ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 5, Nomor 2, Tahun 2016, Halaman 299-309
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PERBANDINGAN KLASIFIKASI PENYAKIT HIPERTENSI MENGGUNAKAN REGRESI LOGISTIK BINER DAN ALGORITMA C4.5
(Studi Kasus UPT Puskesmas Ponjong I, Gunungkidul)

Wella Rumaenda[1], Yuciana Wilandari [2], Diah Safitri[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Jurusan Statistika FSM Universitas Diponegoro
rumaendawella@yahoo.com

ABSTRACT
Hypertension is a major problem in the world today. In Indonesia prevalence of hypertension is still high. There are two types of hypertension based on cause, primary and secondary hypertension. In this thesis focused on the classification of types of hypertension based on the cause using binary logistic regression and C4.5 algorithms with case studies in UPT Puskesmas Ponjong I, Gunungkidul of October-November 2015. Binary logistic regression is a method that describes the relationship between the response variable and several predictor variables with the variable equal to 1 to declare the existence of a characteristic and the value 0 to declare the absence of a characteristic. C4.5 algorithm is one method of classification of data mining is used to create a decision tree. The predictor variables were used in this thesis are gender, age, systolic blood pressure, diastolic blood pressure, treatment history, as well as diseases and or other complaints. Based on this analysis, classification of hypertension by binary logistic regression method obtained value APER=27,4648% and 72,5352% of accuracy, while the value obtained using the algorithm C4.5 APER=35,9155% and the accuracy 64,0845 %. In two different test proportion was found that there were significant differences of the two methods.

Keywords : Types of Hypertension, Classification, C4.5 Algorithm, Biner Logistic Regression, APER

1. PENDAHULUAN
1.1 Latar Belakang
Sampai saat ini, hipertensi atau penyakit darah tinggi masih menjadi masalah utama di dunia, baik di negara maju maupun di negara berkembang. Menurut *World Health Organization* (2013), penyakit yang dijuluki *the silent killer of death* ini merupakan penyebab kematian nomor satu di dunia. Di Indonesia sendiri, menurut Riset Kesehatan Dasar (2013) bahwa prevelensi hipertensi di Indonesia terbilang masih cukup tinggi yaitu sebesar 25,8 %. Berdasarkan data UPT Puskesmas Ponjong I tahun 2015, penyakit hipertensi primer sepanjang tahun 2014, berada diurutan pertama pada sepuluh besar penyakit yang banyak diderita oleh masyarakat, yaitu sebanyak 3.112 jiwa.

Menurut Mansjoer *dkk* (2001), hipertensi sendiri tidak menunjukkan gejala tertentu. Terdapat sekitar 95% kasus hipertensi yang tidak diketahui penyebabnya, sedangkan sisanya ditimbulkan akibat adanya penyakit lain. Menurut Kementerian Kesehatan RI (2013), berdasarkan penyebabnya, hipertensi dibagi menjadi, yaitu hipertensi primer dan hipertensi sekunder. Hipertensi primer adalah suatu kondisi dimana terjadi tekanan darah tinggi yang tidak diketahui penyebabnya secara pasti. Sedangkan hipertensi sekunder merupakan suatu kondisi terjadinya tekanan darah tinggi yang penyebabnya secara spesifik diketahui seperti adanya penyakit lain.

Ada beberapa metode statistika yang dapat digunakan untuk mengetahui faktor-faktor yang mempengaruhi jenis hipertensi, diantaranya yaitu model regresi logistik biner

dan algoritma C4.5. Hosmer dan Lemeshow (2000) mengatakan bahwa model regresi logistik biner merupakan metode regresi logistik yang digunakan untuk menganalisis hubungan antara satu variabel respon dan beberapa variabel prediktor, dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik. Kusrini dan Luthfi (2009) menyebutkan, algoritma C4.5 adalah salah satu metode klasifikasi dari *data mining* yang digunakan untuk mengkonstruksikan pohon keputusan (*decision tree*).

Sehubungan dengan penelitian ini, kedua metode tersebut digunakan karena keduanya dapat mengatasi data yang bertipe kategorik. Selain itu regresi logistik biner digunakan untuk mengidentifikasi dua variabel respon bertipe kategorik dalam penelitian ini variabel hipertensi primer dan sekunder. Sementara algoritma C4.5 memiliki perhitungan sederhana dalam mengklasifikasikan jenis pernyakit hipertensi dan dapat mengatasi data bertipe kontinu.

