ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 5, Nomor 1, Tahun 2016, Halaman 183-192
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

KLASIFIKASI STATUS KERJA PADA ANGKATAN KERJA KOTA SEMARANG TAHUN 2014 MENGGUNAKAN METODE CHAID DAN CART

Novie Eriska Aritonang[1], Agus Rusgiyono[2], Rita Rahmawati[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Jurusan Statistika FSM Universitas Diponegoro

ABSTRACT

The growth of labor will increase along with increasing population. Increasing the number of this labor of course going to have an impact on his status, whether employ or unemployed. The method can be used to classify the status of the labor is CHAID (Chi-squared Automatic Interaction Detection) and CART (Classification and Regression Trees). Both of these methods aim to identify factors that influence employment status. These methods will be applied for Semarang labor data in 2014. Based on CHAID method, the factors that affect the status of the labor is gender, age and status of the completeness of a life partner with accuracy classification results amounted to 72.63%. Factors that affect the status of the labor force with the CART method is gender, age, educational status, and the status of the completeness of a life partner with the accuracy of the classification is 72.79%. Based on proportion test, these methods are same of doing classification employment status.

Keywords: Labor, Classification, CHAID, CART, Accuracy of classification

1. PENDAHULUAN

Peningkatan penduduk yang pesat membawa dampak pada tingkat pertumbuhan angkatan kerja. Peningkatan jumlah angkatan kerja ini berdampak pada status kerjanya, apakah bekerja atau tidak bekerja (pengangguran). Penciptaan lapangan kerja diharapkan menjadi solusi atas dampak peningkatan jumlah angkatan kerja. Ketersediaan data angkatan kerja yang di dalamnya terdapat pengelompokan penduduk bekerja dan tidak bekerja dapat membantu pemerintah mengambil tindakan yang efektif.

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia[7]. Klasifikasi dalam ilmu statistika dapat dilakukan dalam berbagai metode. Metode-metode tersebut diantaranya adalah CHAID dan CART. Kedua metode tersebut memiliki tujuan yang sama yaitu untuk kegiatan klasifikasi. Pada metode CART ini, data akan dieksplorasi untuk mengetahui variabel-variabel independen yang berpengaruh dan mengelompokkan data tersebut ke dalam kategori-kategori yang ada pada variabel dependen. Sedangkan metode CHAID bertujuan untuk menduga variabel-variabel independen yang signifikan terhadap variabel respon atau dependennya. Penulisan tugas akhir ini akan mengaplikasikan kedua metode tersebut dengan permasalahan yang dibahas adalah status kerja pada angkatan kerja Kota Semarang tahun 2014.

2. TINJAUAN PUSTAKA

2.1. Angkatan Kerja

Penduduk yang termasuk angkatan kerja adalah penduduk usia kerja (15 tahun atau lebih) yang bekerja, atau punya pekerjaan namun sementara tidak bekerja dan pengangguran[1]. Bekerja adalah kegiatan ekonomi yang dilakukan oleh seseorang dengan maksud memperoleh atau membantu memperoleh pendapatan atau keuntungan, paling sedikit 1 jam (tidak terputus) dalam seminggu yang lalu. Kegiatan tersebut termasuk pola

kegiatan pekerja tak dibayar yang membantu dalam suatu usaha atau kegiatan ekonomi. Punya pekerjaan tetapi sementara tidak bekerja adalah keadaan dari seseorang selama seminggu yang lalu bekerja tetapi sementara tidak bekerja karena berbagai sebab, seperti sakit, cuti, menunggu panenan, mogok dan sebagainya.

2.2. Skala Pengukuran

Skala pengukuran adalah peraturan penggunaan notasi bilangan dalam pengukuran. Berdasarkan skala pengukurannya, data dibedakan menjadi data nominal, data ordinal, data interval, dan data rasio[2].

