* dan status klasifikasi pasien. Status klasifikasi pasien terdiri dari 5 kelas yaitu pasien sehat (*healthy*), pasien memiliki penyakit jantung dalam kategori ringan (*sick1*), sedang (*sick2*), berat (*sick3*) dan sangat berat (*sick4*).

Fungsi keanggotaan yang digunakan untuk mengubah atribut diperoleh dari internet yang digunakan dalam jurnal dan penelitian yang pernah

dilakukan oleh Ali Adeli dan Mehdi Neshat tahun 2010. Atribut umur dibagi menjadi 4 variabel linguistik yaitu *young* (umur kurang dari 38), *middle* (diantara umur 33 dan 45 tahun), *old* (diantara umur 40 dan 58 tahun) dan *very old* (umur lebih dari 52 tahun).

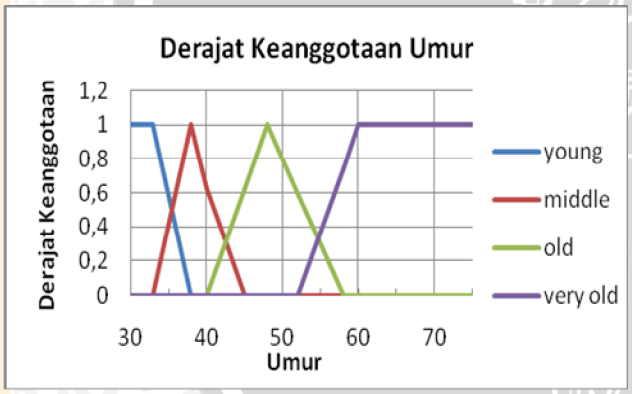
$$\mu_{young}(x) = \begin{cases} 1 & ; \quad x < 33 \\ \frac{38-x}{5} & ; 33 \leq x < 38 \\ 0 & ; \quad x \geq 38 \end{cases} \quad \ldots\ldots (2.17)$$

$$\mu_{mid}(x) = \begin{cases} 0 ; x < 33 \text{ atau } x \geq 45 \\ \frac{x-33}{5} ; 33 \leq x < 38 \\ 1 ; x = 38 \\ \frac{45-x}{7} ; 38 \leq x < 45 \end{cases} \quad \ldots\ldots (2.18)$$

$$\mu_{old}(x) = \begin{cases} 0 ; x < 40 \text{ atau } x \geq 58 \\ \frac{x-40}{8} ; 40 \leq x < 48 \\ 1 ; x = 48 \\ \frac{58-x}{10} ; 48 \leq x < 58 \end{cases} \quad \ldots\ldots (2.19)$$

$$\mu_{veryold}(x) = \begin{cases} 0 & ; \quad x < 52 \\ \frac{x-52}{8} & ; 52 \leq x < 60 \\ 1 & ; \quad x \geq 60 \end{cases} \quad \ldots\ldots (2.20)$$

Himpunan fuzzy untuk setiap variabel linguistik atribut umur menggunakan kurva berbentuk segitiga seperti pada Gambar 2.1.



Gambar 2.1 Fungsi keangotaan Umur

Atribut tekanan darah dibagi menjadi 4 variabel linguistik, yaitu *low* (tekanan darah kurang dari 134 mm Hg), *medium* (diantara tekanan darah 127 – 153 mm Hg), *high* (diantara tekanan darah 142 – 172 mm Hg) dan *very high* (tekanan darah lebih dari 154 mm Hg).

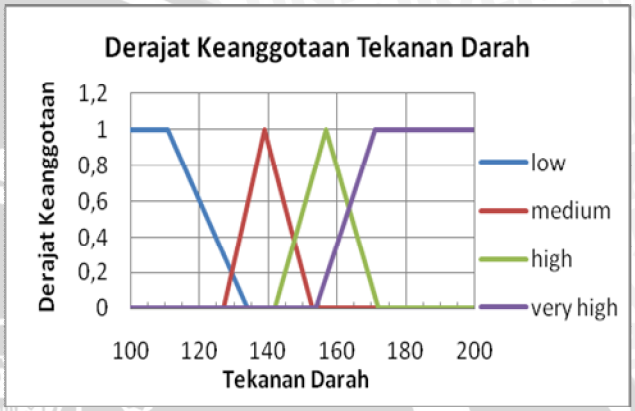
$$\mu_{low}(x) = \begin{cases} 1 & ; \quad x < 111 \\ \frac{134-x}{23} & ; 111 \leq x < 134 \\ 0 & ; \quad x \geq 134 \end{cases} \quad \ldots\ldots (2.21)$$