1.2 Tujuan Penelitian

Tujuan dari penulisan penelitian ini adalah:

1. Menentukan faktor-faktor yang mempengaruhi terjadinya jenis penyakit hipertensi di UPT Puskesmas Ponjong I
2. Mendapatkan permodelan penyebab terjadinya jenis penyakit hipertensi menggunakan metode regresi logistik biner serta mengukur ketepatan klasifikasinya
3. Membentuk pohon klasifikasi menggunakan metode Algoritma C4.5 dan mengukur ketepatan klasifikasinya
4. Membandingkan ketepatan klasifikasi antara metode regresi logistik biner dengan metode Algoritma C4.5.

2. TINJAUAN PUSTAKA

2.1 Hipertensi

Menurut *World Health Organization* (2013), hipertensi merupakan suatu kondisi dimana pembuluh darah terus-menerus mengalami peningkatan tekanan. Adanya peningkatan tekanan pada pembuluh darah mengakibatkan kerja jantung untuk memompa darah semakin keras/cepat. Hipertensi juga dapat didefinisikan sebagai peningkatan tekanan sistolik lebih besar atau sama dengan 140 mmHg dan atau tekanan diastolik sama atau lebih besar 90 mmHg. Menurut Mansjoer *dkk* (2001), berdasarkan penyebabnya hipertensi dibagi menjadi dua golongan, yaitu hipertensi primer adalah hipertensi yang tidak diketahui penyebabnya secara pasti dan hipertensi sekunder adalah hipertensi yang diketahui secara spesifik penyebabnya.

2.2 Model Regresi Logistik Biner

Hosmer dan Lemeshow (2000) menyatakan bahwa regresi logistik biner digunakan untuk menjelaskan hubungan antara beberapa variabel predikor X terhadap variabel respon Y yang bersifat dikotomi atau biner. Dikotomi artinya variabel respon Y hanya bernilai 1 untuk keberadaan suatu karakteristik dan bernilai 0 untuk ketidakberadaan suatu karakteristik. Model probabilitas regresi logistik adalah:

$$\pi(x_i) = \frac{e^{\hat{g}(x_i)}}{1 + e^{\hat{g}(x_i)}}$$

dengan $\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}$

2.2.1 Estimasi Parameter

Menurut Agresti (1990),untuk menentukan estimasi parameter regresi logistik biner dapat digunakan metode *Maximum Likelihood Estimation* (MLE) yang membutuhkan turunan pertama dan turunan kedua dari fungsi likelihood. Hosmer dan Lemeshow (2000) mengatakan, pada dasarnya fungsi maksimum likelihood menggunakan estimasi nilai β untuk memaksimalkan fungsi log likelihoodnya sebagai berikut :

$$L(\beta) = \ln\{l(\beta)\} = \sum_{i=1}^{n} [y_i \ln\{\pi(x_i)\} + (1 - y_i) \ln\{1 - \pi(x_i)\}]$$

Hosmer dan Lemeshow (2000) menyeburkan, untuk mendapatkan nilai β yang memaksimalkan nilai $L(\beta)$, maka $L(\beta)$ diturunkan terhadap β_0, β_1, β_2, ..., β_p dengan hasil sama dengan nol. Sehingga hail turunan pertama dalam bentuk matriks adalah

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x}_i))$$

dengan: $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}$ dan $\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x}_i) = \begin{bmatrix} y_1 - \pi(x_1) \\ y_2 - \pi(x_2) \\ \vdots \\ y_n - \pi(x_i) \end{bmatrix}$

Oleh sebab turunan pertama dari fungsi log likelihood tidak berbentuk *closed form*, maka estimasi parameter dilakukan dengan metode numerik, sehingga dibutuhkan turunan kedua dalam bentuk matriks adalah sebagai berikut: $\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$

dengan: $\boldsymbol{V} = \begin{bmatrix} \pi(x_i)[1-\pi(x_i)] & 0 & \cdots & 0 \\ 0 & \pi(x_i)[1-\pi(x_i)] & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \pi(x_i)[1-\pi(x_i)] \end{bmatrix}$

Menurut Hastie, *et al*. (2009), untuk mendapatkan estimasi parameter digunakan metode Newton Raphson dengan langkah-langkah sebagai berikut:

a. Menentukan nilai taksiran awal untuk $\widehat{\boldsymbol{\beta}} = 0$
b. Menghitung $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x}_i))$ dan invers dari $\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$
c. Menghitung taksiran baru untuk setiap d + 1 dengan rumus:
$$\widehat{\boldsymbol{\beta}}^{(d+1)} = \widehat{\boldsymbol{\beta}}^{(d)} + \{\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\}^{-1}\{\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x}_i))\}$$
d. Proses iterasi berhenti bila didapat hasil yang konvergen, $\widehat{\boldsymbol{\beta}}^{(d+1)} \cong \widehat{\boldsymbol{\beta}}^{(d)}$

2.2.2 Uji Rasio Likelihood

Hosmer dan Lemeshow (2000) mengatakan bahwa, uji rasio likelihood merupakan uji signifikansi parameter secara keseluruhan atau bersama-sama.