2.3. CHAID (*Chi-Squared Automatic Interaction detection*)

Metode CHAID merupakan suatu metode pohon klasifikasi yang pertama kali dikenalkan oleh Dr. G. V. Kass tahun 1980 pada buku *Applied Statistics* dalam sebuah artikel yang berjudul "*An Exploratory Technique for Investigating Large Quantities of Categorical Data*". CHAID merupakan suatu teknik iteratif yang menguji variabel-variabel independen secara individual yang digunakan dalam klasifikasi dan menyusunnya pada tingkat signifikansi statistik *chi-square* terhadap variabel dependennya[3] .

2.3.1 Variabel-variabel dalam Analisis CHAID

Variabel-variabel independen kategori pada CHAID dibedakan menjadi tiga bentuk[3]. Variabel-variabel independen kategorik tersebut yaitu:

a. Variabel Independen Monotonik
   Variabel independen monotonik adalah variabel independen yang kategori di dalamnya dapat digabungkan oleh CHAID hanya jika keduanya berdekatan satu sama lain, yaitu variabel-variabel yang kategorinya mengikuti urutan aslinya (data ordinal).
b. Variabel Independen Bebas (*Free*)
   Variabel independen bebas adalah variabel independen yang kategori di dalamnya dapat digabungkan ketika keduanya berdekatan ataupun tidak (data nominal).
c. Variabel Independen Mengambang (*Floating*)
   Variabel independen mengambang adalah variabel independen yang kategori di dalamnya diperlakukan seperti monotonik kecuali untuk kategori yang terakhir (yaitu *missing value*), yang dapat berkombinasi dengan kategori manapun.

2.3.2 Uji Independensi

Langkah-langkah dalam melakukan uji independensi adalah sebagai berikut:

Hipotesis

H_0 : Tidak terdapat hubungan antara variabel pertama dan variabel kedua

H_1 : Terdapat hubungan antara variabel pertama dan variabel kedua

Statistik uji:

Perhitungan nilai χ^2_{hitung} untuk tabel kontingensi berukuran 2x2 diperoleh dari persamaan koreksi Yates. Persamaan koreksi Yates adalah sebagai berikut:

$$\chi^2_{\text{hitung}} = \frac{N\left(|O_{11}O_{22} - O_{12}O_{21}| - \frac{N}{2}\right)^2}{n_1 n_2 (O_{11}+O_{21})(O_{12}+O_{22})} \quad (1)$$

Untuk tabel kontingensi berukuran rxc, nilai χ^2_{hitung} diperoleh dari persamaan berikut:

$$\chi^2_{\text{hitung}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right] \quad (2)$$

dengan nilai E_{ij} diperoleh dari perhitungan berikut:

$$E_{ij} = \frac{n_{i+}.n_{+j}}{N} \quad (3)$$

Keputusan:

H_0 ditolak jika $\chi^2_{hitung} > \chi^2_{\alpha;(r-1)(c-1)}$ atau dengan membandingkan nilai *sig* dengan α, maka H_0 ditolak jika *sig* < α.

2.3.3 Koreksi Bonferroni

Dalam tahap penggabungan, terdapat kategori-kategori dari variabel independen yang digabung dari a kategori menjadi b kategori karena kategori tersebut tidak signifikan. Maka dari itu nilai *p-value* yang baru merupakan perkalian nilai *p-value* dengan pengali Bonferroni sesuai dengan jenis variabelnya[3] . Pengali Bonferroni untuk masing-masing jenis variabel independennya adalah sebagai berikut:

a. Variabel Independen Monotonik

$$B=\binom{a-1}{b-1} \quad (4)$$

b. Variabel Independen bebas (*Free*)

$$B=\sum_{i=0}^{b-1}(-1)^i \frac{(b-i)^a}{i!(b-i)!} \quad (5)$$

c. Variabel Independen Mengambang (*Floating*)

$$B=\binom{a-2}{b-2}+r\binom{a-2}{b-1} \quad (6)$$

2.4. CART (*Classification and Regression Trees*)

Metode CART dikembangkan oleh Leo Breiman, Jerome H. Freidman, Richard A. Olshen, dan Charles J. Stone. Metode CART merupakan suatu metodologi statistik untuk analisis klasifikasi, baik untuk variabel dependen berbentuk kategorik maupun kontinu. Metode CART akan menghasilkan pohon klasifikasi bila variabel dependennya kategorik dan pohon regresi bila variabel dependennya kontinu. Prinsip kerja dari analisis CART disebut sebagai *binary recursive partitioning*. Istilah "binary" menyatakan bahwa setiap simpul induk akan dipisah menjadi dua simpul anak. Istilah "recursive" mengacu pada proses pemisahan simpul dilakukan. Istilah "partitioning" mengacu pada data dipisah menjadi bagian-bagian atau partisi-pasrtisi yang lebih kecil[5].

