Indones. J. Math. Nat. Sci. 44(1) 2021
Indonesian Journal of Mathematics and Natural Sciences
http://journal.unnes.ac.id/nju/index.php/JM

Ketepatan Klasifikasi Metode Regresi Logistik dan Metode Chaid dengan Pembobotan Sampel

Puspa Juwita*, Sugiman, Putriaji Hendikawati

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia
Gedung D7 Lt.1, Kampus Sekaran Gunungpati, Semarang 50229
E-mail: puspajuwita.ej@gmail.com

Diterima 2 November 2020 Disetujui 10 Maret 2021 Dipublikasikan 30 April 2021

Abstrak

Tujuan penelitian ini adalah menentukan ketepatan metode regresi logistik dan CHAID dengan pembobotan sampel pada klasifikasi status angkatan kerja Kabupaten Temanggung 2015. Populasi dalam penelitian ini adalah angkatan kerja Kabupaten Temanggung 2015. Data dalam penelitian ini diperoleh dari Sakernas Kabupaten Temanggung 2015. Variabel dependen dalam penelitian ini adalah angkatan kerja, sedangkan variabel independennya adalah klasifikasi desa/kelurahan, hubungan dengan kepala rumah tangga, jenis kelamin, umur, status pernikahan, pendidikan, pelatihan kerja, dan pengalaman kerja. Dari analisis regresi logistik diperoleh persamaan, sedangkan anlalisi CHAID menghasilkan pohon klasifikasi. Persamaan dan pohon klasifikasi tersebut dapat digunakan untuk memprediksi variabel dependen. Kesalahan klasifikasi dihitung menggunakan APER (*Apparent Error Rate*), kemudian ketepatan klasifikasi dapat diperoleh dengan rumus 1 – APER. Ketepatan regresi logistik dan CHAID dengan pembobotan sampel secara berturut-turut adalah 96,4% dan 96,6%. Hal ini menunjukkan ketepatan metode CHAID pada klasifikasi status angkatan kerja Kabupaten Temanggung 2015 lebih tinggi dibandingkan regresi logistik.

Kata kunci: regresi logistik, CHAID, pembobotan sampel

Abstract

The purpose of this study is to determine the accuracy of logistic regression and CHAID with sample weighting on Temanggung regency labor status classification in 2015. The population of this study is labor of Temanggung Regency in 2015. The data of this study is obtained from Sakernas of Temanggung Regency in 2015. The dependent variable of this study is labor status, whereas the independent variables of this study are domicile region, relation with family head, gender, age, marriage status, education level, job training, and job experience. Logistic regression analysis results a mathematic equation, and CHAID method result a classification tree. Those result can predict the dependent variable. Classification error is calculated using APER (Apparent Error Rate), then the accuracy can be calculated by 1- APER. Accuracy of logistic regression and CHAID with sample weighting respectively are 96,4% and 96,6%. This show that accuracy of CHAID is greater than logistic regression.

Keywords: logistic regression, CHAID, sample weighting

How to cite:
Juwita P., Sugiman, & Hendikawati P. (2021). Ketepatan Klasifikasi Metode Regresi Logistik dan Metode Chaid dengan Pembobotan Sampel. *Indones. J. Math. Nat. Sci.,* 44(1), 22-33

PENDAHULUAN

Metode klasifikasi banyak digunakan dalam berbagai bidang, seperti pendidikan, pemerintahan, kesehatan, teknologi, maupun sosial. Klasifikasi didefinisikan sebagai pekerjaan mengelompokkan suatu objek ke dalam kategori tertentu. Klasifikasi dapat dilakukan pada data kategorik maupun bukan. Jika data bukan kategorik maka harus diubah dalam bentuk kategorik terlebih dahulu.

Regresi logistik merupakan pendekatan pemodelan matematika yang dapat digunakan untuk mendeskripsikan hubungan beberapa variabel independen dengan variabel dependen dikotomi. Model regresi logistik dibuat untuk mendeskripsikan peluang variabel dependen antara 0 dan 1 (Kleinbaum & Klein, 2010). Berdasarkan penelitian Rahman & Zain (2014) dan Imaslihkah *et al*. (2013), regresi logistik mempunyai ketepatan klasifikasi yang akurat. Menurut Antipov & Pokryshevskaya (2009), regresi logistik sangat menarik karena beberapa hal, yaitu (1) secara konsep sederhana, (2) mudah diinterpretasikan, dan (3) terbukti dapat menyediakan hasil yang akurat dan baik.

Pohon keputusan (*decision tree*) digunakan secara luas sebagai alat bantu prediksi. *Decision tree* mampu mendeteksi dan menghitung efek nonlinear pada variabel dependen dan interaksi diantara variabel independen. CHAID (*Chi-Square Automatic Interaction Detection*) merupakan salah satu pohon keputusan (*decision tree*). Seperti namanya, CHAID menggunakan kriteria uji chi-square untuk membentuk diagram pohon. Pada setiap cabangnya, CHAID melakukan tahap penggabungan (*merging)* dan tahap pemisahan (*splitting)* (Ritschard, 2010). Menurut penelitian Milana (2012) dan Rahayu *et al*. (2015), metode CHAID akurat untuk klasifikasi.

Dalam suatu penelitian, sampel yang representatif terhadap populasi sangat penting. Sampel yang representatif akan meningkatkan keakuratan hasil penelitian. Diharapkan dengan sampel yang terbatas populasi dapat terwakili, sehingga penelitian akan efektif dan efisien. Oleh karena itu diperlukan pembobotan sampel, yaitu pemberian bobot pada data sehingga satu sampel dapat mewakili lebih dari satu data dalam populasi.

