E-ISSN 2527-9378

Jurnal Statistika Industri dan Komputasi
Volume 03, No. 2, Juli 2018, pp. 22-32

Implementasi Metode *Decision Tree* dengan Algoritma ID3 dan C4.5 untuk Mengklasifikasikan Partisipasi Perempuan Nikah dalam Kegiatan Ekonomi Rumah Tangga di DIY

Amiludin Bukhori[1], Noviana Pratiwi[2*]

[1,2]Jurusan Statistika, Fakultas Sains Terapan, Institut Sains dan Teknologi AKPRIND, Yogyakarta
Email : bukhori04@gmail.com[1], novianapratiwi@akprind.ac.id[2]

Abstract : *Data mining is a process in finding various models, summary data and valuable information in a set of data. There are several data analysis techniques in data mining, one of which is classification. A fairly simple method of classification is the decision tree. The Decision tree has the ability to transform large data into tree shapes that represent kinds of easily understood conditions. National Labor Force Survey is the annual employment survey of BPS DIY. The Sakernas's database contains a wealth of employment data, one of which is female workers. The existence of female workers in DIY can not be underestimated. Labor Force Participation Rate (LFPR) of DIY's women in February 2017 amounted to 63.29%. 70.93% of female workers are married women. This shows that the rate of marriage women participation in household economic activities is quite high. In this research the decision tree was built using the ID3 and C4.5 algorithms to clarify the participation of married women in household economic activities in Special Region of Yogyakarta. The results showed that the participation of married women in household economic activities in DIY is better explained by the decision tree with C4.5 algorithm whose accuracy rate is 68.71% than the ID3 algorithm whose accuracy rate is 68.10%. By using C4.5 algorithm it is known that the variable that most influence women to participate in household economic activity is husband job status. There are several conditions that can explain the participation of marriage women in household economic activities in DIY, among which is married women generally work when under 83 years old and have a husband who works as a laborer / employee / employee.*

Key Words : Married Women, Labor Force Participation Rate, Data mining, Decision Tree. ID3 Algorhitm and C4.5 Algorithm.

Abstrak: *Data mining* adalah suatu proses dalam menemukan berbagai model, ringkasan data dan informasi-informasi yang berharga dalam sekumpulan data. Terdapat beberapa teknik analisis data dalam *data mining*, salah satunya klasifikasi. Metode klasifikasi yang cukup sederhana adalah *decision tree*. *Decision tree* memiliki kemampuan untuk mengubah data yang besar menjadi pohon keputusan yang merepresentasikan berbagai kondisi yang mudah dipahami. Survey Angkatan Kerja Nasional (SAKERNAS) merupakan survei tahunan BPS DIY tentang ketenagakerjaan. Database Sakernas mengandung banyak sekali data mengenai ketenagakerjaan salah satanya ialah tentang pekerja perempuan. Keberadaan pekerja perempuan di DIY tidak dapat dipandang sebelah mata. Tingkat Partisipasi Angkatan Kerja (TPAK) perempuan DIY pada bulan Februari 2017 sebesar 63,29%. 70,93% dari tenaga kerja perempuan merupakan perempuan yang telah menikah. Hal ini Menunjukkan bahwa tingkat pertisipasi perumpuan nikah dalam kegiatan ekonomi rumah tangga cukup tinggi. *Dalam penelitian ini decision tree dibangun dengan menggunakan algoritma ID3 dan C4.5* untuk mengkalisifikasikan partisipasi perempuan nikah dalam kegiatan ekonomi rumah tangga di Daerah Istimewa Yogyakarta. Hasil penelitian menunjukkan bahwa partisipasi perempuan nikah dalam kegiatan ekonomi rumah tangga di DIY dijelaskan lebih baik oleh pohon keputusan dengan algoritma C4.5 dengan tingkat akurasi 68,71% dari pada algoritma ID3 dengan tingkat akurasi 68,10%. Dengan algoritma C4.5 diketahui bahwa variabel yang paling mempengaruhi perempuan untuk ikut serta dalam kegiatan ekonomi rumah tangga adalah status pekerjaan suami. Terdapat beberapa kondisi yang bisa menjelaskan tentang partisipasi perempuan nikah dalam kegiatan ekonomi rumah di DIY, diantaranya ialah perempuan nikah umumnya bekerja ketika usia di bawah 83 tahun dan memiliki suami yang berprofesi sebagai buruh/pegawai/karyawan.