$$\mu_{medium}(x) = \begin{cases} 0 ; x < 127 \text{ atau } x \geq 153 \\ \frac{x-127}{12} ; 127 \leq x < 139 \\ 1 ; x = 139 \\ \frac{153-x}{14} ; 139 \leq x < 153 \end{cases} \quad \ldots\ldots (2.22)$$

$$\mu_{high}(x) = \begin{cases} 0 ; x < 142 \text{ atau } x \geq 172 \\ \frac{x-142}{15} ; 142 \leq x < 157 \\ 1 ; x = 157 \\ \frac{172-x}{15} ; 157 \leq x < 172 \end{cases} \quad \ldots\ldots (2.23)$$

$$\mu_{veryhigh}(x) = \begin{cases} 0 & ; \quad x < 154 \\ \frac{x-154}{17} & ; 154 \leq x < 171 \\ 1 & ; \quad x \geq 171 \end{cases} \quad \ldots\ldots (2.24)$$

Himpunan fuzzy untuk setiap variabel linguistik atribut tekanan darah menggunakan kurva berbentuk segitiga seperti Gambar 2.2.



Gambar 2.2 Fungsi keangotaan Tekanan Darah

Atribut kolesterol dibagi menjadi 4 variabel linguistik yaitu *low* (kolesterol kurang dari 197 mg/dL), *medium* (diantara kolesterol 188 – 250 mg/dL), *high* (diantara kolesterol 217 – 307 mg/dL) dan *very high* (kolesterol lebih dari 281 mg/dL).

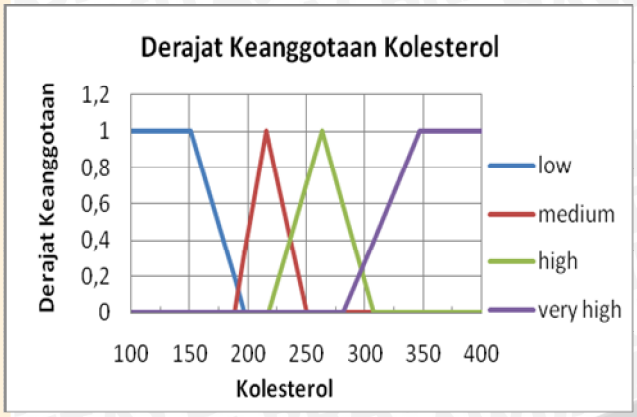
$$\mu_{low}(x) = \begin{cases} 1 & ; \quad x < 151 \\ \frac{197-x}{46} & ; 151 \leq x < 197 \\ 0 & ; \quad x \geq 197 \end{cases} \quad \ldots\ldots (2.25)$$

$$\mu_{medium}(x) = \begin{cases} 0 ; x < 188 \text{ atau } x \geq 250 \\ \frac{x-188}{27} ; 188 \leq x < 215 \\ 1 ; x = 215 \\ \frac{250-x}{35} ; 215 \leq x < 250 \end{cases} \quad \ldots\ldots (2.26)$$

$$\mu_{high}(x) = \begin{cases} 0 ; x < 217 \text{ atau } x \geq 307 \\ \frac{x-217}{46} ; 217 \leq x < 263 \\ 1 ; x = 263 \\ \frac{307-x}{44} ; 263 \leq x < 307 \end{cases} \quad \ldots\ldots (2.27)$$

$$\mu_{veryhigh}(x) = \begin{cases} 0 & ; \quad x < 281 \\ \frac{x-281}{66} & ; 281 \leq x < 347 \\ 1 & ; \quad x \geq 347 \end{cases} \quad \ldots\ldots (2.29)$$

Himpunan fuzzy untuk setiap variabel linguistik atribut kolesterol menggunakan kurva berbentuk segitiga seperti pada Gambar 2.3.