Hipotesis :

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

H_1 : paling sedikit salah satu dari $\beta_j \neq 0$ dengan j = 1,2,...,p

Statistik uji : $G = -2 \ln\left(\frac{likelihood\ tanpa\ variabel\ bebas}{likelihood\ dengan\ variabel\ bebas}\right)$

dengan: ln likelihood tanpa variabel bebas $= \sum_{i=1}^{n}[y_i\, e^{\beta_0} - \ln(1 + e^{\beta_0})]$

ln likelihood dengan variabel bebas

$$= \sum_{i=1}^{n}\left[y_i\, e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} - \ln(1 + e^{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}})\right]$$

Kriteria uji : H_0 ditolak jika nilai $G > \chi^2_{(\alpha,p)}$

2.2.3 Uji Wald

Menurut Hosmer dan Lemeshow (2000), uji Wald merupakan uji signifikansi parameter untuk masing-masing variabel prediktor, yang diperoleh dengan cara mengkuadratkan rasio estimasi parameter dengan estimasi standar errornya.

Hipotesis :

$H_0 : \beta_j = 0$ dengan j= 1, 2, ..., p

$H_1 : \beta_j \neq 0$ dengan j= 1, 2, ..., p

Statistik uji : $W_j = \left\{ \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \right\}^2$

Kriteria uji : H_0 ditolak jika $W_j > \chi^2_{(\alpha,1)}$

2.2.4 Uji *Goodness of Fit*

Hosmer dan Lemeshow (2000), mengatakan bahwa uji *goodness of fit* merupakan uji yang dilakukan untuk mengetahui apakah ada perbedaan antara prediksi dan hasil observasi (model sesuai atau tidak). Statistik uji untuk uji *goodness of fit* adalah sebagai berikut :

Hipotesis :

H_0 : Model sesuai (tidak ada perbedaan antara hasil observasi dengan hasil prediksi)

H_1 : Model tidak sesuai (ada perbedaan antara hasil observasi dengan hasil prediksi)

Statistik uji : $\hat{C} = \sum_{b=1}^{g} \frac{(o_b - n'_b \bar{\pi}_b)^2}{(n'_b \bar{\pi}_b)(1-\bar{\pi}_b)}$

dimana: $o_b = \sum_{i=1}^{n'_b} y_i$ dan $\bar{\pi}_b = \sum_{i=1}^{n'_b} \frac{m_i \hat{\pi}_i}{n'_b}$

g = banyaknya grup

n'_b= banyaknya observasi pada grup ke-b

Kriteria uji: H_0 ditolak jika $\hat{C} > \chi^2_{(\alpha;g-2)}$

2.3 Algoritma C4.5

Menurut Kusrini dan Luthfi (2009), algoritma C4.5 merupakan salah satu algoritma yang dapat dipakai dalam pembentukan pohon keputusan (*decision tree*). Algoritma C4.5 diperkenalkan oleh Quinlan (1993) sebagai versi perbaikan dari Algoritma *Iterative Dichotomiser* 3 (ID3). Menurut Witten *et al*., (2011), algoritma C4.5 memiliki keunggulan dibandingkan dengan ID3 yaitu mampu mengatasi nilai yang hilang (*missing value*), mengatasi data bertipe kontinu, dan melakukan pemangkasan pohon (*prunning trees*).

Menurut Quinlan (1993), algoritma C4.5 menggunakan kriteria *gain* dalam menentukan pemecah *node* pada pohon keputusan. Rokach dan Maimon (2008) menyebutkan, *information gain* atau yang bisa disebut dengan *gain info* adalah kriteria pemisahan yang menggunakan pengukuran *entropy*. *Entropy* adalah rataan jumlah informasi yang dibutuhkan untuk mengidentifikasikan kelas pada kasus ke dalam himpunan *T*. Nilai dari setiap penghitungan *entropy* memiliki satuan *bits* atau *binary digits*. *Entropy* digunakan sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika kumpulan sampel data semakin heterogen, maka nilai *entropy*-nya semakin besar.