Kata kunci : Tingkat Partisipasi Perempuan Nikah, Data Mining, Decision Tree, Algoritma ID3 dan Algoritma C4.5.

*Corresponding author's email: novianapratiwi@akprind.ac.id

1. PENDAHULUAN

Peranan perempuan sebagai mitra yang sejajar dengan laki-laki pada saat ini bukan merupakan suatu hal yang baru. Menurut hukum, perempuan dan laki-laki di Indonesia mempunyai peluang yang sama untuk berpartisipasi dalam proses pembangunan di semua bidang kehidupan. Pernyataan tersebut tertuang dalam UUD 1945 Pasal 27 [5]**.** Secara agregat, partisipasi perempuan dalam kegiatan ekonomi dapat diukur dari Tingkat Partisipasi Angkatan Kerja (TPAK) Perempuan [5]. TPAK merupakan persentase jumlah angkatan kerja terhadap penduduk usia kerja. TPAK perempuan pada bulan Februari 2017 sebesar 63,29%, angka ini mengalami kenaikan 1,19% dari bulan Agustus 2016. Namun, bila dibandingkan dengan bulan Februari 2016 TPAK perempuan mengalami sedikit penurunan sebesar 0,16%, Dapat disimpulkan bahwa lebih dari separuh perempuan angkatan kerja di DIY berpartisipasi di kegiatan ekonomi di DIY. Dari keseluruhan perempuna yang bekerja, 70,93% telah menikah.
Dengan banyaknya perempuan yang bekerja maka akan timbulah masalah dalam internal keluarga seperti hal kepengurusan rumah tangga. Banyak pihak lain yang kemudian mengambi beberapa peran ibu rumah tangga dalam megurusi rumah. Dan hal ini mengakibatkan munculah lapangan pekerjaan baru. Untuk meningkatkan perekonomian makro dan mikro serta kesejahteraan masyarakat maka perlu dilihat klaifikasi TPAK di DIY. Klasifikasi merupakan pengelompokan sampel berdasarkan ciri-ciri pesamaan dan perbedaan dengan menggunakan variable target sebagai kategori(Larose, 2005). *Decision tree* merupakan pohon keputusan yang terdiri dari varibel akar (*root*), tangkai (*nood*) dan daun (*leaf*), daun merupakan ujung pohon keputusan yang nantinya akan menggambarkan hasil klasifikasinya. Pada penelitian ini, *decision tree* menggunakan algoritmaID3dan C4,5.

Sebelumnya telah ada penelitian oleh Vildha Indriyani Riyanto (2015) tentang analisis faktor-faktor yang mempengaruhi partisipasi kerja perempuan dalam kegatan ekonomi rumah tangga di kota Semarang dan di dapat kesimpulan bahwa variabel umur dan jumlah anggota keluarga berpngaruh terhadap partisipasi kerja perempuan nikah di Kota Semarang. Penelitian yang serupa juga perlah dilakukan oleh Devima Christi Mukti Rantau dan Dr. Dra. Ismaini Zain, M. Si. (2013) di Provinsi Jawa Timur dengan kesimpulan pada pemodelan didapatkan lima variabel yang berpengaruh terhadap keputusan partisipasi perempuan kawin dalam ekonomi, yaitu umur, pendidikan akhir, jumlah anak balita, pendidikan akhir suami dan status kerja suami.

2. METODE PENELITIAN

2.1 Objek Penelitian

Penelitian ini dilakukan di Daerah Istimewa Yogyakarta, dengan pokok bahasan adalah tentang partisipasi perempuan dalam kegiatan ekomoni rumah tangga di DIY dengan menggunakan metode *decision tree* algoritma ID3 dan C4.5.

2.2 Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari Badan Pusat Statistik Daerah Istimewa Yogyakarta berupa data SAKERNAS dengan beberapa variabel.

2.3 Variabel Penelitian

Variabel terdiri dari dua yaitu variabel target (Y) yaitu Partisipasi Perempuan Nikah , dan empat variabel prediktor (X) yang terdiri dari Umur, Jumlah Anggota Keluarga, Tingkat Pendidikan, Tingkat Pedidikan Suami dan Status Kerja Suami.