Indonesia merupakan negara berkembang yang telah mengalami kemajuan pesat di bidang ekonomi dan sosial. Semakin banyak penduduk Indonesia yang menikmati standar hidup yang lebih tinggi. Indonesia juga mempunyai potensi pertumbuhan yang kuat, yaitu populasi yang masih muda. Saat ini Indonesia mempunyai tantangan dalam bidang perekonomian, yaitu diversifikasi ekonomi dengan memperkuat kualitas sumber daya manusia sehingga memungkinkan sektor-sektor ekonomi yang padat keterampilan dan padat tenaga kerja untuk terus berkembang (Survei Ekonomi OECD : Indonesia 2016, 2016).

Menurut Sumitro Djojohadikusumo (1994) dalam BPS (2015), masalah pengangguran menjadi tantangan pokok dalam pembangunan ekonomi negara-negara berkembang. berhasil tidaknya suatu usaha untuk menanggulangi hal ini akan mempengaruhi kestabilan sosial politik dalam kehidupan masyarakat dan kontinuitas dalam pembangunan ekonomi jangka panjang. pada tahun 2015 kabupaten temanggung mempunyai tingkat partisipasi angkatan kerja (TPAK) paling tinggi di jawa tengah, yaitu 75,47%. tujuan penelitian ini adalah untuk menentukan ketepatan metode regresi logistik dan metode chaid dengan pembobotan sampel untuk klasifikasi status angkatan kerja Kabupaten Temanggung 2015.

METODE

Sumber Data

Data dalam penelitian ini diperoleh dari Sakernas Kabupaten Temanggung tahun 2015.

Variabel Penelitian

Variabel dalam penelitian ini terdiri dari variabel dependen dan variabel independen. Karena merupakan data kategori, maka diperlukan variabel *dummy* di mana kategori pada setiap variabel diberi kode. Adapun variabel-variabel dalam penelitian ini adalah sebagai berikut.

Variabel Dependen

1. Status Angkatan Kerja (Y)
   a. Pengangguran (0)
   b. Bekerja (1)

Variabel Independen

1. Klasifikasi Desa/Kelurahan (X1)
   a. Pedesaan (0)
   b. Perkotaan (1)
2. Hubungan dengan Kepala Rumah Tangga (X2)
   a. Bukan Kepala Rumah Tangga (0)
   b. Kepala Rumah Tangga (1)
3. Jenis Kelamin (X3)
   a. Perempuan (0)
   b. Laki-laki (1)

4. Umur (X4)
   a. 15 – 24 tahun (0)
   b. 25 – 54 tahun (1)
   c. ≥ 55 tahun (2)
5. Status Pernikahan (X5)
   a. Tidak menikah (0)
   b. Menikah (1)
6. Pendidikan (X6)
   a. ≤ SD sederajat (0)
   b. SLTP sederajat (1)
   c. SLTA sederajat (2)
   d. DI – DIII (3)
   e. ≥ S1 (4)
7. Pelatihan Kerja (X7)
   a. Tidak (0)
   b. Ya (1)
8. Pengalaman Kerja (X8)
   a. Tidak (0)
   b. Ya (1)

Analisis Data

Analisis data menggunakan metode regresi logistik dan CHAID dengan pembobotan sampel.

Pembobotan Sampel

Pembobotan sampel dilakukan dengan cara memberikan bobot pada setiap sampel sehingga satu sampel dapat mewakili lebih dari satu data populasi.

Regresi Logistik

Estimasi Parameter

Untuk membuat model regresi logistik dibutuhkan estimasi parameter. Metode yang digunakan adalah *maximum likelihood*. Prinsip *maximum likelihood* menyatakan bahwa digunakan nilai estimasi yang memaksimalkan fungsi likelihood. Untuk menemukannya, fungsi likelihood diturunkan terhadap koefisien dan disamadenganken dengan dengan 0. Penyelesaian persamaan likelihood membutuhkan bantuan *software* (Hosmer & Lemeshow, 2000).

Uji Signifikansi Serentak

Setelah diperoleh persamaan regresi logistik, maka selanjutnya menguji signifikansi variabel di dalam model. Membandingakn nilai observasi dan nilai prediksi variabel dependen yang diperoleh dari model regresi logistik dengan variabel independen dan tanpa variabel independen. Untuk membandingkannya digunakan uji rasio likelihood (Hosmer & Lemeshow, 2000).

Uji signifikansi serentak digunakan untuk mengetahui apakah variabel independen signifikan secara bersama-sama terhadap variabel dependen.

H_0 : $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$

H_1 : paling sedikit ada satu $\boldsymbol{\beta_j \neq 0}$ dengan $\boldsymbol{j = 1, 2, \ldots, p}$

Tolak H_0 jika $\boldsymbol{G > \chi^2_{(p,\alpha)}}$ atau *p-value* < taraf signifikansi.

Uji Signifikansi Parsial

Uji signifikansi parsial digunakan untuk mengetahui apakah variabel independen secara individu/parsial signifikan terhadap variabel dependen. Untuk menguji signifikansi parsial regresi logistik digunakan uji Wald. Menurut Hosmer & Lemeshow (2000), uji Wald diperoleh dengan membandingkan estimasi maximum likelihood terhadap estimasi standar error.

H_0 : $\boldsymbol{\beta_j = 0}$ dengan $\boldsymbol{j = 1, 2, \ldots, p}$

H_1 : $\boldsymbol{\beta_j \neq 0}$ dengan $\boldsymbol{j = 1, 2, \ldots, p}$

Tolak H_0 jika $\boldsymbol{W_j > \chi^2_{(\alpha,1)}}$ atau *p-value* < taraf signifikansi.