Gambar 2.3 Fungsi keangotaan Kolesterol

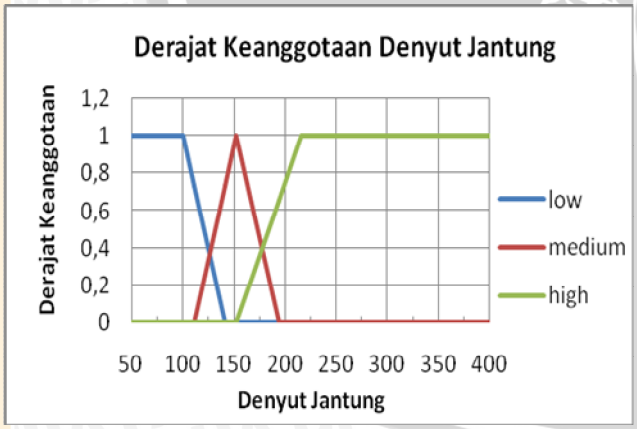
Atribut denyut jantung dibagi menjadi 3 variabel linguistik yaitu *low* (denyut jantung kurang dari 141), *medium* (diantara denyut jantung 111 – 194) dan *high* (denyut jantung lebih dari 152).

$$\mu_{low}(x) = \begin{cases} 1 & ; \quad x < 100 \\ \frac{141-x}{41} & ; 100 \leq x < 141 \\ 0 & ; \quad x \geq 141 \end{cases} \quad \ldots\ldots (2.30)$$

$$\mu_{medium}(x) = \begin{cases} 0 & ; x < 111 \text{ atau } x \geq 194 \\ \frac{x-111}{41} & ; 111 \leq x < 152 \\ 1 & ; x = 152 \\ \frac{194-x}{42} & ; 152 \leq x < 194 \end{cases} \quad \ldots\ldots (2.31)$$

$$\mu_{high}(x) = \begin{cases} 0 & ; \quad x < 152 \\ \frac{x-152}{64} & ; 152 \leq x < 216 \\ 1 & ; \quad x \geq 216 \end{cases} \quad \ldots\ldots (2.32)$$

Himpunan fuzzy untuk setiap variabel linguistik atribut denyut jantung menggunakan kurva berbentuk segitiga seperti pada Gambar 2.4.



Gambar 2.4 Fungsi keangotaan Denyut Jantung

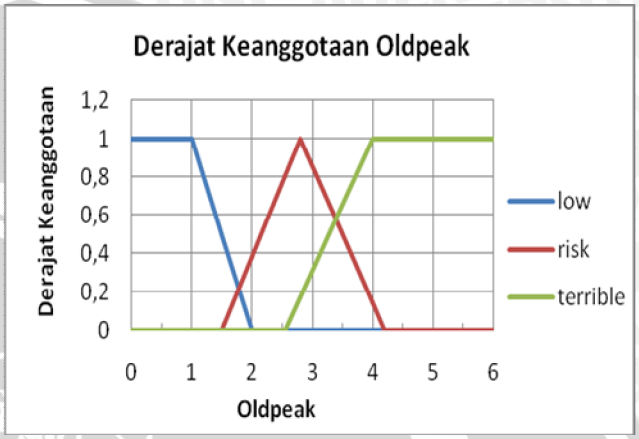
Atribut *oldpeak* dibagi menjadi 3 kelompok atau *linguistic term* yaitu *low* (*oldpeak* kurang dari 2), *risk* (diantara *oldpeak* 1.5 – 4.2) dan *terrible* (*oldpeak* lebih dari 2.55).

$$\mu_{low}(x) = \begin{cases} 1 & ; \quad x < 1 \\ \frac{2-x}{1} & ; 1 \leq x < 2 \\ 0 & ; \quad x \geq 2 \end{cases} \quad \ldots\ldots (2.33)$$

$$\mu_{risk}(x) = \begin{cases} 0 & ; x < 1.5 \text{ atau } x \geq 4.2 \\ \frac{x-1.5}{1.3} & ; 1.5 \leq x < 2.8 \\ 1 & ; x = 2.8 \\ \frac{4.2-x}{1.4} & ; 2.8 \leq x < 4.2 \end{cases} \quad \ldots\ldots (2.34)$$

$$\mu_{terrible}(x) = \begin{cases} 0 & ; \quad x < 2.55 \\ \frac{x-2.55}{1.45} & ; 2.55 \leq x < 4 \\ 1 & ; \quad x \geq 4 \end{cases} \quad \ldots\ldots (2.35)$$

Himpunan fuzzy setiap variabel linguistik atribut denyut jantung menggunakan kurva berbentuk segitiga seperti pada Gambar 2.5.



