Perbandingan Kinerja Metode Klasifikasi *Chi-square Automatic Interaction Detection* (CHAID) dengan Metode Klasifikasi Algoritma C4.5 pada Studi Kasus : Penderita Diabetes Melitus Tipe 2 Di Samarinda Tahun 2015

Performance Comparison of Chi-square Automatic Interaction Detection (CHAID) Classification Method with C4.5 Algorithm Classification Method in Case Study: Type 2 Diabetes Mellitus Patients In Samarinda 2015

Muhammad Faisal[1], Yuki Novia Nasution[2], Fidia Deny TA[3]
[1]Mahasiswa Program Studi Statistika FMIPA Universitas Mulawarman
[2,3]Jurusan Matematika FMIPA Universitas Mulawarman
Email : muhammadfsl05@yahoo.com[1]

Abstract

C4.5 algorithm is tree classification where tree branches can be more than two. In C4.5 algorithm, the decision tree is based on entropy and gain criterias. Chi-Squared Automatic Interaction Detection (CHAID) classification method is a methods which is used to divide data to become a smaller groups based on categorical dependent and independent variables. The purpose of this research is to determine the classification process by C4.5 algorithm and CHAID method for DM type 2 patients. Risk factors for diabetes type 2 are Decline, Age, Gender, Status of Obesity, Diet, and Sports Activity based on the availability of source data from the Hospital of Abdul Wahab Sjahranie Samarinda. The results show that factors which significantly affect the DM type 2 patients are Obesity and Sport Activity. While by using CHAID, the factors which significantly affect the patients are Decline, Obesity, Diet and Sports Activity. The Classification result accuracy of the C4.5 algorithm is 90% and 94% for CHAID method.

Keywords : C4.5 algorithms, Diabetes Mellitus type 2, CHAID Method

Pendahuluan

Data adalah catatan atas kumpulan fakta. Dalam penggunaan sehari-hari data berarti suatu pernyataan yang dapat diterima. Data dapat diperoleh, disimpan, diolah, dipakai dan sebagainya. Salah satu bentuk dari pengolahan suatu data yaitu *data mining*. *Datamining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Datamining* juga merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk menguraikan, mengidentifikasi informasi yang bermanfaat, dan pengetahuan yang terakit dari berbagai *database* besar (Efraim, dkk, 2005).

Klasifikasi merupakan pengelompokkan sampel berdasarkan ciri-ciri persamaan dan perbedaan dengan menggunakan variabel terikat berupa kategori. Ada beberapa macam pengklasifikasian dalam *data mining* yaitu *decision tree*, *naivebayes*, *support vector machine* (SVM), dan lain-lain (Larose, 2005).

Decision tree (pohon keputusan) adalah pohon klasifikasi yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari permasalahan. Dalam *decision tree*,daerah pengambilan keputusan yang sebelumnya kompleks dapat diubah menjadi lebih sederhana.Banyak algoritma yang dapat digunakan dalam pembentukan *decision tree* yaitu *Iterative Dichotomiser 3* (ID3), *Classification and Regression Trees* (CART), C4.5, *Chi-square Automatic Interaction Detection* (CHAID) dan lain-lain. Algoritma adalah urutan langkah-langkah yang logis untuk menyelesaikan suatu masalah (Prasetyo, 2012).

Algoritma ID3 pertama kali diperkenalkan oleh Quinlan pada Tahun 1986 yang digunakan untuk menginduksi *decision tree*. Algoritma ID3 dapat bekerja baik pada semua fitur yang mempunyai tipe data kategorik (nominal atau ordinal). Namun dalam perkembangannya, algoritma ID3 mengalami perbaikan menjadi Algoritma C4.5. Perbaikan yang ada pada Algoritma C4.5 adalah dapat menangani fitur dengan tipe numerik (interval atau rasio), melakukan pemotongan (*pruning*) *decision tree*, dan penurunan (*deriving)rule set* (Prasetyo, 2012).