Uji Kesesuaian Model

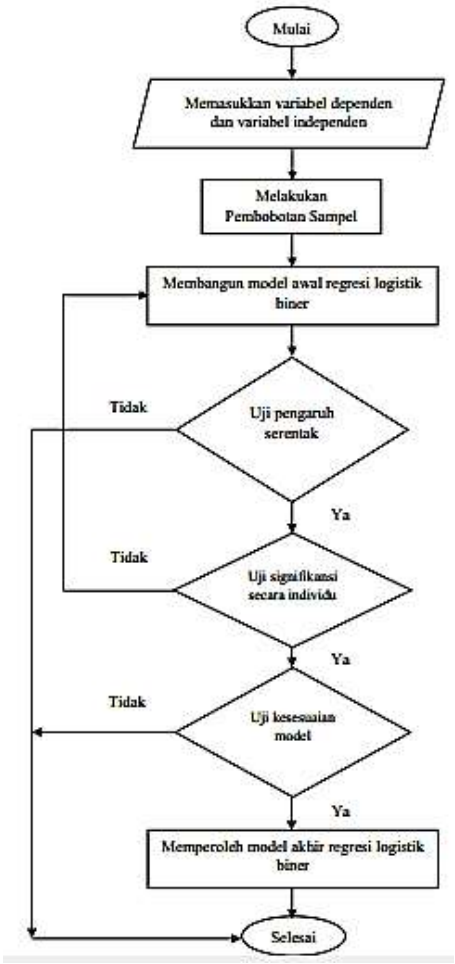
Uji kesesuaian model digunakan untuk menguji apakah model sudah sesuai, yaitu tidak ada perbedaaan signifikan antara hasil observasi dengan hasil prediksi model. Statistik uji yang digunakan adalah uji Hosmer Lemeshow. Menurut Hosmer & Lemeshow (2000), uji kesesuaian model mengetahui keefektifan model dalam menjelaskan variabel dependen.

H_0 : tidak terdapat perbedaan antara hasil observasi dengan hasil prediksi

H_1 : terdapat perbedaan antara hasil observasi dengan hasil prediksi

Tolak $\boldsymbol{H_0}$ jika $\hat{\boldsymbol{c}} > \boldsymbol{\chi^2_{(\alpha,g-2)}}$ atau $\boldsymbol{H_0}$ ditolak jika p-value < $\boldsymbol{\alpha}$.

Diagram alir metode regresi logistik dengan pembobotan sampel dapat dilihat pada Gambar 1.



Gambar 1. Diagram Alir Regresi Logistik dengan Pembobotan Sampel

Chaid

Tahap Penggabungan (Merging)

Pada tahap penggabungan dilakukan uji signifikansi pasangan kategori pada setiap variabel independen terhadap variabel dependen. Kriteria pengujian menggunakan uji chi-square. Pasangan kategori variabel independen yang signifikan terhadap variabel dependen tidak digabung, sedangkan yang tidak signifikan akan digabung menjadi sebuah kategori gabungan.

H_0 : tidak terdapat hubungan antara variabel pertama dan variabel kedua

H_1 : terdapat hubungan antara variabel pertama dan variabel kedua

Tolah H_0 jika $\boldsymbol{\chi^2_{hitung} > \chi^2_{\alpha,(r-1)(c-1)}}$.

Tahap Pemisahan (Splitting)

Pada pada tahap pemisahan, dipilih variabel independen terbaik untuk memisahkan simpul/cabang pohon CHAID. Kriteria pemilihan variabel pemisah terbaik adalah uji chi-square.

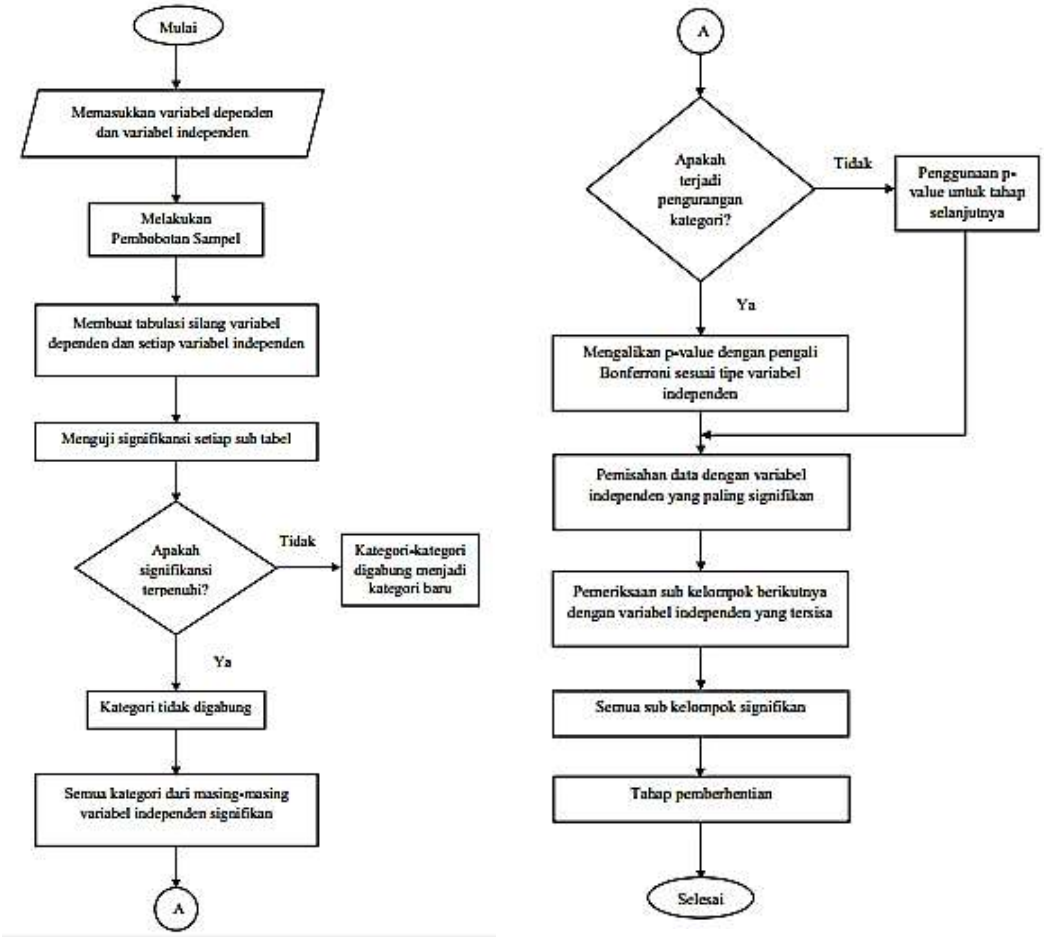
Tahap Penghentian (Stopping)