2.4.1 Proses Pemecahan *Node*

Proses pemecahan pada masing-masing simpul induk didasarkan pada *goodness of split* (kriteria pemecahan terbaik)[6]. *Goodness of split* adalah suatu evaluasi pemilahan oleh pemilah s pada simpul t. Jika sebuah pemilah s dalam simpul t dibagi ke dalam t_R adalah P_R, dan ke dalam t_L dengan proporsi banyaknya objek yang dimasukkan ke dalam t_L adalah P_L, maka didefinisikan *decrease impurity* (pengurangan keragaman) adalah sebagai berikut:

$$\Delta i(s,t) = I(t) - P_R I(t_R) - P_L I(t_L) \quad (7)$$

Suatu pemilah s akan digunakan untuk memecah simpul t menjadi sua buah simpul yaitu simpul anak kiri dan simpul anak kanan jika s memaksimalkan nilai

$$\Delta i(s^*,t) = \max_s \Delta i(s,t) \quad (8)$$

Goodness of split berdasarkan pada fungsi *impurity* (fungsi keragaman). Fungsi keragaman yang digunakan dalam penelitian ini adalah indeks Gini (*Gini index*). Indeks Gini dirumuskan sebagai berikut:

$$I(t) = \sum_{i \neq j} p(Ai|t)p(Aj|t) \quad (9)$$

dimana, I(t) = fungsi keragaman indeks Gini
p(Ai|t) = peluang kelas i pada *node* t
p(Aj|t) = peluang kelas j pada *node* t

2.4.2 Pelabelan Kelas

Pelabelan kelas adalah suatu proses dimana setiap simpul pada kelas tertentu diindentifikasi[6]. Pelabelan kelas didasarkan atas jumlah anggota kelas terbanyak dirumuskan sebagai berikut:

$$P(j_0|t) = \max_j P(j|t) = \max_j \frac{N_j(t)}{N(t)} \quad (10)$$

dimana $P(j_0|t)$ = peluang kelas j_0 pada *node* t, $N_j(t)$ adalah banyaknya pengamatan di kelas j pada *node* t, dan N(t) adalah banyaknya pengamatan pada *node* t.

2.4.3 Proses Penghentian Pemecahan

Proses pemecahan akan berhenti ketika hanya ada satu pengamatan yang terdapat pada simpul terakhir, semua pengamatan yang berada dalam simpul merupakan aggota kelas yang sama (homogen), dan proses pemecahan akan berhenti apabila peneliti telah mendefinisikan sebelumnya batas akhir pembentukan pohon[5].

2.4.4 Proses Pemangkasan Pohon

Pemangkasan pohon bertujuan untuk mencegah terbentuknya pohon yang besar dan sangat kompleks. Metode yang digunakan dalam proses pemangkasan pohon didasarkan pada *minimal cost complexity prunning*, yaitu:

$$R(T) = \sum_{t\in\tilde{T}} r(t)p(t) = \sum_{t\in\tilde{T}} R(t) \quad (11)$$

$\tilde{T}$ adalah simpul-simpul akhir dari pohon klasifikasi T. *Tree misclassification cost* atau *tree resubstitution cost* (proporsi kesalahan pada sub pohon) dinotasikan dengan R(T).

$$r(t) = 1 - \max_j p(j|t) \quad (12)$$

Simpul *misclassification cost* atau r(t) adalah probabilitas kesalahan dalam melakukan klasifikasi. P(t) adalah peluang sebuah obyek akan berada dalam simpul t.