Menurut Ruggieri (2002), *information gain* atribut *a* dari suatu himpunan *T* dapat dihitung sebagai berikut. Jika sebuah atribut *a* adalah diskret dari suatu himpunan kasus *T* dan T_1, ..., T_s adalah sub-himpunan dari *T* yang terdiri dari kasus-kasus yang nilainya

sudah diketahui maka untuk mendapatkan *information gain* dari atribut *a* atau *Gain*(*a*) dibutuhkan *entropy* keseluruhan kelas atau *info*(*T*) dan *entropy* masing-masing atribut pada himpunan *T* atau *info*(*Ti*). Rumus dari *Gain*(*a*) adalah sebagai berikut:

$$Gain(a) = info(T) - \sum_{i=1}^{s} \frac{|T_i|}{|T|} \times info(T_i)$$

dimana nilai *entropy* keseluruhan kelas:

$$info(T) = -\sum_{j=1}^{n} \frac{freq(C_j,T)}{|T|} \times {}^{2}log\left(\frac{freq(C_j,T)}{|T|}\right)$$

sedangkan nilai *entropy* untuk setiap atribut *i*:

$$info(T_i) = -\sum_{j=1}^{n} \frac{freq(C_j,T_i)}{|T_i|} \times {}^{2}log\left(\frac{freq(C_j,T_i)}{|T_i|}\right)$$

keterangan:

$|T|$ = Banyaknya kasus dalam himpunan *T*

$|T_i|$ = Banyaknya kasus dalam sub-himpunan T_i

$freq(C_j,T)$ = Banyak dari kasus-kasus dalam himpunan *T* yang memiliki kelas C_j

Jika *a* adalah atribut kontinu maka kasus dalam *T* dengan nilai atribut tersebut diurutkan dari yang terkecil sampai terbesar. Dimisalkan nilai hasil pengurutan adalah $w_1, ..., w_m$, dan nilai $v = \frac{(w_i+w_{i+1})}{2}$ dimana $i \in [1, m\text{-}1]$ dan pemisahan yang terjadi untuk atribut bertipe kontinu adalah:

$$T_1^v = \{w_j | w_j \leq v\} \; dan \; T_2^v = \{w_j | w_j > v\}$$

Untuk setiap nilai *v*, *gain info* dari *gain* dihitung dengan mempertimbangkan prosedur pemisahan di atas. *Information gain* untuk *a* didefinisikan sebagai nilai maksimum dari semua *gain* dan nilai *v* merupakan sebagai nilai ambang batas untuk atribut kontinu.

2.4 Ketepatan Klasifikasi

Menurut Johnson dan Wichern (2007), terjadinya kesalahan klasifikasi suatu observasi merupakan hal yang sangat mungkin terjadi. Hal ini dikarenakan terkadang terdapat beberapa observasi yang tidak berasal dari kelompok tertentu tetapi dimasukkan ke dalam kelompok tersebut. Perhitungan nilai *Apparent Error Rates* (APER) dapat dilakukan dengan menggunakan matriks konfusi sebagai berikut:

Tabel 1. Matrik Konfusi

Kelompok Aktual	Kelompok Prediksi		Jumlah Observasi
	1	2	
1	n_{11}	n_{12}	n_1
2	n_{21}	n_{22}	n_2

Maka nilai APER dapat dihitung dengan rumus: $APER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$

Menurut Sugiarto (2000), untuk mendapatkan model terbaik digunakan uji beda dua proporsi dengan langkah sebagai berikut:
Hipotesis

H_0 : $PR_1 = PR_2$ (tidak ada perbedaan signifikan dari kedua metode)
H_1 : $PR_1 \neq PR_2$ (ada perbedaan signifikan dari kedua metode)
Taraf Signifikansi: $\alpha = 5\%$
Statitik Uji:

$$Z_{hitung} = \frac{PR_1 - PR_2}{\sqrt{(PR_{gab}(1-PR_{gab}) x (\frac{1}{n_1}+\frac{1}{n_2})}}$$

dengan :
PR_1 = Proporsi metode regresi logistik biner
PR_2 = Proporsi metode algoritma C4.5
PR_{gab} = Proporsi gabungan yaitu $\frac{n_1 PR_1 + n_2 PR_2}{n_1 + n_2}$
n_1 = ukuran sampel pada metode regresi logistik biner
n_2 = ukuran sampel pada metode algoritma C4.5
Kriteria Uji : H_0 ditolak jika $Z_{hitung} > Z_{\alpha/2}$ atau $Z_{hitung} < - Z_{\alpha/2}$

Jika H_0 diterima, maka tidak ada perbedaan yang signifikan antara sistem klasifikasi metode regresi logistik biner dengan metode algoritma C4.5. Sistem klasifikasi terbaik adalah sistem klasifikasi yang mempunyai nilai akurasi paling tinggi.