Pada setiap simpul pohon, dilakukan tahap penggabungan dan tahap pemisahan. Tahap penggabungan dan tahap pemisahan pada diagram pohon terus dilakukan sampai memenuhi kriteria tahap penghentian (*stopping*). Jika pembentukan diagram pohon memenuhi satu atau lebih kriteria berikut ini maka dilakukan tahap penghentian.

a. Jika sebuah simpul menjadi murni, yaitu semua kasus dalam simpul tersebut mempunyai nilai variabel dependen yang sama

b. Jika semua kasus dalam sebuah simpul mempunyai nilai yang sama untuk setiap variabel independen
c. Jika kedalaman pohon mencapai batas kedalaman pohon maksimal yang ditentukan
d. Jika ukuran simpul kurang dari ukuran simpul minimal yang ditentukan
e. Jika pemisahan sebuah simpul menghasilkan child node (simpul anak) yang ukurannya kurang dari ukuran simpul anak minimal, simpul anak yang mempunyai terlalu sedikit kasus (< ukuran simpul anak minimal) akan digabungkan dengan simpul anak yang paling mirip yang diukur menggunakan p-value terbesar. Akan tetapi, jika ukuran child node yang dihasilkan adalah 1, pertumbuhan diagram pohon akan dihentikan.

Diagram alir metode CHAID dengan pembobotan sampel dapat dilihat pada gambar 2.



Gambar 2. Diagram Alir Metode CHAID dengan Pembobotan Sampel

HASIL DAN PEMBAHASAN

Regresi Logistik

Analisis regresi logistik dilakukan sampai semua variabel independen dalam persamaan signifikan (secara serentak dan secara parsial) terhadap variabel dependen. Jika masih terdapat variabel independen yang tidak signifikan, maka analisis regresi akan terus dilakukan (hanya dengan variabel independen yang signifikan).

Model Pertama

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

dengan

$$g(x) = 36{,}067 - 0{,}071x_1(1) - 0{,}235x_2(1) + 0{,}300x_3(1) - 18{,}118x_4(1) - 15{,}662x_4(2) - 2{,}124x_5(1) + 5{,}554x_6(1) + 4{,}687x_6(2) + 3{,}474x_6(3) + 1{,}450x_6(4) - 17{,}611x_7(1) - 1{,}475x_8(1)$$

Uji Signifikansi Serentak Model Pertama

Hasil uji signifikansi serentak model pertama dapat dilihat pada Tabel 1. Berdasarkan Tabel 1, nilai uji rasio likelihood adalah 29.030,031 dengan nilai signifikansi 0,000. Keputusan pengujian adalah menolak H_0. Dapat ditarik kesimpulan bahwa paling sedikit terdapat satu variabel independen yang signifikan terhadap variabel status angkatan kerja.

Tabel 1. Hasil Uji Signifikansi Serentak Model Pertama (Omnibus Tests of Model Coefficients)

		Chi-square	df	Sig.
Step 1	Step	29030,031	12	,000
	Block	29030,031	12	,000
	Model	29030,031	12	,000

Uji Signifikansi Parsial Model Pertama

Output hasil uji signifikansi parsial model pertama dengan menggunakan software komputer dapat dilihat pada Gambar 3. Berdasarkan Gambar 3, dapat dilihat nilai uji Wald dan nilai signifikansinya untuk setiap variabel independen. Keputusan pengujian adalah menolak H_0 untuk variabel hubungan dengan kepala rumah tangga, jenis kelamin, status pernikahan, pendidikan, dan pengalaman kerja. Dapat ditarik kesimpulan bahwa hanya variabel-variabel tersebut yang signifikan terhadap variabel status angkatan kerja. Oleh karena itu, perlu dilakukan analisis regresi logistik kedua.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1(1)	-,071	,051	1,901	1	,168	,932
	X2(1)	-,235	,106	4,946	1	,026	,791
	X3(1)	,300	,039	57,997	1	,000	1,350
	X4			1591,704	2	,000	
	X4(1)	-18,118	193,923	,009	1	,926	,000
	X4(2)	-15,662	193,923	,007	1	,936	,000
	X5(1)	-2,124	,065	1080,442	1	,000	,120
	X6			4539,341	4	,000	
	X6(1)	5,554	,107	2684,545	1	,000	258,240
	X6(2)	4,687	,076	3805,342	1	,000	108,540
	X6(3)	3,474	,064	2922,871	1	,000	32,269
	X6(4)	1,450	,105	191,105	1	,000	4,263
	X7(1)	-17,611	507,062	,001	1	,972	,000
	X8(1)	-1,475	,040	1385,837	1	,000	,229
	Constant	36,067	542,880	,004	1	,947	4,611E15

a. Variable(s) entered on step 1: X1, X2, X3, X4, X5, X6, X7, X8.

Gambar 3. Output Hasil Uji Signifikansi Parsial Model Pertama

Model Kedua

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

dengan

$$g(x) = 3{,}870 - 1{,}095x_2(1) + 0{,}485x_3(1) - 3{,}602x_5(1) + 5{,}464x_6(1) + 3{,}640x_6(2) + 2{,}780x_6(3) + 0{,}844x_6(4) - 1{,}569x_8(1)$$

Uji Signifikansi Serentak Model Kedua

Hasil uji signifikasni serentak model kedua dapat dilihat pada Tabel 2.