2.4 Metode Analisis Data

2.4.1 *Datamining*

Data mining adalah suatu proses dalam menemukan berbagai model, ringkasan data dan informasi-informasi yang berharga didalam sekumpulan data. Berbagai disiplin ilmu dan teknolgi melakukan pendekatan terhadap *data mining*, salah satunya statistik yang menjadikan *data mining* lebih optimal dalam menyerap dan menggambarkan informasi dan melakukan pendugaan. *Data mining* memiliki dua fungsi yaitu analisis deskriptif dan analisis prediktif. Analisis deskriptif adalah analisis yang merepresentasikan informasi yang ditemukan tanpa harus melakukan pemodelan dengan hasil yang lebih spesifik, pada analisis deskriptif tidak terdapat variabel target yang akan diprediksi..

The Cross Industry Standard Process for Data Mining (CRISP-DM, 1996) menyediakan sebuah kerangka kerja yang umum dan dikembangkan dengan baik untuk menjalankan proyek data mining. CRISP-DM merumuskan enam langkah yang khas dalam proses data mining[7]yaitu :

1 *Problem Understanding* (Pemahaman Masalah).
2 *Data Understanding* (Pemahaman Data).
3 *Data Preparation* (Persiapan Data).
4 *Modeling* (Pemodelan).
5 *Evaluation* (Evaluasi Model).
6 *Deployment* (Penyebaran

2.4.2 *Normalitas Multivariate*

Pengujian *normalitas* yang dimaksud adalah *multivariate Gaussian distribution*, uji *multivariate Gaussian distribution* menguji beberapa variabel sekaligus. Shapiro Wilk merupakan test yang optimal untuk test univariat dengan sampe kurang dari 50,dengan modifikasi oleh Rosyton, *Shapiro-Wilk* dapat digunakan untuk test multivariate pada sampel besar (3 < n < 5000). Rumus metode *Shapiro-Wilk* adalah sebagai berikut [5]:

$$W(i) = \frac{\left\{\sum_{j=1}^{N} a_j (x_{n-i+1} - x_j)\right\}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)} \quad (1)$$

Nilai setiap *i*, *W(i)* dapat ditransformasi ke dalam bentuk standar normal *G(W(i)* dengan menggunakan system S_B Johnson:

$$G\big(W(i)\big) = \gamma + \delta \log\left\{\frac{W(i) - \varepsilon}{1 - W(i)}\right\} \quad (2)$$

Dimana γ, δ dan ε dapat ditemukan di Tabel Shapiro Wilk (1968) dengan nilai n = 50, For n > 50, nilai dari γ, δ dan ε dapat ditemukan pada Shapiro and Francia (1972) dan Royston (1972). Untuk menguji normalitas multivariate selanjutnya dilakukan perhitung :

$$M_1 = -2 \sum_{i=1}^{P} \ln\{\Phi(\mathrm{G}(\mathrm{W(i)}))\} \quad (3)$$

dimana

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} exp\left\{-\frac{1}{2}t^2\right\} dt$$

M_1 merupakan test statistic untuk menguji normalitas multivariat, dimana M_1 berdistribusi $\chi^2_{2,p}$.

2.4.3 Deteksi Outlier

Dalam penelitian ini deteksi outlier menggunakan teknik statistika yaitu dengan menggunakan jarak mahalnobis. Jarak *mahalanobis (The Mahalanobis Distance)* untuk tiap observasi dapat dihitung dan akan menunjukkan jarak sebuah observasi dari rata-rata semua variabel dalam sebuah ruang multidimensional. Jarak *mahalanobis (The Mahalanobis Distance)* dapat dihitung dengan rumus [2]:

$$M_i = \left(\sum_{i=1}^{n} (x_i - \bar{x})^T V_2^{-1} (x_i - \bar{x}) \right)^{1/2} \quad (4)$$

Dengan V_n merupakan matriks *covarian* dengan rumus:

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (x_i - \bar{x})^T \quad (5)$$

2.4.4 *Missing Value*

Missing value merupakan kekosongan data yang terjadi akibat tidak adanya informasi dari suatu objek bukan karena kesalahan individu dalam proses pengumpulan data. Metode yang digunakan untuk menangani *missing value* pada penelitian ini adalah dengan cara memberi nilai pada observasi dengan *missing value* menggunakan niai sentralitas statistic (*mean*, *median*, *modus*). Penanganan missing value tergantung pada skala data yang digunakan, jika data berskala nominal atau ordinal maka menggunakan modus,jika data berskala interval atau rasio makan menggunakan mean jika data berdistribusi normal atau median jika data tidak berdistribusi normal.