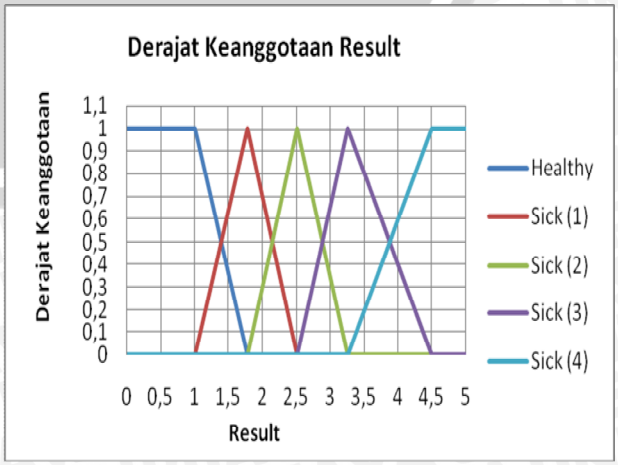
Gambar 2.5 Fungsi keangotaan Oldpeak

Variabel output dibagi menjadi 5 kelas yaitu *healty* (sehat), *s1*(kategori ringan), *s2*(kategori sedang), *s3*(kategori berat) dan *s4*(kategori sangat berat). Fuzzy set dan range untuk variabel output dapt dilihat pada Tabel 2.1.

Tabel 2.1. Klasifikasi variabel output

Output Field	Range	Fuzzy Sets
Result	<1.78	Healthy
	1-2.51	Sick (s1)
	1.78-3.25	Sick (s2)
	2.51-4.5	Sick (s3)
	3.25>	Sick (s4)

Derajat keanggotaannya dapat dilihat pada Gambar 2.6.



Gambar 2.6 Fungsi keangotaan Variabel Output

2.7 Akurasi

Akurasi adalah nilai derajat kedekatan dari pengukuran kuantitas untuk nilai sebenarnya (*true*). Nilai akurasi didapatkan dari hasil *rule* yang dihasikan dari perhitungan *decision tree* kemudian di uji coba kan pada data testing dan menghasilkan derajat keakuratan dari *rule* tersebut setelah di uji coba kan pada data testing. Tingkat akurasi diperoleh dengan perhitungan sesuai dengan persamaan 2.36 (Nugraha, 2006).

$$Akurasi\ (\%) = \frac{\sum data\ uji\ benar}{\sum total\ data\ uji} x\ 100\% \quad ...(2.36)$$

3. METODOLOGI DAN PERANCANGAN

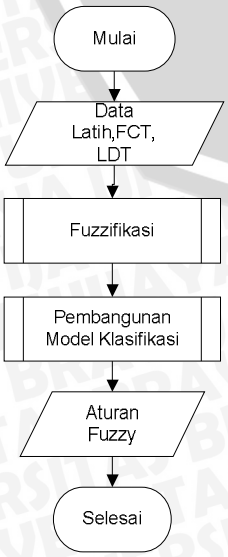
3.1 Deskripsi Umum Sistem

Sistem yang dibangun adalah sistem klasifikasi penyakit jantung yang mengimplementasikan proses pembangunan model klasifikasi menggunakan metode *fuzzy decision tree* dengan algoritma C4.5. Sistem ini bertujuan memberikan hasil klasifikasi pasien tehadap penyakit jantung berdasarkan beberapa parameter yang digunakan.

Sistem dibagi menjadi dua bagian yaitu pelatihan dan pengujian. Sistem ini akan menguji keakuratan hasil klasifikasi *dataset* terhadap data sebenarnya. Parameter uji yang digunakan untuk menganalisis metode *fuzzy decision tree* dengan algoritma C4.5 adalah akurasi klasifikasi.

3.2 Perancangan Proses

Sistem memiliki dua proses utama, yaitu proses pembentukan aturan klasifikasi dan proses pengujian. Perancangan alur sistem untuk kedua proses diatas, dapat dilihat pada Gambar 3.1 dan 3.2. Proses pembentukan aturan klasifikasi merupakan proses untuk mendapatkan sejumlah aturan *fuzzy* menggunakan metode *fuzzy decision tree* dengan algoritma C4.5. Proses ini terdiri dari 2 subproses yaitu transformasi ke data *fuzzy* (fuzzifikasi) dan pembangunan model klasifikasi.

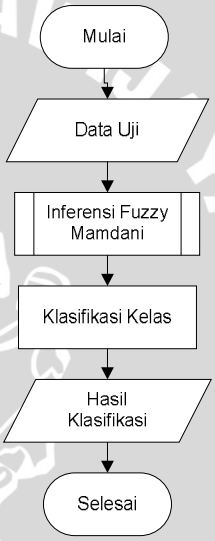


Gambar 3.1 *Flowchart* Proses Pembentukan Aturan

Berikut ini adalah penjelasan alur proses diatas:

1. Sistem mendapatkan data input berupa data latih, FCT dan LDT. Data latih terdiri dari 6 atribut yaitu umur, tekanan darah, kolesterol, denyut jantung, *oldpeak* dan kelas.
2. Mengubah data latih kedalam linguistik dan derajat keanggotaan.
3. Membangun model klasifikasi data latih dengan *fuzzy* C4.5.
4. Keluarannya berupa aturan *fuzzy*.