Tabel 2. Hasil Uji Signifikasni Serentak Model Kedua (Omnibus Tests of Model Coefficients)

		Chi-square	df	Sig.
Step 1	Step	25952,666	8	,000
	Block	25952,666	8	,000
	Model	25952,666	8	,000

Berdasarkan Tabel 2, nilai uji rasio likelihood adalah 25.952,666 dengan nilai signifikansi 0,000. Keputusan pengujian adalah menolak H_0. Dapat ditarik kesimpulan bahwa paling sedikit terdapat satu variabel independen yang signifikan terhadap variabel status angkatan kerja.

Uji Signifikansi Parsial Model Kedua

Output hasil uji signifikansi parsial model kedua dengan menggunakan software komputer dapat dilihat pada Gambar 4.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X2(1)	-1,095	,091	143,218	1	,000	,335
	X3(1)	,485	,036	180,279	1	,000	1,624
	X5(1)	-3,602	,055	4214,760	1	,000	,027
	X6			5395,084	4	,000	
	X6(1)	5,464	,095	3305,874	1	,000	235,985
	X6(2)	3,640	,058	3884,754	1	,000	38,084
	X6(3)	2,780	,054	2686,529	1	,000	16,120
	X6(4)	,844	,087	94,672	1	,000	2,326
	X8(1)	-1,569	,037	1770,419	1	,000	,208
	Constant	3,870	,092	1760,464	1	,000	47,918

a. Variable(s) entered on step 1: X2, X3, X5, X6, X8.

Gambar 4 Output Hasil Uji Signifikansi Parsial Model Kedua

Berdasarkan Gambar 4, dapat dilihat nilai uji Wald dan nilai signifikansinya untuk setiap variabel independen. Keputusan pengujian adalah menolak H_0 untuk variabel hubungan dengan kepala rumah tangga, jenis kelamin, status pernikahan, pendidikan, dan pengalaman kerja. Dapat ditarik kesimpulan bahwa variabel-variabel tersebut signifikan terhadap variabel status angkatan kerja.

Uji Kesesuaian Model

Dari analisis regresi logistik kedua, semua variabel independen dalam model telah signifikan terhadap variabel status angkatan kerja. Untuk menguji kesesuaian model, dilakukan uji Hosmer Lemeshow terhadap model regresi logistik kedua. Berdasarkan penelitian Graubard *et al.*, uji Hosmer Lemeshow tidak dapat dimodifikasi untuk kasus pembobotan sampel. Karena pada penelitian ini digunakan pembobotan sampel maka uji Hosmer Lemeshow tidak perlu dilakukan.

***Odd Ratio* (OR)**

Persamaan regresi logistik dapat diinterpretasikan dengan menggunakan nilai *odd ratio*. *Odd ratio* (OR) merupakan perbandingan antara peluang sukses (y=1) dengan peluang gagal (y=0). Pada penelitian ini, *odd ratio* merupakan perbandingan antara peluang bekerja dengan peluang pengangguran. Berikut ini merupakan interpretasi persamaan regresi logistik kedua.

1. Nilai *odd ratio* variabel hubungan dengan kepala rumah tangga adalah 0,335. Seorang angkatan kerja bukan kepala rumah tangga cenderung untuk memiliki status bekerja sebanyak 0,335 kali lipat dibandingkan kepala rumah tangga. Karena koefisien negatif maka hubungan variabel status angkatan kerja dan hubungan dengan kepala rumah tangga berbanding terbalik.
2. Nilai odd ratio variabel pendidikan kategori ≤ SD sederajat adalah 235,985. Seorang angkatan kerja yang mempunyai pendidikan ≤ SD sederajat cenderung untuk memiliki status bekerja sebanyak 235,985 kali lipat dibandingkan seorang yang mempunyai pendidikan > SD sederajat. Karena koefisien positif maka hubungan variabel status angkatan kerja dan pendidikan berbanding lurus.
3. Nilai odd ratio variabel pendidikan kategori SLTP sederajat adalah 38,084. Seorang angkatan kerja yang mempunyai pendidikan SLTP sederajat cenderung untuk memiliki status bekerja sebanyak 38,084 kali lipat dibandingkan seorang yang mempunyai pendidikan > SLTP sederajat. Karena koefisien positif maka hubungan variabel status angkatan kerja dan pendidikan berbanding lurus.
4. Dst.

Ketepatan Klasifikasi Regresi Logistik

Ketepatan regresi logistik dapat diukur menggunakan perhitungan 1 – APER (*Apparent Error Rate*).

1. Bekerja

   Pada contoh perhitungan prediksi variabel status angkatan kerja sebagai bekerja, digunakan sampel ke-1. Untuk mengetahui nilai prediksi variabel dependen sampel ke-1, nilai variabel independen disubstitusikan ke dalam persamaan regresi logistik. Untuk mensubstitusikan ke dalam persamaan regresi logistik, perlu memperhatikan Gambar 5.

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
X6	Tidak Tamat SD atau SD sederajat	217	1,000	,000	,000	,000
	SLTP sederajat	74	,000	1,000	,000	,000
	SLTA sederajat	78	,000	,000	1,000	,000
	DI - DIII	7	,000	,000	,000	1,000
	S1 ke atas	24	,000	,000	,000	,000
X8	Tidak	133	1,000			
	Ya	267	,000			
X3	Perempuan	174	1,000			
	Laki-laki	226	,000			
X5	Tidak Menikah	93	1,000			
	Menikah	307	,000			
X2	Bukan Kepala Rumah Tangga	233	1,000			
	Kepala Rumah Tangga	167	,000			

Gambar 5. Output Pengolahan Dengan Software SPSS Kode Variabel Dummy

$$g(x) = 3{,}870 - 1{,}095(1) + 0{,}485(0) - 3{,}602(1) + 5{,}464(1) + 3{,}640(0) + 2{,}780(0) + 0{,}844(0) - 1{,}569(0) = 4{,}637$$

sehingga

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{4{,}637}}{1 + e^{4{,}637}} = 0{,}99040$$

Karena $\pi(x) > 0{,}5$, maka sampel pertama diprediksi masuk kelas 1 (bekerja).