2.4.5 Algoritma ID3

Algoritma ID3 melakukan prosedur pencarian secara menyeluruh pada semua kemungkinan pohon keputusan. Penentuan penempatan varabel-variabel prediktor pada *root* (akar) dan *node* (tangkai) dilakukan dengan mengevaluasi semua atribut yang ada dengan menggunakan suatu ukuran statistik yang banyak digunakan yaitu information gain untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan kumpulan data.

Pada prakteknya ukuran pohon keputusan ditentukan oleh dataset yang digunakan, pohon keputusan bisa menjadi terlalu besar dan sulit untuk diinterprestasikan jika dataset yang digunakan berukuran besar. Dalam mengatasi pohon yang terlalu besar maka dilakukan *pruning*. Terdapat dua teknik *pruning* yaitu *pre pruning* dan *post pruning*. Pada penelitian ini mengunakan *pre pruning*, *pre pruning* merupakan teknik yang efisien karena dilakukan saat proses pohon keputusan dibangun [6].*Pre pruning* dilakukan dengan cara berhenti memecah *subset* data latih pada *node* tertentu. Berikut ini parameter dalam *pre pruning*:

a. *Minimal gain*
b. Minimal leaf size
c. Minimal size for split

Secara ringkas, langkah-langkah analisis algoritma ID3 dapat digambarkan sebagai berikut (Defiyanti dalam Sidette *et al*, 2014):

1. Input data yang akan digunakan, tentukan variabel target dan variabel prediktor.
2. Hitung *entropy* dan *information gain* dari setiap atribut. Nilai *entropy* setiap variabel prediktor digunakan untuk menghitung *information gain* yang akan digunakan untuk pemilihan variabel yang digunakan sebagar akar dari pohon keputusan (*decision tree*).
3. Bentuk smpul yang berisi atribut tersebut.
4. Ulangi proses perhitungan sampai semua data telah termasuk dalam kelas yang sama. Atribut yang telah dipilih sebagai simpul tidak diikutkan lagi dalam perhitungan selanjutnya.

2.4.6 AlgoritmaC4.5

Algoritma C4.5 diperkenalkan oleh Quinlan sebagai versi perbaikan dari ID3. Pada algoritma C4.5 terdapat beberapa perbaikan yang dilakukan untuk memperbaiki kekurangan yang terdapat pada agoritma ID3, beberapa perbaikan yang dilakukan yaitu mampu menangani data berskala interval atau rasio, mampu menangani *missing value* dan mampu melakukan *post pruning*.

Pada penelitian ini mengunakan *pre pruning*, *pre pruning* merupakan teknik yang efisien karena dilakukan saat proses pohon keputusan dibangun [6]**.***Pre pruning* dilakukan dengan cara berhenti memecah *subset* data latih pada *node* tertentu. Berikut ini parameter dalam *pre pruning*:

1. *Maximal dept Minimal*
2. *Minimal gain*
3. Minimal leaf size
4. Minimal size for split
5. *Number of prepruning alternatives*

2.4.7 *Entropy*

Entropy merupakan distribusi probabilitas dalam teori informasi dan diadopsi ke dalam algoritma untuk mengukur tingkat homogenitas distribusi kelas dari sebuah himpunan data (*data set*). Rumus *entropy* adalah sebagai berikut [4]:

$$Entropy(S) = -\sum_{i=1}^{n} \mathrm{p}_i \, log_2 p_i \quad (6)$$

dengan:

S : Himpunan Kasus
n : Jumlah partisi *S*
pi : Proporsi dari Si terhadap *S*

Jika terdapat kasus yang berbentuk 0 log 0 maka hasilnya dianggap 0 **(Nowozi, 2012).**