$$p(j,t) = \frac{N_j(t)}{N} \quad (13)$$

$$p(t) = \sum_j p(j,t) = p(1,t) + p(2,t) + \ldots + p(j,t) = \frac{N(t)}{N} \quad (14)$$

Peluang bahwa sebuah objek adalah anggota kelas j dan jika diketahui objek ini berada dalam simpul t disimbolkan dengan p(j|t) yang dirumuskan sebagai berikut:

$$p(j|t) = \frac{p(j,t)}{p(t)} = \frac{\frac{N_j(t)}{N}}{\frac{N(t)}{N}} = \frac{N_j(t)}{N} \times \frac{N}{N(t)} = \frac{N_j(t)}{N(t)} \quad (15)$$

Untuk memperoleh pohon hasil proses pemangkasan, perlu memperhatikan t_R yang merupakan simpul anak kanan dan t_L yang merupakan simpul anak kiri yang merupakan hasil dari pemilahan oleh simpul induk t. Apabila t, t_R, dan t_L memenuhi persamaan $R(t) = R(t_L) + R(t_R)$, maka t_L dan t_R dipangkas[6].

2.5. Ukuran Kinerja Klasifikasi

Kegiatan klasifikasi perlu diukur kinerjanya. Pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (*confusion matrix*)[7].

		Kelas hasil prediksi (j)	
		Kelas=1	Kelas=0
Kelas asli (i)	Kelas=1	f_{11}	f_{10}
	Kelas=0	f_{01}	f_{00}

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (16)$$

$$\text{Laju eror} = \frac{\text{Jumlah data yang diprediksi secara salah}}{\text{Jumlah prediksi yang dilakukan}} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (17)$$

Dimana data dari masing-masing kelas yang diprediksi secara benar yaitu ($f_{11} + f_{00}$), dan data yang diklasifikasikan secara salah yaitu ($f_{10} + f_{01}$).

Untuk mengetahui ketepatan klasifikasi dari masing-masing metode, digunakan uji beda dua proporsi. Proporsi masing-masing metode didapatkan dari perhitungan nilai akurasinya. Langkah-langkah dalam melakukan uji beda dua proporsi adalah sebagai berikut[2]:

Hipotesis:

H_0: $P_1 = P_2$ (tidak ada perbedaan signifikan dari kedua metode)

H_1: $P_1 \neq P_2$ (ada perbedaan signifikan dari kedua metode)

Taraf signifikansi: 0,05

Statistik uji:

$$Z_{hitung} = \frac{P_1 - P_2}{\sqrt{P(1-P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (18)$$

dengan:

P_1= Proporsi metode CHAID

P_2= Proporsi metode CART

n_1= Ukuran sampel pada metode CHAID

n_2= Ukuran sampel pada metode CART

P = Proporsi gabungan yaitu $\frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$

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

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data status kerja di Kota Semarang tahun 2014. Data diperoleh dari Survei Angkatan Kerja Nasional (SAKERNAS) yang dilakukan oleh Badan Pusat Statistik Provinsi Jawa Tengah. Variabel yang digunakan dalam penelitian ini adalah variabel dependen (Y) yaitu status kerja dan enam variabel (X) yaitu Status Hubungan dalam Rumah Tangga (X_1), Jenis Kelamin (X_2), Usia (X_3), Status Kelengkapan Pasangan Hidup (X_4), Status Pendidikan (X_5), dan Status Pelatihan Kerja (X_6). Software yang digunakan adalah *SPSS 16* dan *Ms. Excel 2013*. Langkah-langkah pada metode CHAID adalah sebagai berikut:

1. Memasukkan data dengan menetapkan variabel dependen dan variabel independen.
2. Membuat tabulasi silang untuk setiap kategori-kategori variabel dependen dengan kategori-kategori variabel independen.
3. Melakukan penggabungan terhadap kategori-kategori dalam variabel independen yang memiliki nilai *chi-square* terkecil.
4. Pemilihan variabel independen yang paling signifikan sebagai *split* untuk membentuk sub kelompok. Proses pemilihan variabel untuk memisah terus berjalan hingga semua sub kelompok telah dianalisis.
5. Melakukan interpretasi terhadap pohon klasifikasi yang terbentuk dan mengukur ketepatan klasifikasinya.