2. Pengangguran

   Pada contoh perhitungan prediksi variabel status angkatan kerja sebagai pengangguran, digunakan sampel ke-44. Perhitungan dilakukan dengan cara yang sama seperti pada kasus bekerja.

$$g(x) = 3{,}870 - 1{,}095(1) + 0{,}485(1) - 3{,}602(1) + 5{,}464(0) + 3{,}640(0) + 2{,}780(0) + 0{,}844(0) - 1{,}569(1) = -1{,}911$$

sehingga

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{e^{-1{,}911}}{1 + e^{-1{,}911}} = 0{,}12887$$

Karena $\pi(x) < 0{,}5$, maka sampel pertama diprediksi masuk kelas 0 (pengangguran).

Ketepatan klasifikasi dapat dihitung menggunakan rumus 1 – APER. Adapun perhitungannya adalah sebagai berikut.

$$APER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{4.314 + 605}{2.145 + 4.314 + 605 + 130.786} = 0{,}036$$

$$1 - APER = 1 - 0{,}036 = 0{,}964$$

Ketepatan metode regresi logistik dengan pembobotan sampel dalam klasifikasi status angkatan kerja Kabupaten Temanggung tahun 2015 adalah 96,4%.

CHAID

Tahap Penggabungan (Merging)

Tahap penggabungan untuk variabel umur akan dijelaskan pada bagian ini. Sebelum melakukan tahap penggabungan, terlebih dahulu ditentukan apakah variabel umur merupakan

variabel nominal atau ordinal. Jika variabel umur merupakan variabel nominal, maka pasangan kategori yang dapat dibentuk adalah dua kategori manapun yang dapat dibentuk. Jika variabel umur merupakan variabel ordinal, maka pasangan kategori yang dapat dibentuk adalah dua kategori yang berurutan. Pada penelitian ini, variabel umur merupakan variabel ordinal sehingga hanya kategori yang berurutan yang dapat dibentuk.

Tabel 3. Tabel Silang Status Angkatan Kerja dan Umur

Umur	Status Angkatan Kerja		Total
	Pengangguran	Bekerja	
15-24 tahun	5.209	15.597	20.806
25-54 tahun	1.250	84.425	85.675
≥ 55 tahun	0	31.369	31.369
Total	6.459	131.391	137.850

Dari Tabel 3, dapat diketahui statistik umur terhadap status angkatan kerja. Pasangan kategori yang dapat dibentuk adalah (15-24 tahun dan 25-54 tahun) dan (25-54 tahun dan ≥ 55 tahun). Langkah selanjutnya adalah membuat sub tabel dari tabel 1, yaitu tabel silang dari setiap pasang kategori variabel umur terhadap variabel status angkatan kerja. Sub tabel yang pertama dapat dilihat pada Tabel 4.

Tabel 4. Sub Tabel Silang Status Angkatan Kerja dan Umur

Umur	Status Angkatan Kerja		Total
	Pengangguran	Bekerja	
15-24 tahun	5.209	15.597	20.806
25-54 tahun	1.250	84.425	85.675
Total	6.459	100.022	106.481

Untuk melakukan penggabungan kategori, digunakan statistik uji chi-square.

H_0 : tidak terdapat hubungan antara variabel umur dan variabel status angkatan kerja

H_1 : terdapat hubungan antara variabel umur dan variabel status angkatan kerja

Tolak H_0 jika $\chi^2_{hitung} > \chi^2_{\alpha,(r-1)(c-1)}$

Taraf signifikansi : 5%

Perhitungan :

$$E_{11} = \frac{n_{1.} \cdot n_{.1}}{n} = \frac{20.806 \cdot 6.459}{106.481} = 1.262{,}065$$

$$E_{12} = \frac{n_{1.} \cdot n_{.2}}{n} = \frac{20.806 \cdot 100.022}{106.481} = 19.543{,}935$$

$$E_{21} = \frac{n_{2.} \cdot n_{.1}}{n} = \frac{85.675 \cdot 6.459}{106.481} = 5.196{,}935$$

$$E_{22} = \frac{n_{2.} \cdot n_{.2}}{n} = \frac{85.675 \cdot 100.022}{106.481} = 80.478{,}065$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

$$= \frac{(n_{11} - E_{11})^2}{E_{11}} + \frac{(n_{12} - E_{12})^2}{E_{12}} + \frac{(n_{21} - E_{21})^2}{E_{21}} + \frac{(n_{22} - E_{22})^2}{E_{22}}$$

$$= \frac{(5.209 - 1.262{,}065)^2}{1.262{,}065} + \frac{(15.597 - 19.543{,}935)^2}{19.543{,}935} + \frac{(1.250 - 5.196{,}935)^2}{5.196{,}935} + \frac{(84.425 - 80.478{,}065)^2}{80.478{,}065}$$

$$= \frac{15.578.295{,}894}{1.262{,}065} + \frac{15.578.295{,}894}{E_{ij}} + \frac{15.578.295{,}894}{5.196{,}935} + \frac{5.196{,}935}{80.478{,}065}$$

$$= 12.343{,}497 + 797{,}091 + 2.997{,}593 + 193{,}572 = 16.331{,}753$$

Keputusan :
Nilai $\chi^2_{0,05;(2-1)(2-1)}$ adalah 3,841. Keputusan pengujian adalah menolak H_0, artinya terdapat hubungan antara variabel umur dan variabel status angatan kerja. Sehingga pasangan ketegori pada Tabel 4 tidak digabung.