Metode CHAID merupakan salah satu tipe dari metode *Automatic Interaction Detection* (AID). Metode AID adalah salah satu teknik untuk menganalisis kelompok data berukuran besar dengan membaginya menjadi sub-sub kelompok yang tidak saling tumpah tindih yang diperuntukkan bagi data dengan variabel independen berskala rasio atau interval. Metode ini terutama dikembangkan untuk menelusuri keterkaitan struktural dalam data survei (Fielding, 1997). Metode CHAID merupakan teknik eksplorasi untuk menganalisis sekumpulan data yang berukuran besar dan cukup efisien untuk menduga variabel independen yang paling signifikan terhadap variabel dependen.

Penelitian ini terkait dengan penelitian sebelumnya Riskyawandi (2015) telah melakukan penelitian tentang metode regresi

logistik biner dan metode CHAID dengan studi kasus kepuasan pelanggan pengguna kartu Telkomsel terhadap pelayanan grapari Telkomsel di Kota Samarinda. Sementara itu, Chair (2016) telah melakukan penelitian tentang Aplikasi Klasifikasi Algoritma C4.5 dengan studi kasus Masa Studi Mahasiswa Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2008 yang mengatakan keunggulan algoritma C4.5 adalah memiliki tingkat akurasi klasifikasi yang tinggi.

Diabetes Melitus

Diabetes Melitus (DM) adalah penyakit kronis yang ditandai dengan hiperglikemia, disertai kelainan metabolik sebagai defek sekresi insulin (sel beta pankreas rusak = insulitis), atau kerja insulin terganggu, atau keduanya. Hiperglikemia kronis menyebabkan rentetan kerusakan dan disfungsi berbagai jaringan dan berbagai organ: mata, ginjal, saraf, jantung, dan pembuluh darah. Gejala hiperglikemia berat menyebabkan poliuri, polidipsi, polifagi dan berat badan menurun. Konsekuensi berat adalah ketoasidosis dan sindroma nonketotik hiperosmolar (Kosasih, 2008).

Klasifikasi Diabetes Melitus

1) Golongan 1
2) Golongan 2
3) Golongan 3
4) Golongan 4

Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh Quinlan pada Tahun 1996 sebagai versi perbaikan dari ID3. Dalam ID3, induksi *Decision tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal), sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan. Perbaikannya yaitu tidak hanya dapat menangani fitur bertipe kategorikal, tetapi juga dapat menangani fitur dengan tipe numerik, serta juga dapat melakukan pemotongan (*pruning*) *decision tree*,dan penurunan (*deriving*) *rule set*. Algoritma C4.5 juga menggunakan kriteria *gain* dalam menentukan fitur yang menjadi pemecah *node* pada pohon yang diinduksi.

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Kusrini dan Luthfy, 2009):

1. Pemilihan variabel akar

Untuk memilih variabel sebagai akar, didasarkan pada nilai *gain* tertinggi dari variabel-variabel yang ada. Berikut adalah cara untuk menghitung nilai *gain*:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)$$

dengan:

S : Himpunan kasus
S_i : Himpunan kasus pada partisi ke i
A : Variabel
n : Jumlah partisi atribut A
$|S_i|$: Jumlah kasus pada partisi ke i
$|S|$: Jumlah kasus dalam S

Sebelum mendapatkan nilai *Gain*, dicari terlebih dahulu nilai *Entropy*. *Entropy* adalah informasi mengenai proporsi pembagian kelas, nilai *entropy* berkisar mulai dari 0 sampai dengan 1, jika nilai *entropy* = 0, maka menandakan jumlah sampel hanya berada di salah satu kelas, sedangkan jika nilai *entropy* = 1, maka menandakan jumlah sampel berada di masing-masing kelas dengan jumlah yang sama. Adapun rumus dasar dari perhitungan *Entropy* adalah sebagai berikut:

$$\text{Entropy}(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

dengan:

S : Himpunan Kasus
n : Jumlah partisi S
p_i : Proporsi dari S_i terhadap S

2. Penentuan cabang untuk masing-masing nilai

Untuk penentuan cabang sama seperti mencari variabel akar yaitu didasarkan pada nilai *gain* tertinggi dari variabel-variabel yang ada.