Sedangkan langkah-langkah pada metode CART adalah sebagi berikut:

1. Memasukkan data dengan menetapkan variabel dependen dan variabel independen.
2. Melakukan pembentukan pohon klasifikasi berdasarkan algoritma CART dengan menggunakan *software SPSS 16* dengan tahapan pembentukan pohon klasifikasi adalah sebagai berikut:
   a. Proses pemecahan *node* atau simpul
   b. Proses pelabelan kelas

c. Proses penghentian pemecahan
d. Proses pemangkasan pohon

3. Melakukan interpretasi terhadap pohon klasifikasi yang terbentuk dan mengukur ketepatan klasifikasinya.

4. HASIL DAN PEMBAHASAN

4.1. Metode CHAID

4.1.1 Penggabungan Kategori

Variabel yang memiliki lebih dari dua kategori pada pembahasan ini adalah status pendidikan. Kategori 1 untuk status pendidikan <= SD, kategori 2 untuk pendidikan SMP-SMA, dan kategori 3 untuk pendidikan D1-S3. Hasil pengujian statistik *chi-square* yang sudah dilakukan dapat dilihat dalam tabel daftar keputusan di bawah ini:

Tabel 1. Nilai statistik *chi-square* status pendidikan dan status kerja

Kategori status kerja	Kategori Status Pendidikan	Nilai χ^2_{hitung}	*Sig 2-tailed*	Keputusan
1 dan 2	1 dan 2	0,191	0,662	H_0 diterima
1 dan 2	2 dan 3	6,814	0,009	H_0 ditolak

Dari Tabel 1 diperoleh hasil bahwa nilai χ untuk <=SD (1) dan SMP-SMA (2) H_0 diterima sehingga kategori 1 dan 2 digabung menjadi kategori baru karena tidak signifikan. Selanjutnya melakukan pengujian yang sama untuk kategori gabungan dengan kategori 3 dan didapatkan bahwa kedua variabel saling bebas sehingga penggabungan telah maksimal. Pengali Bonferroni untuk variabel bebas adalah nilai perhitungan dari:

B=$\binom{a-1}{b-1}$ =$\binom{3-1}{2-1}$ = $\binom{2}{1}$ = 2

Maka nilai uji signifikansi dari hasil penggabungan kategori adalah perkalian nilai *sig (2-tailed)* dengan nilai koreksi Bonferroni, (0,009)(2) = 0,018.

Keputusan yang diambil adalah menolak H_0 karena niai *p-value* terkoreksi tetap lebih kecil dari nilai α=5%, artinya variabel status pendidikan untuk kategori campuran dan kategori 3 tidak saling bebas.

4.1.2 Uji independensi variabel independen dengan variabel dependen

Uji independensi dilakukan untuk menentukan variabel independen yang paling signifikan pertama kali sebagai pemilah utama. Hasil dari uji independensi dapat dilihat pada tabel di bawah ini:

Tabel 2. Uji independensi variabel independen dan variabel dependen

Status Kerja	Variabel Independen	Kategori Variabel Independen	Nilai *chi-square*	*Sig (2-tailed)*	Keputusan
1 dan 2	Status hubungan dalam RT	1 dan 2	49,445	0,000	Ho ditolak
1 dan 2	**Jenis kelamin**	**1 dan 2**	**113,076**	**0,000**	**Ho ditolak**
1 dan 2	Usia	1 dan 2	38,143	0,000	Ho ditolak
1 dan 2	Status kelengkapan pasangan hidup	1 dan 2	3,872	0,049	Ho ditolak
1 dan 2	Status pendidikan	Gabungan 1,2 dan 3	6,583	0,009	Ho ditolak
1 dan 2	Status pelatihan kerja	1 dan 2	0,026	0,872	Ho diterima

Dari tabel di atas diperoleh variabel jenis kelamin memiliki nilai *chi-square* terbesar yaitu 113,076 dan *p-value* 0,000. Variabel jenis kelamin merupakan pemilah utama karena memiliki nilai *chi-square* terbesar dibandingkan variabel lainnya. Proses pemilahan terus dilakukan pada setiap simpul selama masih terdapat variabel-variabel independen yang signifikan.