3. METODOLOGI PENELITIAN

Sumber data yang digunakan dalam penelitian ini adalah data sekunder yaitu data pasien hipertensi di UPT Puskesmas Ponjong 1, Gunungkidul selama bulan Oktober sampai November 2015. Jumlah populasi sebanyak 3.112 jiwa, dan jumlah sampel sebanyak 354 jiwa.

Variabel yang digunakan dalam penelitian ini adalah jenis hipertensi (primer dan sekunder) sebagai variabel respon (Y),sedangkan variabel prediktor (X) adalah jenis kelamin, umur, tekanan darah sistolik, tekanan darah diastolik, riwayat berobat (rutin dan tidak tentu), dan penyakit lain (ada dan tidak ada).

4. HASIL DAN PEMBAHASAN

4.1 Penyakit Hipertensi di UPT Puskesmas Ponjong I Bulan Oktober-November 2015

Berdasarkan data UPT Puskesmas Ponjong I, Gunungkidul bulan Oktober-November 2015 untuk penyakit hipertensi diperoleh informasi sebagai berikut:

Tabel 2. Data Pasien Hipertensi

Jenis Hipertensi	Jumlah	Persentase
Primer	266	75%
Sekunder	88	25%
Total	354	100%

Berdasarkan Tabel 2, dapat dilihat bahwa jumlah pasien yang menderita hipertensi primer lebih banyak dibandingkan dengan pasien yang menderita hipertensi sekunder. Jumlah pasien yang menderita hipertensi primer adalah sebanyak 266 jiwa atau 75% dari

keseluruhan, sedangkan sisanya 25% atau 88 jiwa merupakan pasien yang menderita hipertensi sekunder.

4.2 Pengklasifikasian Regresi Logistik Biner

Untuk membagi data *training* dan *testing* dilakukan beberapa kali percobaan dengan melihat hasil akurasi yang paling tinggi. Pada penelitian ini, data dipartisi 60% untuk data *training* atau sebanyak 212 data dan 40% data *testing* atau sebanyak 142 data.

Dari hasil pengolahan data *training* diperoleh model regresi logistik sebagai berikut:
Model awal :

$$\pi(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

dengan: $\hat{g}(x) = -2{,}404 - 0{,}877(X_{11}) + 0{,}021(X_2) + 0{,}025(X_3) - 0{,}013(X_4) + 0{,}433(X_{51}) - 0{,}749(X_{61})$

Setelah dilakukan uji sigifikansi parameter, ada tiga variabel yang tidak mempengaruhi model yaitu variabel umur, tekanan darah diastolik, dan riwayat berobat, sehingga perlu dilakukan uji signifikansi parameter yang kedua untuk menguji variabel jenis kelamin, tekanan darah sistolik, dan penyakit lain.

4.2.1 Uji Rasio Likelihood

Tabel 3. Uji Rasio Likelihood

Uji ke-	G	Nilai Tabel Chi-Square	Keputusan
1	20,911	$\chi^2_{(0,05;6)} = 12{,}592$	H_0 ditolak
2	15,186	$\chi^2_{(0,05;3)} = 7{,}815$	H_0 ditolak

Berdasarkan Tabel 3, diketahui bahwa pada taraf signifikansi 5%, variabel prediktor secara bersama-sama mempengaruhi model.

4.2.2 Uji Wald

Tabel 4. Uji Wald Pertama

Variabel	Wald		Keputusan
X_{11}	5,637	3,841	H_0 ditolak
X_2	2,558	3,841	H_0 diterima
X_3	5,497	3,841	H_0 ditolak
X_4	0,676	3,841	H_0 diterima
X_{51}	1,426	3,841	H_0 diterima
X_{61}	4,492	3,841	H_0 ditolak

Tabel 5. Uji Wald Kedua

Variabel	Wald	$\chi^2_{(0.05;1)}$	Keputusan
X_{11}	4,525	3,841	H_0 ditolak
X_3	5,834	3,841	H_0 ditolak
X_{61}	3,925	3,841	H_0 ditolak

Berdasarkan Tabel 4 dan Tabel 5, dapat dilihat bahwa pada taraf signifikansi 5%, variabel prediktor yang signifikan mempengaruhi model adalah X_{11} (jenis kelamin), X_3

(tekanan darah sistolik), dan X_{61} (penyakit lain). Sedangkan variabel X_2 (umur), X_4 (tekanan darah diastolik), dan X_{51} (riwayat berobat) tidak signifikan mempengaruhi model.