Setelah terbentuk aturan *fuzzy* dari proses pembentukan aturan klasifikasi, proses selanjutnya adalah proses pengujian. Proses pengujian merupakan proses untuk mengetahui tingkat keakuratan aturan yang terbentuk. Proses ini terdiri dari 2 subproses yaitu proses inferensi *fuzzy* mamdani dan proses klasifikasi kelas sebagaimana yang ditunjukan pada *flowchart* 3.3 berikut.



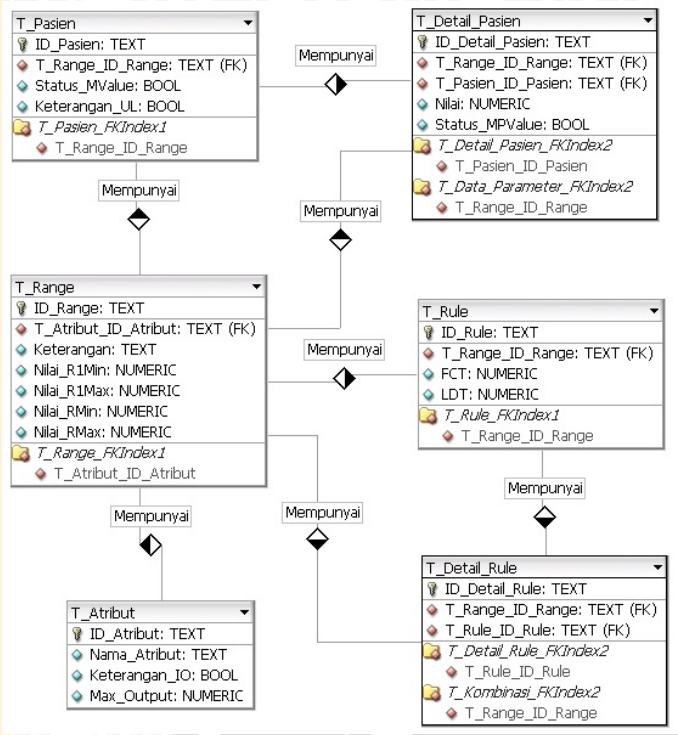
Gambar 3.2 *Flowchart* Proses Pengujian

Berikut ini adalah penjelasan alur proses diatas:

1. Sistem mendapatkan data input berupa data uji yang terdiri dari atribut umur, tekanan darah, kolesterol, denyut jantung, *oldpeak*.
2. Melakukan proses pengambilan keputusan dengan metode inferensi *fuzzy* mamdani.
3. Mengklasifikasi kelas sesuai nilai crisp yang dihasilkan dari proses inferensi *fuzzy* dalam *range* tertentu.
4. Keluarannya berupa hasil klasifikasi penyakit jantung.

3.3 Perancangan Database

Dalam penelitian ini, dibutuhkan sebuah database dalam pembentukan aturan klasifikasi. Database yang digunakan adalah Ms.Acsess yang dihubungkan dengan Delphi. Diagram relasi dari rancangan basis data ini dapat dilihat pada Gambar 3.3.



Gambar 3.3 Diagram Relasi Database

3.4 Perancangan Uji Coba

Pengujian dilakukan secara berulang dengan mengubah-ubah nilai FCT dan LDT. Dari proses pengujian akan didapatkan sejumlah aturan yang terbentuk. Pengujian akurasi menggunakan data latih yang telah ditetapkan sebelumnya. Dari aturan-aturan yang terbentuk, kemudian dilakukan perbandingan hasil klasifikasi yang dihasilkan oleh *fuzzy decision tree* C4.5 dengan data sebenarnya yang menggunakan penghitungan akurasi. Tingkat akurasi dinyatakan dalam bentuk persen (%). Perhitungannya dapat dinyatakan dengan rumus 2.36. Hasil dari perhitungan akurasi kemudian disimpan dalam Tabel 3.1

Tabel 3.14 Rancangan Pengujian Akurasi

FCT (%)	LDT (%)			

4. IMPLEMENTASI DAN PEMBAHASAN

4.1 Sistematika Pengujian

Pada pengujian tingkat akurasi, kelas output yang dihasilkan dibandingkan dengan kelas output pada data asli. Untuk mendapatkan kelas output, dilakukan proses inferensi mamdani yang disesuaikan dengan *rule-rule* yang terbentuk dalam proses pengujian sebelumnya. Pengujian ini dilakukan sebanyak 40 kali pada 3 jumlah data latih yang berbeda yaitu 70, 140 dan 210 data. Tingkat akurasi ini kemudian di rata-rata berdasarkan kombinasi nilai FCT dan LDT nya untuk mendapatkan tingkat akurasi sistem.