4.1.3 Hasil Klasifikasi

Klasifikasi status kerja yang telah dilakukan perlu diuji tingkat akurasinya dalam melakukan pengelompokkan data.

Tabel 3. Matriks konfusi hasil klasifikasi

		Kelas hasil prediksi (j)	
		Bekerja	Tidak bekerja
Kelas asli (i)	Bekerja	868	36
	Tidak bekerja	311	53

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}} = \frac{868+53}{868+311+36+53} = 0{,}7263$$

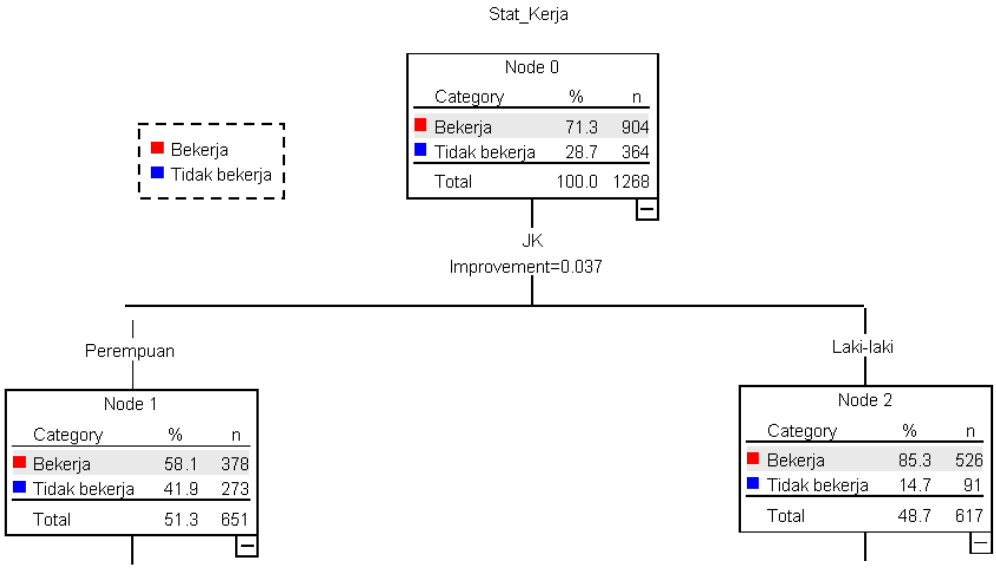
$$\text{Laju eror} = \frac{\text{Jumlah data yang diprediksi secara salah}}{\text{Jumlah prediksi yang dilakukan}} = \frac{311+36}{868+311+36+53} = 0{,}2737$$

Akurasi hasil klasifikasi data status kerja pada angkatan kerja di Kota Semarang tahun 2014 sebesar 0,7263 atau 72,63% dengan laju eror sebesar 0,2737 atau 27,37%.

4.2. Metode CART

4.2.1 Proses Pemecahan Simpul

Proses pembentukan pohon dimulai dengan menentukan pemilah utama pada simpul induk yang akan dipecah menjadi simpul anak kiri dan simpul anak kanan yang disebut pengurangan keragaman. Kandidat pemilah yang ada akan dipilih salah satunya berdasarkan nilai *goodness of split* terbesar dengan kriteria pemilahan menggunakan indeks Gini. Variabel jenis kelamin merupakan pemilah utama yang terpilih karena memiliki nilai *goodness of split* terbesar. Proses pemecahan simpul digambarkan seperti gambar di bawah ini:



Nilai *goodness of split* untuk masing-masing pemilah dapat dilihat pada tabel di bawah ini:

Tabel 4. Nilai *goodness of split* masing-masing pemilah

No	Variabel	Simpul kiri	Simpul kanan	*Goodness of split*	N kiri	N kanan
1	**Jenis kelamin**	**Perempuan**	**Laki-laki**	**0,03693**	**651**	**617**
2	Status hubungan dalam RT	Bukan kepala RT	Kepala RT	0,01626	816	452
3	Usia	Bukan usia produktif	Usia produktif	0,01258	414	854
4	Status pendidikan	<=SD	SMP-SMA, D1-S3	0,00001	351	917
		SMP-SMA	<=SD, D1-S3	0,00096	736	532
		D1-S3	<=SD, SMP-SMA	0,00227	181	1087
5	Stat kelengkapan pasangan hidup	Tidak beristri/suami	Beristri/suami	0,00134	383	885
6	Status pelatihan kerja	Tidak pernah	Pernah	0,00002551	1198	70

Proses pemecahan simpul terus berjalan terhadap semua simpul dan berhenti jika hanya ada satu pengamatan yang terdapat pada simpul terakhir, semua pengamatan yang berada dalam simpul merupakan aggota kelas yang sama (homogen), dan bila peneliti telah mendefinisikan sebelumnya batas akhir pembentukan pohon. Dalam proses pemecahan simpul juga berlangsung proses pelabelan kelas pada masing-masing simpul yang didasarkan atas peluang terbesar dari setiap kelas.

4.2.2 Proses Pemangkasan Pohon

Pemangkasan pohon bertujuan untuk menyederhanakan bentuk dari pohon maksimal yang terbentuk. Proses pemangkasan pohon dimulai dengan mengambil t_L yang merupakan simpul anak kiri dan t_R yang merupakan simpul anak kanan. Simpul anak kiri (t_L) dan simpul anak kanan (t_R) akan dipangkas apabila memenuhi persamaan $R(t) = R(t_L) + R(t_R)$. Proses pemangkasan pohon menghasilkan pohon hasil pemangkasan. Simpul-simpul yang mengalami proses pemangkasan dan yang tidak mengalami proses pemangkasan dapat dilihat pada Tabel 5.

Tabel 5. Hasil proses pemangkasan pohon

No	Simpul terkait	R(t)	$R(t_L) + R(t_R)$	Keterangan
1	Simpul 1 (t), simpul 3 (t_L), dan simpul 4 (t_R)	0,21530	021215	Tidak dipangkas
2	Simpul 2 (t), simpul 5 (t_L), dan simpul 6 (t_R)	0,07177	0,07177	Pangkas
3	Simpul 3 (t), simpul 7 (t_L), dan simpul 8 (t_R)	0,12855	0,12855	Pangkas
4	Simpul 4 (t), simpul 9 (t_L), dan simpul 10 (t_R)	0,08360	0,07729	Tidak dipangkas
5	Simpul 9 (t), simpul 19 (t_L), dan simpul 20 (t_R)	0,07413	0,06861	Tidak dipangkas
6	Simpul 10 (t), simpul 21 (t_L), dan simpul 22 (t_R)	0,00315	0,00315	Pangkas
7	Simpul 19 (t), simpul 37 (t_L), dan simpul 38 (t_R)	0,04259	0,04259	Pangkas
8	Simpul 20 (t), simpul 39 (t_L), dan simpul 40 (t_R)	0,02603	0,02445	Pangkas

4.2.3 Hasil Klasifikasi

Klasifikasi status kerja yang telah dilakukan perlu diuji tingkat akurasinya dalam melakukan pengelompokkan data.

Tabel 6. Matriks konfusi hasil klasifikasi

		Kelas hasil prediksi (j)	
		Bekerja	Tidak bekerja
Kelas asli (i)	Bekerja	871	33
	Tidak bekerja	312	52

$$\text{Akurasi}=\frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah prediksi yang dilakukan}}=\frac{871+52}{871+312+33+52}=0{,}7279$$

$$\text{Laju eror}=\frac{\text{Jumlah data yang diprediksi secara salah}}{\text{Jumlah prediksi yang dilakukan}}=\frac{312+33}{871+312+33+52}=0{,}2721$$

Akurasi hasil klasifikasi data status kerja pada angkatan kerja di Kota Semarang tahun 2014 sebesar 0,7279 atau 72,79% dengan laju eror sebesar 0,2721 atau 27,21%.