4.2.3 Uji *Goodness of Fit*

Berdasarkan uji *goodness of fit* diperoleh hasil bahwa pada taraf signifikansi 5%, H_0 diterima karena nilai $\hat{C} = 12{,}714 < \chi^2_{(0{,}05;8)} = 15{,}507$. Jadi model regresi logistik biner yang terbentuk sesuai.

4.2.4 Model Akhir

Setelah dilakukan uji signifikansi terhadap model baik secara keseluruhan maupun individu, diperoleh model akhir sebagai berikut:

$$\pi(x) = \frac{\exp(-1{,}852 - 0{,}759(X_{11}) + 0{,}022(X_3) - 0{,}688(X_{61})}{1 + \exp(-1{,}852 - 0{,}759(X_{11}) + 0{,}022(X_3) - 0{,}688(X_{61})}$$

4.2.5 Ketepatan Klasifikasi

Pada penyusunan tabel klasifikasi, maka perlu dihitung probabilitas dari data *testing*, sehingga didapatkan matriks konfusi sebagai berikut:

Tabel 6. Hasil Ketepatan Klasifikasi Regresi Logistik Biner

	Prediksi	
Observasi	Sekunder	Primer
Sekunder	1	0
Primer	39	102

Berdasarkan Tabel 6, dapat diketahui bahwa pada pengklasifikasian jenis hipertensi di UPT Puskesmas Ponjong I, Gunungkidul didapatkan nilai APER sebesar 27,4648% dengan ketepatan klasifikasi sebesar 72,5352%.

4.3 Pengklasifikasian Algoritma C4.5

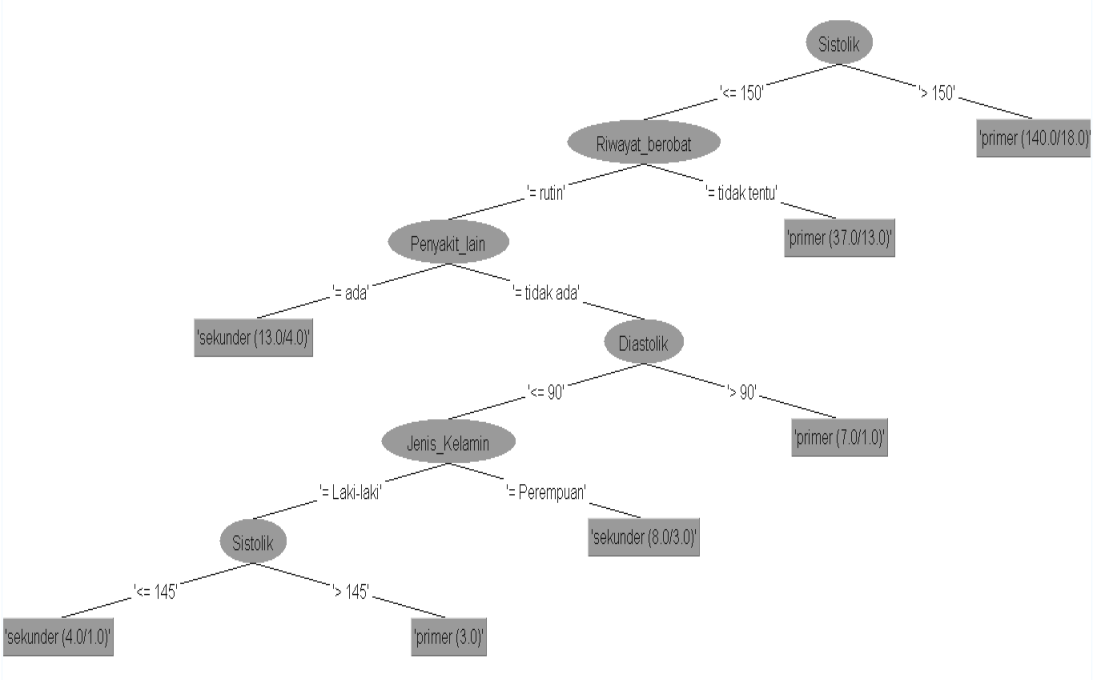
4.3.1 Pembentukan Pohon Keputusan Algoritma C4.5