3. Kelas dibagi dalam cabang dan apabila cabang mempunyai dua kelas maka yang dipilih kelas yang terbanyak.
4. Proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing.

Ada dua metode dalam melakukan pemangkasan dalam pohon keputusan, yaitu :

1. *Prepruning* yaitu menghentikan pembangunan suatu *subtree* lebih awal, yaitu dengan memutuskan untuk tidak lebih jauh mempartisi *data training*. Pada pendekatan *prepruning*, sebuah pohon dipangkas dengan cara menghentikan pembangunannya jika partisi yang akan dibuat dianggap tidak signifikan.
2. *Postpruning* yaitu menyederhanakan pohon dengan cara membuang beberapa cabang *subtree* setelah pohon selesai dibangun. Metode *postpruning* ini merupakan metode *standard* untuk algoritma C4.5.

Pemangkasan pohon juga dapat digunakan untuk mengatasi *overfitting*. *Overfitting* terjadi karena ada *noise data training*, yaitu data yang tidak relevan sehingga mengakibatkan pohon memiliki *subtree* yang panjang dan tidak seimbang. Misal internal node memiliki kelas YA = 5 dan TIDAK = 1. Data yang berada pada kelas TIDAK merupakan *noise*, sehingga apabila data tersebut diolah akan menghasilkan pohon dengan *subtree* yang panjang. *Overfitting* juga dapat terjadi karena *data training* yang sedikit.

Metode CHAID

CHAID adalah singkatan dari *Chi-Squared Automatic Interaction Detection*. CHAID pertama kali diperkenalkan dalam sebuah artikel yang berjudul *"An Exploratory Technique For Investigating Large Quantities Of Categorical Data"* oleh G. V. Kass tahun 1980. CHAID merupakan salah satu tipe dari metode AID (*Automatic Iteration Detection*). Metode AID adalah suatu teknik untuk menganalisis kelompok data berukuran besar dengan membaginya menjadi sub-sub kelompok yang tidak saling tumpang tindih (Kass, 1980). Teknik pemecahan (*splitting*) kelompok menjadi beberapa sub kelompok sehingga diperoleh sub-sub kelompok yang secara maksimal berbeda.

Algoritma CHAID

Algoritma CHAID digunakan untuk melakukan pemisahan dan penggabungan kategori-kategori dalam variabel yang dipakai dalam analisisnya. Magdison dalam Bagozzi (1994), menerangkan bahwa langkah-langkah analisis CHAID secara garis besar dapat dibagi menjadi tiga tahap yaitu :

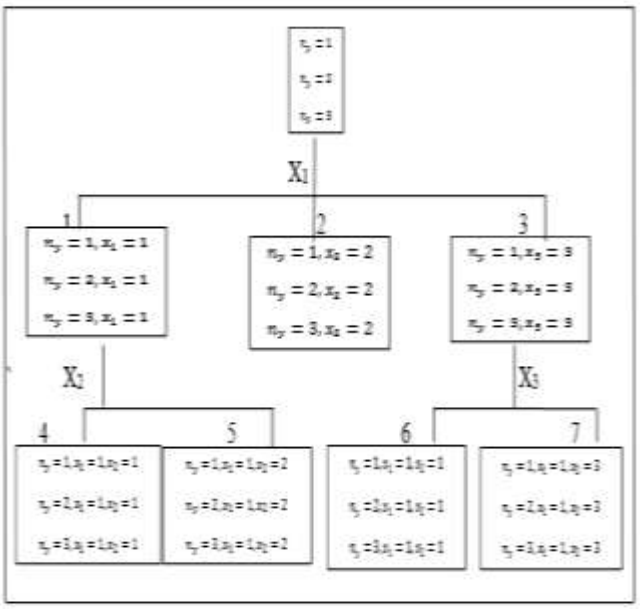
1. **Penggabungan**
   Tahap pertama dalam algoritma CHAID adalah penggabungan (*merging*). Pada tahap ini akan diperiksa signifikansi dari masing-masing kategori variabel bebas terhadap variabel terikat.
2. **Pemisahan**
   Tahap *splitting* memilih variabel bebas yang mana akan digunakan sebagai *split node* (pemisah *node*) yang terbaik. Pemilihan dikerjakan dengan membandingkan *p-value* (dari tahap *merging*) pada setiap variabel bebas.
3. **Penghentian**
   Tahap *stopping* dilakukan jika proses perumbuhan pohon harus dihentikan jika tidk ada lagi variabel bebas yang signifikan menunjukan perbedaan terhadap variabel terikat.