4.3. Evaluasi Ketepatan Klasifikasi

Untuk mengetahui metode yang tepat dalam melakukan klasifikasi status kerja dari angkatan kerja, maka dilakukan evaluasi ketepatan klasifikasi dengan melakukan uji beda dua proporsi.

Hipotesis

H_0: $P_1 = P_2$ (tidak ada perbedaan signifikan dari kedua metode)

H_1: $P_1 \neq P_2$ (ada perbedaan signifikan dari kedua metode)

Taraf signifikansi: 0,05

Statistik uji:

$$P=\frac{(1268\ x\ 0{,}72634)+(1268\ x\ 0{,}72792)}{1268+1268}=\frac{1844{,}00168}{2536}=0{,}72713$$

$$Z_{\text{hitung}}=\frac{0{,}72634-0{,}72792}{\sqrt{0{,}72713(1-0{,}72713)\left(\frac{1}{1268}+\frac{1}{1268}\right)}}$$

$= -0{,}08931$

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

Keputusan: Karena $Z_{hitung} = -0{,}08931 > -Z_{\alpha/2} = -1{,}96$, maka H_0 diterima.

Kesimpulan: Pada taraf signifikansi 0,05 didapatkan bahwa tidak ada perbedaan signifikan dari kedua metode. Dengan kata lain, tidak terdapat perbedaan antara metode CHAID dan CART dalam melakukan klasifikasi status kerja dari angkatan kerja. Jadi, kedua metode metode ini mempunyai ketepatan yang relatif sama.

5. KESIMPULAN

Berdasarkan hasil dan pembahasan diperoleh kesimpulan hasil klasifikasi status kerja pada angkatan kerja Kota Semarang tahun 2014 menggunakan metode CHAID dan CART yaitu:

1 Banyak kelas yang dihasilkan dari proses klasifikasi dengan metode CHAID adalah 8 kelas. Delapan kelas yang dihasilkan merupakan simpul akhir yang akan merepresentasikan karakteristik dari angkatan kerja yang bekerja dan tidak bekerja. Ketepatan klasifikasi yang dihasilkan dengan metode CHAID adalah 72,63%.

2 Banyak kelas yang dihasilkan dari proses klasifikasi dengan metode CART adalah 5 kelas. Lima kelas ini merupakan simpul akhir yang merepresentasikan karakteristik angkatan kerja yang bekerja dan tidak bekerja yang didapatkan setelah pohon berhasil dipangkas. Ketepatan klasifikasi yang dihasilkan dengan metode CART adalah 72,79%.

3 Ketepatan hasil klasifikasi dengan metode CART lebih tinggi dibandingkan dengan menggunakan metode CHAID. Ketepatan hasil klasifikasi dengan metode CART adalah sebesar 72,79% sedangkan dengan metode CHAID 72,63%. Dari uji proporsi yang dilakukan didapatkan bahwa tidak terdapat perbedaan yang signifikan dari kedua metode ini. Jadi, kedua metode mempunyai ketepatan yang relatif sama.

DAFTAR PUSTAKA

[1] Badan Pusat Statistik (BPS) Provinsi Jawa Tengah. 2015. *Profil Ketenagakerjaan Provinsi Jawa Tengah Tahun 2014*.

[2] Kass, G.V. 1980. *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. *Applied Statistics 29*, No. 2; 119-127

[3] Gallagher, C.A., Monroe, H. M., Fish, J. L. 2000. *An Iterative Approach to Classification Analysis*. www.casact.org/library/ratemaking/90dp237.pdf. (diakses tanggal 15 November 2014).

[4] Conover, W.J. 1971. *Practical Nonparametric Statistics.* John Wiley & Sons.Inc.

[5] Lewis, R. J. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis*. Presented at the 2000 Anual Meeting of Society For Academy Emergency Medicine in San Fransisco, California.

[6] Breiman, L., Friedman, J.H., Olshen, R.A. dan Stone, C.J. (1984). *Classification And Regression Tree*. New York, NY: Chapman And Hall

[7] Prasetyo, E. 2012. *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: C.V Andi Offset.