Pembentukan pohon keputusan menghasilkan sebanyak 13 simpul, yang terdiri dari satu simpul akar (tekanan darah sistolik), 5 simpul keputusan, dan 7 simpul daun. Berikut ini adalah hasil perhitungan mencari nilai *entropy* dan *information gain* dari semua atribut untuk menentukan simpul akar:

Tabel 7. Nilai *Information Gain* untuk Simpul Akar

No	Atribut	Gain (dalam bits)
1	Jenis Kelamin	0,0134
2	**Sistolik**	**0,07957**
3	Diastolik	0,00569
4	Riwayat Berobat	0,000671
5	Penyakit Lain	0,010468
6	Umur	0,046731

Berdasarkan Tabel 7, dapat diketahui bahwa dalam penelitian ini atribut tekanan darah sistolik terpilih sebagai pemilah pada simpul akar karena memiliki nilai *information gain* terbesar diantara atribut lainnya.



Gambar 1. Pohon Klasifikasi Algoritma C4.5

Berdasarkan pohon keputusan pada Gambar 1, banyaknya keseluruhan simpul yang terbentuk adalah 13 simpul, yang terdiri dari satu simpul akar, 5 simpul keputusan, dan 7 simpul daun. Sedangkan variabel yang berpengaruh adalah tekanan darah sistolik, riwayat berobat, penyakit lain, tekanan darah diastolik, dan jenis kelamin.

4.2.2 Ketepatan Klasifikasi

Berikut matriks konfusi pada perhitungan ketepatan klasifikasi algoritma C4.5 :

Tabel 8. Ketepatan Klasifikasi Algoritma C4.5

	Prediksi	
Observasi	Sekunder	Primer
Sekunder	3	37
Primer	14	88

Berdasarkan Tabel 8, dapat diketahui bahwa pada pengklasifikasian jenis hipertensi di UPT Puskesmas Ponjong I, Gunungkidul didapatkan nilai APER sebesar 35,9155% dengan ketepatan klasifikasi sebesar 64,0845%.

4.4 Perbandingan Ketepatan Klasifikasi

Ketepatan klasifikasi menggunakan regresi logistik biner dan algoritma C4.5 sebagai berikut:

Tabel 9. Perbandingan Ketepatan Klasifikasi

	Regresi Logistik Biner	Algoritma C4.5
APER	27,4648%	35,9155%
1-APER	72,5352%	64,0845%

Berdasarkan Tabel 9, diperoleh nilai ketepatan klasifikasi menggunakan metode regresi logistik biner sebesar 72,5352% dengan nilai laju error sebesar 27,4648%. Sedangkan nilai ketepatan klasifikasi menggunakan metode algoritma C4.5 adalah sebesar 64,0845% dengan nilai laju error 35,9155%.

Untuk mendapatkan metode terbaik, maka dilakukan evaluasi ketepatan klasifikasi dengan melakukan uji beda dua proporsi sebagai berikut :

Hipotesis:

$H_0 : PR_1 = PR_2$ (tidak ada perbedaan signifikan dari kedua metode)

$H_1 : PR_1 \neq PR_2$ (ada perbedaan signifikan dari kedua metode)

Taraf Signifikansi: $\alpha = 5\%$

Statitik Uji:

$$PR_{gab} = \frac{142x0{,}725352 + 142x0{,}640845}{142 + 142} = \frac{194}{284} = 0{,}683099$$

$$Z_{hitung} = \frac{PR_1 - PR_2}{\sqrt{\left(PR_{gab}(1 - PR_{gab})x\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right.}}$$

$$= \frac{0{,}725352 - 0{,}640845}{\sqrt{\left(0{,}683099(1 - 0{,}683099)x\left(\frac{1}{142} + \frac{1}{142}\right)\right.}}$$

$$= 27{,}71682$$

Kriteria Uji: H_0 ditolak jika $Z_{hitung} < -Z_{\alpha/2}$ atau $Z_{hitung} > Z_{\alpha/2}$

Keputusan: Karena $Z_{hitung} = 27{,}71682 > Z_{\alpha/2} = 1{,}960$ maka H_0 ditolak.

Kesimpulan:

Jadi pada taraf signifikansi 5% disimpulkan bahwa ada perbedaan signifikan dari kedua metode. Dengan kata lain, metode regresi logistik biner lebih baik dalam mengklasifikasikan jenis penyakit hipertensi di UPT Puskesmas Ponjong I, Gunungkidul dibandingkan dengan metode algoritma C4.5.