4.2 Analisa Hasil

Proses pengujian menggunakan 3 macam data uji dengan jumlah data yang berbeda. Akurasi yang dihasilkan memiliki nilai yang berbeda untuk setiap kombinasi nilai FCT dan LDT. Hal ini disebabkan karena pada pelatihan yang dilakukan sebelumnya menghasilkan *rule-rule* dengan kelas output yang berbeda dengan kelas output pada data uji yang sebenarnya, sehingga ketidakcocokan ini mempengaruhi nilai akurasi klasifikasi. Semakin besar ketidakcocokan dengan *rule*, maka akurasi akan semakin menurun. Hasil pengujian pada 3 macam data uji, dapat diambil nilai rata-rata akurasi klasifikasi seperti ditunjukkan pada Tabel 4.1 sebagai berikut.

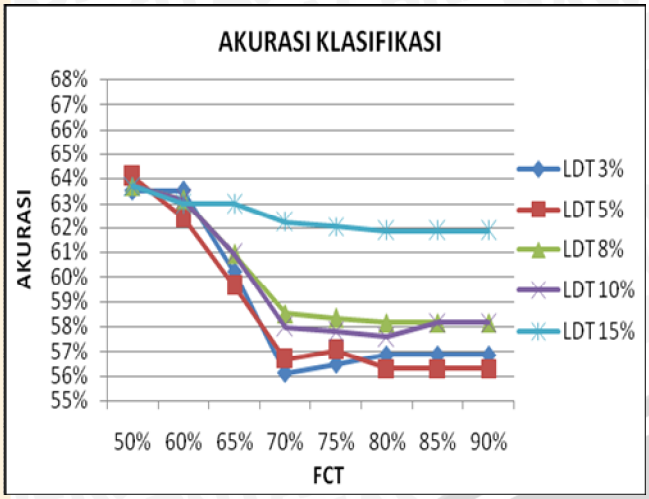
Tabel 4.1 Rata-rata Pengujian Akurasi Klasifikasi

FCT	Akurasi				
	LDT				
	3 %	5 %	8 %	10 %	15 %
50 %	63,52%	64,07%	63,70%	63,70%	63,70%
60 %	63,52%	62,41%	63,15%	63,15%	62,96%
65 %	60,19%	59,63%	60,92%	60,92%	62,96%
70 %	56,11%	56,67%	58,52%	57,96%	62,22%
75 %	56,48%	57,04%	58,33%	57,78%	62,04%
80 %	56,85%	56,30%	58,15%	57,59%	61,85%
85 %	56,85%	56,30%	58,15%	58,15%	61,85%
90 %	56,85%	56,30%	58,15%	58,15%	61,85%

Dari Tabel 4.1 diatas, dapat dilihat bahwa kinerja algoritma FC4.5 mengalami penurunan jika nilai FCT semakin besar dan atau nilai LDT yang semakin kecil, walaupun penurunan yang terjadi tidaklah signifikan sehingga masih dapat ditoleransi. Kondisi ini disebabkan karena terjadinya *overfitting*. *Overfitting* adalah terlalu tingginya nilai FCT yang digunakan pada saat pelatihan, sehingga *tree* akan terus diekspansi sampai betul-betul sesuai dengan pelatihan. Akibatnya *tree* memiliki node-node yang mengandung data yang mengalami kesalahan klasifikasi.

Nilai FCT terlalu tinggi atau nilai LDT terlalu rendah dapat menghasilkan *tree* dengan ukuran yang besar dan *rule* yang dihasilkan banyak dan bervariasi karena *tree* akan diekspansi sampai *leaf-node* terdalam atau sampai tidak ada atribut lagi. Sebaliknya, nilai FCT yang terlalu rendah dan atau nilai LDT yang terlalu tinggi akan menghasilkan *tree* dengan ukuran yang kecil sehingga *rule* yang dihasilkan juga sedikit. Hal ini terjadi karena *tree* yang sedang dibangun mengalami *pruning* atau pemotongan.