2.4.8 *Information Gain*

Information gain ini diukur dengan menghitung selisih antara entropi *data set* sebelum dan sesudah pembagian (*splitting*) dilakukan. Rumus *information gain* adalah sebagai berikut [4] :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|s|} \; x \; Entropy(Si) \quad (7)$$

dengan:

S : Himpunan kasus
Si : Himpunan kasus pada partisi ke *i*
A : Variabel
n : Jumlah partisi atribut *A*
|*Si*| : Jumlah kasus pada partisi ke *i*
|*S*| : Jumlah kasus dalam *S*

2.4.9 *Split Info*

Split Info merupakan entropi dari seluruh distribusi probabilitas *subset* setelah dilakukan pemartisian [1]. Rumus *Split Info* adalah sebagai berikut [4]:

$$Split\ Info(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (8)$$

dimana:

S : ruang (data) sample yang digunakan untuk training.
A : atribut.
S_i : jumlah sample untuk atribut i

2.4.10 *Gain Ratio*

Perhitungan *information gain* masih memiliki sejumlah kekurangan. Salah satu kekurangan yang mungkin terjadi adalah pemilihan atribut yang tidak relevan sebagai pemartisi yang terbaik pada suatu simpul. *Gain ratio* merupakan normalisasi dari *information gain* yang memperhitungkan entropi dari distribusi probabilitas subset setelah dilakukan proses partisi. Rumus *gain ratio* adalah sebagai berikut [4]:

$$Gain\ Ratio(A) = \frac{Information\ Gain\ (S,A)}{Split\ Info\ (S,A)} \quad (9)$$

dimana :

S	: Himpunan Kasus
A	: Variabel
$Information\ Gain\ (S, A)$	: *Information Gain* pada atribut A
$Split\ Info\ (S, A)$	: *Split Info* pada atribut A

2.4.11 K-FordValidation

Cross validation adalah metode yang umum digunakan untuk mengevaluasi kinerja *classifier* dengan menggunakan keseluruhan data. Salah satu pendekatan pada metode *cross validation* yaitu *k-ford validation*. Pada motode *k-validation* keseluruhan data dibagi nenjadi k partisi yang berukuran kira-kira sama. Selama proses, salah satu dari partisi dipilih untuk data *testing*, sedangkan sisa pertisi lainnya digunakan untuk data *training*. Sebagai contoh apabila terdapat subset D_1, D_2, ..., D_k, maka untuk iterasi pertama, subset D_1 digunakan sebagai *data testing* sedangkan sisanya D_2, ..., D_k digunakan sebagai *data training* untuk memperoleh model pertama, begitu juga untuk iterasi kedua, maka D_2 digunakan untuk *data testing* dan sisanya D_1, D_3, ..., D_k digunakan untuk *data training*, begitu seterusnya sampai subset terakhir D_k.

2.4.12 Confusion Matriks

Confusion Matrix merupakan tabel pencatat hasil kerja klasifikasi. Tabel 1 merupakan contoh tabel confusion matriks klasifikasi dua kelas, hanya ada 2 kelas yaitu kelas 0 dan kelas 1.

Tabel 1 Contoh Tabel *Confusion Matrix*

		Prediksi(j)	
		Kelas = 1	Kelas = 0
Aktual(i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Keterangan :

f_{11} = jumlah data kelas 1 yang diklasifikasikan secara benar pada kelas 1.
f_{01} = jumlah data kelas 0 yang diklasifikasikan secara salah pada kelas 1.
f_{10} = jumlah data kelas 1 yang diklasifikasikan secara salah pada kelas 0.
f_{00} = jumlah data kelas 0 yang diklasifikasikan secara benar pada kelas 0.

Salah satu pesyaratan standar yang telah didefinisikan untuk matriks klasifikasi dua kelas :

Untuk menghitung akurasi digunakan formula :

$$Akurasi = \frac{f_{11} + f_{00}}{f_{11} + f_{10}\ f_{01} + f_{00}} \quad (10)$$

Untuk menghitung error(kesalahan prediksi) digunakan formula :

$$Kesalahan\ prediksi = \frac{f_{10} + f_{01}}{f_{11} + f_{10}\ f_{01} + f_{00}} \quad (11)$$