Tabel 5 merupakan sub tabel kedua dari tabel silang variabel status angkatan kerja dan variabel umur.

Tabel 5. Sub Tabel Silang Status Angkatan Kerja dan Umur

Umur	Status angkatan kerja		Total
	Pengangguran	Bekerja	
25-54 tahun	1.250	84.425	85.675
≥ 55 tahun	0	31.369	31.369
Total	1.250	115.794	117.044

Dengan cara yang sama, dilakukan uji chi-square pada Tabel 5. Diperoleh nilai statistik uji chi-square untuk Tabel 5 adalah 462,615.

Keputusan :
Nilai $\chi^2_{0,05;(2-1)(2-1)}$ adalah 3,841. Keputusan pengujian adalah menolak H0, artinya terdapat hubungan antara variabel umur dan variabel status angkatan kerja. Sehingga pasangan ketegori pada Tabel 5 tidak digabung.

Tahap Pemisahan (Splitting)

Pada tahap pemisahan, simpul akan dipisah menggunakan sebuah variabel independen. Simpul ini akan dibagi menjadi beberapa bagian berdasarkan kategori variabel independen pemisah simpul. Pada bagian ini akan dijelaskan tahap pemisahan dari root node. Untuk mengetahui variabel independen mana yang akan digunakan untuk memisahkan simpul root node, digunakan statistik uji chi-square. Variabel independen yang digunakan untuk memisahkan simpul adalah yang mempunyai nilai statistik uji chi-square terbesar dan signifikan terhadap variabel dependen.

Sebelum melakukan uji chi-square, terlebih dahulu dibentuk tabel silang setiap variabel independen terhadap variabel status angkatan kerja. Setelah itu, dilakukan uji chi-square pada setiap tabel silang. Adapun langkah uji chi-square sama seperti sebelumnya. Hasil uji chi-square dari setiap variabel independen terhadap variabel status angkatan kerja dapat dilihat pada Tabel 6. Berdasarkan Tabel 6, nilai statistik uji chi-square terbesar terdapat pada variabel umur dan nilai sig. variabel umur kurang dari taraf signifikansi (signifikan terhadap variabel status angkatan kerja). Sehingga variabel umur digunakan sebagai variabel pemisah pada root node.

Tabel 6. Hasil Uji Chi-Square Setiap Variabel Independen terhadap Variabel Status Angkatan Kerja

Variabel Independen	Nilai Statistik Uji Chi-Square	Sig.
Klasifikasi Desa/Kelurahan	1.992,557	0,000
Hubungan dengan Kepala Rumah Tangga	3.599,135	0,000
Jenis Kelamin	7,795	0,005
Umur	22.833,188	0,000
Status Pernikahan	15.009,156	0,000
Pendidikan	13.730,950	0,000
Pelatihan Kerja	263,629	0,000
Pengalaman Kerja	4.197,282	0,000

Sekarang telah terbentuk diagram pohon dengan tiga cabang, yaitu 15-24 tahun, 25-54 tahun, dan ≥ 55 tahun. Langkah penggabungan dan pemisahan dilakukan lagi pada setiap cabang yang terbentuk berdasarkan variabel umur sampai kriteria penghentian terpenuhi.

Tahap Penghentian (Stopping)

Pada tahap penghentian, pembentukan diagram pohon dihentikan karena memenuhi satu atau lebih kriteria penghentian.

Ketepatan Metode CHAID

Pohon CHAID yang telah terbentuk berisi variabel-variabel independen, yaitu umur, pendidikan, pengalaman kerja, status pernikahan, dan klasifikasi desa/kelurahan. Pelabelan kelas dilakukan pada setiap simpul akhir (*terminal node*) untuk mengetahui termasuk dalam kategori variabel dependen yang mana suatu simpul akhir. Pelabelan kelas dilakukan berdasarkan persentase terbesar kategori variabel dependen dalam simpul akhir. Misalnya dalam suatu simpul akhir kategori A sebesar 30% dan kategori B sebesar 70%, maka simpul akhir ini masuk ke dalam kategori B.

Dari analisis data yang telah dilakukan dihasilkan pohon CHAID yang mempunyai 14 simpul akhir.

1. Simpul 4. Seorang angkatan kerja dikategorikan bekerja jika berumur 15-24 tahun dan pendidikan ≤ SD sederajat.
2. Simpul 12. Seorang angkatan kerja dikategorikan bekerja jika berumur 15-24 tahun, pendidikan SLTP sederajat, dan mempunyai pengalaman kerja.
3. Simpul 13. Seorang angkatan kerja dikategorikan bekerja jika berumur 15-24 tahun, pendidikan SLTP sederajat, dan tidak mempunyai pengalaman kerja.
4. Simpul 14. Seorang angkatan kerja dikategorikan bekerja jika berumur 15-24 tahun, pendidikan SLTA sederajat, dan mempunyai pengalaman kerja.
5. Simpul 15. Seorang angkatan kerja dikategorikan bekerja jika berumur 15-24 tahun, pendidikan SLTA sederajat, dan tidak mempunyai pengalaman kerja.
6. Simpul 7. Seorang angkatan kerja dikategorikan pengangguran jika berumur 15-24 tahun dan pendidikan DI-DIII, ≥ S1.
7. Dst

Setelah dilakukan pelabelan kelas pada setiap simpul akhir, langkah selanjutnya adalah mengklasifikasikan setiap sampel apakah termasuk dalam kategori bekerja atau pengangguran. Berikut ini merupakan contoh prediksi variabel status angkatan kerja sebagai bekerja atau pengangguran.