5. KESIMPULAN

Berdasarkan analisis dan pembahasan yang telah diuraikan di atas, maka dapat diambil kesimpulan sebagai berikut:

1. Pada regresi logistik biner, faktor-faktor yang mempengaruhi terjadinya jenis penyakit hipertensi adalah jenis kelamin, tekanan darah sistolik, serta penyakit lain.
2. Pohon keputusan yang terbentuk menggunakan algoritma C4.5 menghasilkan pohon sebanyak 13 simpul, yang terdiri dari sebuah simpul akar yaitu atribut tekanan darah sistolik, 5 simpul keputusan, serta 7 simpul daun. Faktor-faktor yang mempengaruhi terjadinya jenis penyakit hipertensi adalah tekanan darah sistolik, riwayat berobat, penyakit lain, tekanan darah diastolik, dan jenis kelamin, sedangkan faktor umur tidak berpengaruh.
3. Berdasarkan hasil analisis regresi logistik biner diperoleh nilai APER=27,4648% dan ketepatan klasifikasi sebesar 72,5352%, sedangkan pada algoritma C4.5 diperoleh nilai APER=35,9155% dengan ketepatan klasifikasi sebesar 64,0845%. Dari uji beda proporsi dihasilkan bahwa ada perbedaan signifikan antara kedua

metode, sehingga dapat dikatakan bahwa regresi logistik biner lebih baik dibandingkan algoritma C4.5 dalam mengklasifikasikan jenis penyakit hipertensi di UPT Puskesmas Ponjong I, Gunungkidul.

DAFTAR PUSTAKA

Agresti, A. 1990. *Categorical Data Analysis.*New York: John Wiley & Sons.

Hastie, T.,Tibshirani, R. And Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining. Inference and Prediction Second Edition*. New York: Springer Science Bussiness Media.

Hosmer, D. W. and Lemeshow S. 2000. *Applied Logistic Regression*. United States of American: Sons Inc.

Johnson, R. A. and Wichern. D. W., 2007. *Applied Multivariate Statistical Analysis.* Sixth Edition. New Jersey: Prentice Hall International, Inc.

Kementerian Kesehatan RI. 2013. *Pusat Data dan Informasi Kementerian Kesehatan RI Hipertensi*.(https://www.google.co.id/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiu8u7m5ezJAhUCnqYKHfROA8oQFggdMAA&url=http%3A%2F%2Fwww.depkes.go.id%2Fdownload.php%3Ffile%3Ddownload%2Fpusdatin%2Finfodatin%2Finfodatinhipertensi.pdf&usg=AFQjCNHWLiHieCeL1Ksg4Tr_yxZ10Ky7Cg diakses pada tanggal 21 Desember 2015)

Kusrini dan Luthfi. 2009. *Algoritma Data Mining*. Yogyakarta: Andi Offset.

Mansjoer, A. dkk. 2001. *Kapita Selekta Kedokteran Edisi Ketiga*. Jakarta: Media Aesculapius Fakultas Kedokteran Universitas Indonesia.

Quinlan, J. R., 1993. *C4.5: Programs For Machine Learning*. San Mateo: Morgan Kaufmann Publisher, Inc.

Riset Kesehatan Dasar. 2013. *Riset Kesehatan Dasar 2013*. (http://labmandat.litbang.depkes.go.id/images/download/laporan/RKD/2013/Laporan_riskesdas_2013_final.pdf diakses pada tanggal 21 Desember 2015)

Ruggieri, S., 2002. *Efficient C4*.5. (http://www.di.unipi.it/~ruggieri/ Papers/ec45.pdf, diakses pada tanggal 21 Desember 2015).

Rokach, L. and Maimon, O., 2008. *Data Mining With Decision Trees: Theory and Applications*. Singapura: World Scientific Publishing Co. Pte. Ltd.

Sugiarto, D.S. 2000. *Metode Statistika*. Jakarta: Gramedia Pustaka Utama.

UPT Puskesmas Ponjong I. 2015. *Profil Kesehatan UPT Puskesmas Ponjong I Dinas Kesehatan Kabupaten Gunungkidul Tahun 2015 (Data tahun 2014)*. Yogyakarta: UPT Puskesmas Ponjong I.

Witten, I. H., Frank, E., Hall, M. A., 2011. *DATA MINING Practical Machine Learning Tools and Teachniques.* Second Edition. California: Morgan Kaufman.

World Health Organization. 2013. *A Global Brief of Hypertension*. (http://www.who.int/cardiovasculardiseases/publications/global_brief_hypertension/en/ diakses pada tanggal 21 Desember 2015)