Grafik perbandingan akurasi klasifikasi menggunakan data uji dapat ditunjukkan pada Gambar 4.1 sebagai berikut.



Gambar 4.1 Grafik Tingkat Akurasi Klasifikasi

Akurasi tertinggi pada pengujian akurasi dicapai pada nilai FCT sebesar 50% dengan nilai LDT sebesar 5% yaitu 64,07%. Sedangkan akurasi terendah adalah 56,11% pada nilai FCT sebesar 70% dengan nilai LDT sebesar 3%. Akurasi yang diperoleh hanya berkisar antara 64% sampai 56%, hal ini dapat tergolong akurasi yang rendah jika dibandingkan dengan metode klasifikasi yang lain seperti *fuzzy decision tree* dengan algoritma ID3 pada data diabetes dengan akurasi sebesar 94,15%. Setelah dilakukan analisa, hal ini disebabkan karena ketidakcocokan kelas output *rule* yang terbentuk dengan kelas output data sebenarnya. Berikut ini adalah beberapa contoh *rule* yang terbentuk pada 10 data latih.

1. IF tekanan darah low AND oldpeak low THEN healthy
2. IF tekanan darah medium AND oldpeak low THEN healthy
3. IF tekanan darah high AND oldpeak low THEN sick1
4. IF tekanan darah very high AND oldpeak low THEN sick1
5. IF kolesterol low AND oldpeak risk THEN sick1
6. IF kolesterol medium AND oldpeak risk THEN sick4
7. IF kolesterol high AND oldpeak risk THEN sick3
8. IF oldpeak terrible THEN sick3

Dari *rule* diatas jika diberikan sebuah data pengujian umur 48, tekanan darah 124, kolesterol 274, denyut jantung 166, *oldpeak* 0,5 dan kelas outputnya sick3, apabila dibandingkan dengan *rule* yang terbentuk diatas maka hasil klasifikasinya harusnya masuk kelas sehat (*rule* 1). Ketidakcocokan kelas output *rule* yang terbentuk dengan data sebenarnya inilah yang menyebabkan turunnya akurasi.

Selain faktor ketidakcocokan *rule*, faktor data penyakit jantung yang digunakan untuk pelatihan dan pengujian terdapat beberapa data yang jelek, atau hasil klasifikasinya meleset jauh tidak sesuai dengan perbandingan dari kelima parameter yang digunakan. Misalnya, parameter masih tergolong kategori normal atau sehat tetapi kelas output datanya termasuk kelas sick3. Hal ini juga bisa menurunkan akurasi, karena nantinya *rule* yang terbentuk akan masuk kedalam kelas sick3, sehingga pada saat pengujian jika terdapat data yang sama atau kelas outputnya dalam kategori sehat maka kelas output klasifikasinya akan menghasilkan kelas sick3.

Terdapat beberapa hasil akurasi yang sama untuk kombinasi FCT dan LDT, seperti pada FCT 80% dan 90% dengan masing-masing kombinasi nilai LDT dari 3% sampai 15% Persamaan nilai akurasi ini disebabkan karena perbedaan *rule* yang terbentuk tidak berbeda jauh sehingga perbedaan *rule* tersebut tidak mempengaruhi proses pengujian pada data uji yang digunakan.

Dari keseluruhan uji coba, nilai FCT dan LDT sangat berpengaruh terhadap *rule* yang dihasilkan. Semakin bervariasi *rule* maka mempengaruhi besarnya akurasi pada proses pengujian. Semakin besar nilai FCT belum tentu semakin besar tingkat akurasi klasifikasi. Begitu juga dengan perubahan nilai LDT yang semakin rendah, akurasi klasifikasi juga belum tentu akurasinya tinggi.

5. PENUTUP

5.1 Kesimpulan

Setelah melakukan penelitian maka dapat disimpulkan bahwa :

1. Metode pembangkitan aturan dengan *Fuzzy C4.5* dapat diimplementasikan untuk klasifikasi pada data penyakit jantung. Teknik yang dilakukan untuk mengawali adalah dengan pembentukan himpunan *fuzzy* pada data latih, kemudian pembentukan tree dengan algoritma C4.5 dan menghasilkan aturan-aturan. Aturan yang telah terbentuk mengalami proses pengujian dengan menggunakan metode inferensi Mamdani. Hasil dari proses defuzzifikasi pada inferensi Mamdani inilah yang digunakan untuk menentukan kelas output klasifikasi.
2. Nilai FCT dan LDT sangat berpengaruh terhadap *rule* yang dihasilkan. FCT yang terlalu tinggi dapat menyebabkan turunnya nilai akurasi, begitu juga dengan nilai LDT yang terlalu rendah juga dapat menyebabkan akurasi menurun. Hal ini disebabkan *tree* terus diekspansi sampai betul-betul sesuai dengan pelatihan. Akibatnya *tree* memiliki node-node yang mengandung data yang mengalami kesalahan klasifikasi.
3. Tingkat akurasi tertinggi diperoleh sebesar 64,07% dengan nilai FCT 50% dan nilai LDT 5%. Sedangkan akurasi terendah adalah 56,11% pada nilai FCT sebesar 70% dengan nilai LDT sebesar 3%.