1. Bekerja

   Pada contoh prediksi status angkatan kerja sebagai bekerja digunakan sampel ke-2. Untuk mengkategorikan sampel, dilihat variabel independen yang masuk dalam diagram pohon CHAID. Variabel independen yang masuk dalam diagram pohon CHAID adalah umur, pendidikan, pengalaman kerja, status pernikahan, dan klasifikasi desa/kelurahan. Pada sampel ke-2, nilai variabel untuk variabel independen tersebut berturut-turut adalah 15-24 tahun, SLTP sederajat, tidak mempunyai pengalaman kerja, tidak menikah, dan bertempat tinggal di pedesaan. Maka sampel ke-2 termasuk dalam simpul akhir 13. Sampel ke-2 dikategorikan sebagai bekerja.
2. Pengangguran

Pada contoh prediksi status angkatan kerja sebagai pengangguran digunakan sampel ke-393. Untuk mengkategorikan sampel, dilihat variabel independen yang masuk dalam diagram pohon CHAID. Variabel independen yang masuk dalam diagram pohon CHAID adalah umur, pendidikan, pengalaman kerja, status pernikahan, dan klasifikasi desa/kelurahan. Pada sampel ke-393, nilai variabel untuk variabel independen tersebut berturut-turut adalah 15-24 tahun, ≥ S1, mempunyai pengalaman kerja, tidak menikah, dan bertempat tinggal di perkotaan. Maka sampel ke-2 termasuk dalam simpul akhir 7. Sampel ke-393 dikategorikan sebagai pengangguran.

Prediksi variabel status angkatan kerja dilakukan pada 400 sampel. Setelah dilakukan prediksi nilai variabel status angkatan kerja semua sampel, diperoleh hasil klasifikasi seperti pada Tabel 7.

Tabel 7. Tabel Klasifikasi Metode CHAID

Observasu	Prediksi		Total
	Pengangguran	Bekerja	
Pengangguran	1.745	4.714	6.459
Bekerja	0	131.391	131.391
Total	1.745	136.105	137.850

Ketepatan klasifikasi dapat dihitung menggunakan rumus 1 – APER. Adapun perhitungannya adalah sebagai berikut.

$$APER = \frac{4.714 + 0}{1.745 + 4.714 + 0 + 131.391} = 0{,}034$$

$$1 - APER = 1 - 0{,}034 = 0{,}966$$

Berdasarkan perhitungan di atas, diperoleh ketepatan klasifikasi metode metode CHAID adalah sebesar 0,966 atau 96,6%.

Ketepatan klasifikasi metode regresi logistik dan metode CHAID dalam mengklasifikasikan status angkatan kerja Kabupaten Temanggung 2015 dapat dilihat pada Tabel 8.

Tabel 8. Tabel Ketepatan Metode Regresi Logistik dan Metode CHAID

	APER	Ketepatan Klasifikasi
Metode Regresi Logistik	3,6%	96,4%
Metode CHAID	3,4%	96,6%

Dari hasil dan pembahasan penelitian, dapat ditarik kesimpulan bahwa metode CHAID mempunyai ketepatan klasifikasi yang lebih tinggi dalam mengklasifikasikan status angkatan kerja Kabupaten Temanggung tahun 2015.

SIMPULAN

Ketepatan metode regresi logistik dengan pembobotan sampel pada klasifikasi status angkatan kerja Kabupaten Temanggung 2015 adalah 96,4%. Ketepatan metode CHAID dengan pembobotan sampel pada klasifikasi status angkatan kerja Kabupaten Temanggung 2015 adalah 96,6%. Metode CHAID mempunyai ketepatan lebih tinggi pada klasifikasi status angkatan kerja Kabupaten Temanggung 2015. Metode regresi logistik dan CHAID dengan pembobotan sampel dapat digunakan untuk klasifikasi karena mempunyai ketepatan yang akurat, sehingga dapat dimanfaatkan instansi-instansi untuk membantu pengambilan keputusan/kebijakan.

DAFTAR PUSTAKA

Antipov, E. & Pokryshevskaya, E. (2010). Applying CHAID for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement and Analysis for Marketing*, 18, 109-117.

BPS. (2015). *Profil ketenagakerjaan Kabupaten Temanggung 2015*. Temanggung: BPS Kabupaten Temanggung.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied logistic regression* (2^{nd} *edition*). New York: John Willey and Sons Inc.

Imaslihkah, S., Ratna, M., & Ratnasari, V. (2013). Analisis regresi logistik ordinal terhadap faktor-faktor yang mempengaruhi prediksi kelulusan mahasiswa S1 di ITS Surabaya. *Jurnal Sains dan Seni POMITS* **2**: 177-182.

Kleinbaum, D.G. & Klein, M. (2010). *Logistic regression : A Self-Learning Text* (3^{rd} *edition*). New York: John Willey and Sons Inc.

Milana, N. & Abadyo. (2012). *CHAID untuk mengkalsifikasikan status mahasiswa setelah lulus perkuliahan*. Skripsi. Universitas Negeri Malang

Rahayu, R.S., Mukid M.A., & Wuryandari, T. (2015). Identifikasi faktor-faktor yang mempengaruhui terjadinya preeklampsia dengan metode CHAID. *Jurnal Gaussian*, 4(2), 383-392

Rahman, F.R. & Zain, I. (2014). Analisis regresi logistik biner untuk mengidentifikasi faktor-faktor yang mempengaruhi status penerimaan beras keluarga miskin (raskin) di Kecamatan Gunung Anyar. ITS.

Ritschard, G. (2010). *CHAID and Earlier Supervised Tree Methods*. Geneva: Universitas Geneva.

OECD. (2016). Survei ekonomi OECD: Indonesia 2016. Tersedia di https://www.oecd.org/eco/surveys/indonesia-2016-OECD-economic-survey-overview-bahasa.pdf [diakses pada 11-2-2017].