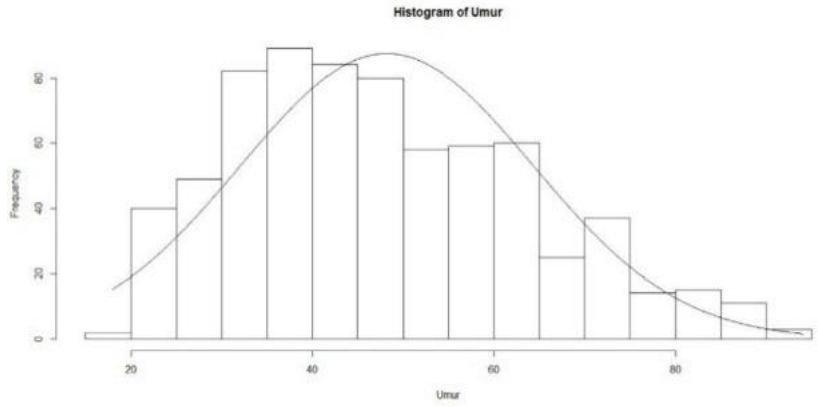
3. HASIL DAN PEMBAHASAN

3.1 Persiapan Data

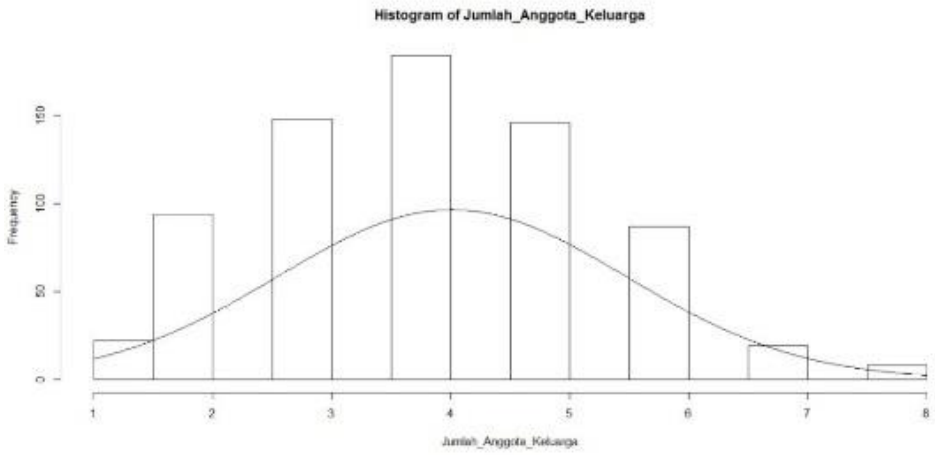
3.2.1 Analisis Deskriptif

Dari 708 perempuan yang sudah menikah, perempuan yang bekerja memiliki persentase sekitar 68% atau sebanyak 484 orang. Sedangkan yang tidak bekerja memiliki persentase sebesar 22,24% atau sebanyak 224 orang. Hal ini menunjukkan bahwa partisipasi ekonomi perempuan nikah di DIY cukup tinggi. mayoritas perempuan nikah yang bekerja berumur 33 – 52 tahun, mayoritas perempuan nikah di DIY terdiri dari perempuan nikah yang tidak memiliki ijazah pendidikan dan berpendidikan akhir sekolah dasar, mayoritas rumah tangga memiliki jumlah anggota 3 sampai 4 orang dan mayoritas rumah tangga memiliki memiliki suami yang berpofesi sebagai buruh/karyawan/pegawai.

3.2.2 Deteksi Ourliers



Gambar 1 Variabel umur



Gambar 2 Variabel Jumlah Anggota Keluarga

Secara visual sebaran data variabel umur dan jumlah anggota keluarga(kanan) tidak memusat pada rata-rata melainkan berpusat melenceng ke kiri atau bisa disebut *skewness*.