5.2 Saran

Akurasi yang didapat dengan menggunakan *fuzzy decision tree* algoritma C4.5 pada data haberman's survival penyakit jantung hanya sekitar 59,92%. Akurasi ini termasuk rendah jika dibandingkan dengan algoritma *fuzzy decision tree* dengan algoritma ID3 pada data diabetes yang memiliki akurasi 94,15%. Oleh karena itu disarankan untuk menambahkan metode lain dalam penelitian selanjutnya atau menggunakan metode lain yang diketahui memiliki akurasi yang bagus.

DAFTAR PUSTAKA

Adeli, Ali dan Mehdi Neshat. 2010. *A Fuzzy Expert System for Heart Disease Diagnosis*. Proceedings of the Internasional MultiConference of Engineers and Computer Scientist 2010 Vol I, IMECS 2010, March 17-19, 2010, Hongkong.

F. Romansyah, I, Sitanggang S, Nurdiati S. 2009. *Fuzzy Decision Tree dengan Algoritma ID3 pada Data Diabetes*. Internetworking Indonesia Journal. Vol.1, No.2: Special Issue on Data Mining.

Han, Jiawei dan Micheline Khamber. 2001. *Data Mining : Concepts and Technique*. Morgan Kaufmann Publisher: San Francisco, USA.

Khairina, Indah Kuntum. 2007. *Penggunaan Pohon Keputusan*. Diakses http://www.informatika.org diakses tanggal 15 November 2011.

Kantardzic, Mehmed. 2005. *Data Mining Concepts, Models, Methods and Algorithms*. IEEE Press. John Wiley & Sons, Inc.

Kusumadewi, Sri . 2002. *Analisis & Desain Sistem Fuzzy Menggunakan Toolbox MATLAB*. Yogyakarta: Graha Ilmu.

Kusumadewi, Sri dan Hari Purnomo. 2004. *Aplikasi Logika Fuzzy untuk Mendukung Keputusan*. Yogyakarta : Graha Ilmu.

Lee, Kwang H. 2005. *First Course on Fuzzy Theory and Applications*. Advanced Institute of Science and Technology, KAIST. Republic of South Korea.

Liang, G. 2005. *A Comparative Study of Three Decision Tree Algorithms: ID3, Fuzzy ID3 and Probalistic Fuzzy ID3*. Informatics & Economics Erasmus University Rotterdam, The Netherlands.

Moertini, Veronica. 2002. *Data Mining Sebagai Solusi Bisnis*. Integral, Vol.7 no.1, April.

Nugraha, Dany, dkk. 2006. *Diagnosis Gangguan Sistem Urinari pada Anjing dan Kucing Menggunakan VFI 5*. Institut Pertanian Bogor.

Schulman, Steven. 2004. *Development and Validation of a Patient Questionnaire to Determine New York Heart Association Classification*. Journal of Cardiac Failure Vol. 10 No.3 2004

Soeharto, Imam. 2004. *Penyakit jantung koroner dan serangan jantung*. Jakarta : PT. Gramedia Pustaka Utama.

Sutikno. 2000. *Perbandingan Metode Defuzzifikasi Aturan Mamdani Pada Sistem Kendali Logika Fuzzy*. Semarang: Universitas Diponegoro.

Tangel, Martin Leonard. 2008. *Sistem Penghitung Pengunjung Menggunakan Teori Pengukuran Fuzzy, Laporan Tugas Akhir*, Fakultas Ilmu Komputer, Universitas Indonesia, Depok.

Tokumaru, Masataka and Noriaki Muranaka. 2009. *Product-Impression Analysis Using Fuzzy C4.5 Decision Tree*. Journal of Advanced Computational Intelligence Vol.13 No.6, 2009 and Intelligent Informatics

Wang, Xiaomeng. 2003. *Information Measures in Fuzzy Decision Trees*. Germany : Department of Computer Science University of Magdeburg.