Diagram Pohon Klasifikasi CHAID

Hasil pembentukan segmen dalam CHAID akan ditampilkan dalam sebuah diagram pohon. Diagram pohon dimulai dari *root node* (node akar) melalui tiga tahap tersebut pada setiap *node* yang terbentuk dan secara berulang. Secara umum diagram pohon CHAID (Lehmann dan Eherler, 2001).

Metodologi Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder. Pengumpulan data sekunder adalah teknik pengumpulan data yang diperoleh dari instansi terkait. Adapun cara pengambilan data dilakukan dengan menggunakan data sampel dari data rekam medis Rumah Sakit Abdul Wahab Sjahranie Samarinda.



Gambar 1. Diagram Pohon Metode CHAID

Variabel Penelitian

Variabel dalam penelitian ini meliputi variabel independen dan variabel dependen. Variabel independen dalam penelitian ini adalah enam faktor resiko diabetes melitus tipe 2 yang yang terdiri dari

1. Riwayat Keluarga

Riwayat dalam penelitian ini menunjukkan apakah pasien memiliki garis keturunan diabetes. Riwayat keluarga dikategorikan menjadi tidak memiliki keturunan diabetes (0), dan memiliki keturunan diabetes (1).

2. Umur

Umur dalam penelitian ini memiliki skala data rasio yang merupakan umur pasien penderita Diabetes Melitus tipe 2. Umur dikategorikan menjadi:

$$\text{umur} = \begin{cases} 0, & \text{jika pasien adalah lansia} \\ 1, & \text{jika pasien adalah bukan lansia} \end{cases}$$

Pengertian lansia menurut UU No.13 Tahun 1998: Lansia adalah seseorang yang mencapai umur 60 tahun ke atas.

3. Jenis Kelamin

Jenis kelamin dalam penelitian ini terdiri dari adalah laki-laki (diberi kode 0) dan perempuan (diberi kode 1).

4. Obesitas

Obesitas memperlihatkan apakah pasien menderita obesitas atau tidak. Skala data untuk variabel obesitas ini dikategorikan menjadi tidak menderita obesitas (dikategorikan 0) dan menderita obesitas (dikategorikan menjadi 1). Untuk menentukan apakah pasien mengalami obesitas atau tidak didasarkan pada nilai *Body Mass Index* (BMI).

Tabel 1 Kategori klasifikasi nilai indeks BMI

Kategori	Nilai BMI
Underweight	BMI < 18,5
Normal	18.5 ≤ BMI< 23
Overweight	23 ≤ BMI < 25
Obese	BMI ≥ 25

untuk mendapatkan nilai indeks BMI didapatkan dengan persamaan:

$$\text{BMI} = \frac{\text{Berat Badan (Kg)}}{(\text{Tinggi Badan (m)})^2}$$

5. Pola Makan

Pola makan memperlihatkan pola makan pasien apakah sudah memenuhi dengan pola diet untuk mencegah diabetes (dikategorikan menjadi 0) dan tidak memenuhi kriteria sehat (dikategorikan menjadi 1). Kriteria yang digunakan adalah asupan kolesterol pada penderita harus kurang dari 300 mg qd. Asupan serat 25 mg/hari, dan meningkatkan serat pangan yang larut maupun tak larut, tidak mengkonsumsi suplemen niasin yang berlebihan karena dapat meningkatkan glukosa darah (Hartono, 2004). Jadwal makanan harus teratur, jumlah kalori dari makanan harus dibatasi, dan makanan dengan indeks glikemik yang tinggi harus dibatasi. Dalam penelitian ini tidak ditampilkan makanan dengan indeks glikemiknya, makanan dengan indeks glikemiknya dapat dilihat di Foster dan Miller pada buku *The Authoritative Source of Glycemic Index Values for 1,200 foods*.

6. Olah Raga

Variabel ini memberi informasi tentang keaktifan pasien apakah aktif berolahraga (dikategorikan menjadi 0) dan kurang (dikategorikan menjadi 1). Aktifitas pasien dikatakan kurang jika frekuensi olah raga dalam satu minggu sebanyak kurang dari 3 kali dengan durasi 30 menit tiap kali olah raga, dan aktif jika frekuensi olah raga dalam satu minggu sebanyak 3 kali dengan durasi 30 menit.

Sedangkan variabel dependen dalam penelitian ini adalah status pasien apakah menderita diabetes tipe 2 atau tidak. Pasien yang menderita Diabetes Melitus tipe 2 dikategorikan 1, dan pasien yang tidak menderita Diabetes Melitus tipe 2 dikategorikan 0.

Metode Analisis Data

Langkah-langkah dalam menganalisis menggunakan metode CHAIDmenggunakan α sebesar 0,05 adalah sebagai berikut :

1. Analisis CHAID

Dalam analisis menggunakan metode CHAID ada beberapa langkah-langkah yaitu sebagai berikut:

a. Penggabungan (*Merging*)

Pada tahap ini variabel independen digabungkan menjadi dua kategori dari semua kemungkinan penggabungan dan dicari untuk nilai uji *Chi-square* terkecil.

b. Pemisahan (*Splitting*)

Pada tahap ini memilih variabel independen yang akan digunakan sebagai *split node* (pemisah node) yang terbaik menggunakan α sebesar 0,05. Pemilihan dikerjakan dengan membandingkan *p-value* pada setiap variabel independen.

c. Penghentian (*Stopping*)

Tahap ini dilakukan jika proses pertumbuhan pohon harus dihentikan ketika pertumbuhan sudah mencapai batas kedalamannya maka pohon klasifikasi dihentikan.

d. Diagram Pohon

Proses metode CHAID keseluruhan untuk pembentukan segmen akan ditampilkan dalam sebuah diagram pohon.

Semua proses di dalam metode CHAID menggunakan bantuan *software* SPSS 20.

Langkah-langkah dalam menganalisis menggunakan metode C4.5 adalah sebagai berikut :

1. Penentuan *node* akar

Langkah pertama untuk membangun pohon keputusan dengan algoritma C4.5 yaitu pemilihan *node* akar, dimana menggunakan persamaan (2.2)dan (2.3). Selanjutnya penentuan cabang untuk masing-masing *node* dengan cara yang sama seperti mencari *node* akar. Untuk menentukan cabang *node* pada metode C4.5 peneliti menggunakan *software Microsoft Excel 2016*

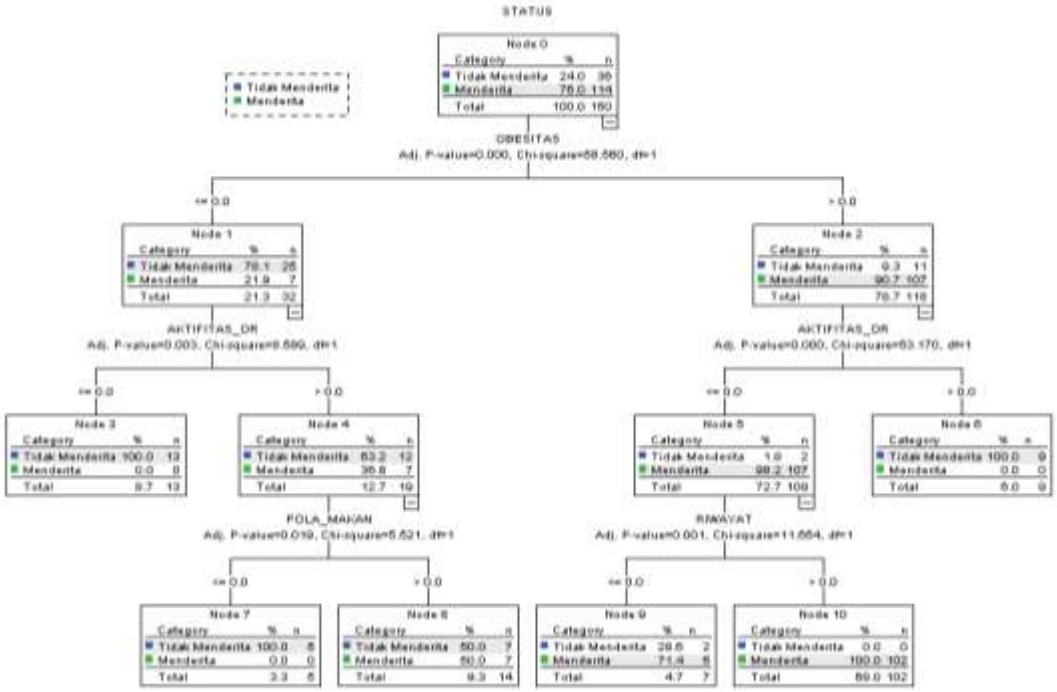
2. Penentuan pohon klasifikasi maksimal

Kelas dibagi dalam cabang dan apabila cabang mempunyai dua kelas maka yang dipilih kelas yang terbanyak dan proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing. Untuk pembentukan pohon maksimal peneliti menggunakan *software* WEKA.

Hasil dan Pembahasan

Analisis CHAID

Telah diketahui bahwa dalam penelitian ini menggunakan 6 variabel bebas dan dalam analisis ini menggunakan skala Guttman, maka hanya memiliki nilai dua kategorik, sehingga tidak terjadi tahap penggabungan.Proses metode CHAID dapat di klasifikasikan dengan diagram pohon klasifikasi CHAID (CHAID Classification Tree).Terlihat jelas pada Gambar 3. node-node serta hasil uji Chi-Square yang didapat pada setiap variabel bebas yang berperan mempengaruhi status pasien (Y), dimana terdapat variabel riwayat(X_1), obesitas (X_4), pola makan (X_5) dan aktivitas olahraga (X_6) yang mempengaruh variabel status pasien (Y)



Gambar 2. Diagram Pohon Klasifikasi Analsis CHAID

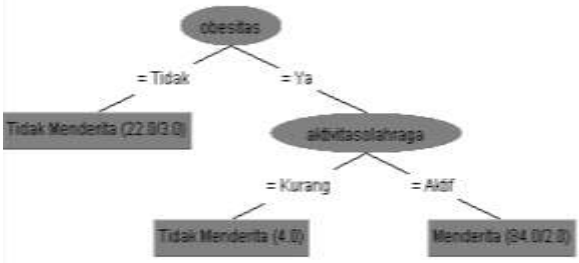
Ketepatan Klasifikasi CHAID

Tabel 2 ketepatan klasifikasi CHAID

	Persentase
APER	6 %
Ketepatan Klasifikasi	94 %

Metode C4.5

Dalam proses pembentukan pohon klasifikasi, dengan empat alur yaitu pemilihan variabel akar, penentuan cabang, kasus dibagi dalam cabang dan proses diulang sampai setiap cabang memiliki kelasyang sama. Pohon klasifikasi yang telah terbentuk akan memunculkan beberapa aturan (*rule*) sebanyak kelas yang terbentuk. Adapun data yang digunakan untuk proses pembentukan pohon klasifikasi ada 110 sampel (*data training*), sedangkan 40 sampel sisanya untuk *data testing* pohon klasifikasi yang terbentuk.Tahap pertama dalam pembentukan pohon klasifikasi adalah pemilihan *node* akar. Perhitungan untuk menentukan *node* akar menggunakan Persamaan (2.2) untuk menentukan nilai *gain* dan Persamaan (2.3) untuk menentukan nilai *entropy*. Variabel yang digunakan untuk menentukan *node* akar adalah Riwayat Keluarga (X_1), Umur (X_2), Jenis Kelamin (X_3), Obesitas (X_4), Pola makan (X_5), dan Olahraga (X_6)



Gambar 3. *Decision Tree C4.5*

Ketepatan Klasifikasi C4.5

Tabel 3 ketepatan klasifikasi C4.5

	Persentase
APER	10 %
Ketepatan Klasifikasi	90 %

Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang dilakukan, diperoleh kesimpulan sebagai berikut:

1. Faktor-faktor yang berpengaruh pada Penderita Diabetes Melitus tipe 2 di Rumah Sakit Abdul Wahab Sjahranie Samarinda tahun 2015 dengan menggunakan metode CHAID adalah Riwayat(X_1), Obesitas (X_4), Pola Makan (X_5) dan Aktivitas Olahraga (X_6).
2. Faktor-faktor yang berpengaruh pada Penderita Diabetes Melitus tipe 2 di Rumah Sakit Abdul Wahab Sjahranie Samarinda tahun 2015 dengan menggunakan metode

C4.5 adalah Obesitas (X_4)dan Aktivitas Olahraga (X_6).

3. Hasil ketepatan klasifikasi pada metode CHAID dan C4.5, diperoleh nilai persentase
4. ketepatan klasifikasi metode C4.5 sebesar 90% dan nilai persentase ketepatan klasifikasi metode CHAID sebesar 94%. Dengan demikian metode CHAID merupakan metode yang lebih baik dibandingkan algoritma C4.5 dalam pengklasifikasian data Penderita Diabetes Melitus tipe 2 di Rumah Sakit Abdul Wahab Sjahranie Samarinda tahun 2015**.**

Daftar Pustaka

Efraim, Turban, J. E. Aronson., dan T. P. Liang. (2005). *Decision Support Systems and Intelligent System.* Yogyakarta : Andi Offset.

Fielding, A. (1997). *Binary Segmentation: The Automatic Iteraction Detector and Related Technique for Exploring Data Structure.* London, New York, Toronto: John Wiley & Sons.

Kosasih, E.N. (2008). *Tafsiran Hasil Pemeriksaan Laboratorium Klinik*. Jakarta: Karisma Publising Group

Kusrini, dan E. T. Luthfy. (2009). *Algoritma Data Mining*. Yogyakarta: AndiOffset.

Larose. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*.New Jersey : John Willey & Sons.

Lehmann,T. and Eherler, D. *Responder Profiling with CHAID and Dependency Analysis*. www.informatik.uni-freiburg.de/-ml/eekdd/WS-Proceedings/w10lehmann.pdf. Tanggal akses : 12 Januari 2017

Prasetyo, E. (2012). *Data Mining : Konsep dan Aplikasi menggunakan Matlab*.Yogyakarta : Andi Offset

Riskyawandi, R. (2015). *Metode Regresi Logistik dan Metode Chi-Squared Automatic Interaction Detection (CHAID) (Studi Kasus kepuasan Pelanggan Pengguna Kartu Telkomsel Terhadap Pelayanan Grapari Telkomsel di Kota Samarinda, Kalimantan Timur)*. Eksponensial 6(1). 65-70.