3.2.3 *Missing Value*

Tabel 2 *Missing Value*

Jumlah	Partisipasi permpuan nikah	Umur	Anggota keluarga	Tingkat Pendidikan	Tingkan pendidikan suami	Status kerja suami
539	1	1	1	1	1	1
169	1	1	1	1	0	0
0	0	0	0	0	169	169

Tabel 2 menunjukan bahwa terdapat dua pola data, pola pertama menunjukkan bahwa terdapat 529 pengamatan tanpa *missing value* dari keenam variabel, pola kedua, terdapat nilai yang hilang pada variabel tingkat pendidikan suami dan status kerja suami. Ada 169

pengamatan yang terdapat *missing value* pada variabel tingkat pendidikan suami dan status kerja suami

3.2 Cleaning Data

3.2.1 Outliers

Deteksi outlier dengan ,menggunakan mahalanobis distance, teknik ini mesyaratkan datanya berdistriobusi normal multivariate, oleh karena itu dilakukan uji normalitas multivariate dengan menggunakan metode *Shapiro-wilk.*

Tabel 3 Uji Normalitas

Variabel	W	Sig.
Umur	0,99165	0,000504
Jumlah Anggota Keluarga		

Tabel 3 menunjukkan nilai signifikansi kedua variabel adalah 0,000504 sedangkan alpha yang digunakan 0,05. Nilai signifikansi lebih kecil dari pada alpha maka keputusannya tolak H_0 atau data tidak berdistribusi normal multivariate.

Data umur dan jumlah anggota keluarga tidak berdistribusi normal multivariate, maka dilakukan tranasformasi. Transformasi yang dilakukan adalah *square root transformation* karena datanya *skewness*.

Tabel 4Uji Normalitas

Variabel	W	Sig.
Umur	0,99668	0,1503
Jumlah Anggota Keluarga		

Tabel 4 menunjukkan nilai signifikansi kedua variabel adalah 0,153 sedangkan alpha yang digunakan 0,05. Nilai signifikansi lebih besar dari pada alpha maka keputusannya H_0 tidak ditolak atau data berdistribusi normal multivariate.

Setelah data berdistribusi normal multivariate,selanjutnya dilakukan pendeteskian outliers dengan menggunakan *malanobis distance*. Berdasarkan hasil dari uji menggunakan *mahalanobis distance* dengan nilai chi square 5,99, terdapat 35 nilai *outlier* dari data yang digunakan. Penanganan terhadap data *outlier* tersebut dengan cara mengapusnya dari dalam data penelitian. Penanganan dari data outlier yaitu dengan cara menghapusnya dari data.

3.2.2 Missing Value

Berdasarkan tabel 2 terdapat *missing value* pada variabel tingkat pendidikan suami dan status pekerjaan suami. Kedua variabel tersebut berskala ordinal oleh karena itu *missing value* diisi dengan menggunakan nilai yang sering muncul(modus). Modus dari tingkat pendidikan yaitu 5 yang berarti sekolah dasar/madrasah ibtidaiyah dan modus dari variabel status pekerjaan suami adalah 5 yang berarti buruh/karyawan/pegawai.

3.3 Transformasi Data

3.3.1 Discretization

Dalam penelitian ini, pohon keputusan dibangun menggunakan algoritma ID3 dan algoritma C4.5, algoritma ID3 mengharuskan setiap variabel berbentuk data kategorik. Terdapat dua variabel berbentuk numerik yaitu umur dan jumlah anggota rumah tangga. Transformasi data dilakukan dengan tujuan untuk merubah data numerik menjadi kategorikal dan meringkas kategori. Berikut merupakan *transformasi* data dengan menggunakan kategori data yang telah ada, umur dikategorikan berdasarkan kategori yang dibuat oleh Badan Pusat

Statistik (BPS) dan jumlah anggota rumah tangga dikategorikan berdasarkan kategori yang dibuat oleh Badan Koordinasi Keluarga Berencana Naisonal (BKKBN).

Tabel 5 Discretization Data

Variabel	Kategori	Keterangan
Umur (X1)	1. Produktif 2. Tidak produktif	15 – 64 tahun Dibawah 15 tahun dan 65 tahun ke atas
Jumlah Anggota Rumah Tangga (X2)	1. Rumah tangga kecil 2. Rumah tangga sedang 3. Rumah tangga besar	< 5 orang 5 – 7 orang > 7 orang

3.3.2 Penskalaan Ulang

Variabel tingkat pendidikan dan tingkat pendidikan suami terdiri dari 14 Kategori, Kondisi ini akan membuat pohon keputusan yang sangat lebar. Oleh karena itu, untuk mempermudah analasis, kategori variabel tingkat pendidikan dan tingkat pendidikan suami perlu diringkas atau dilakukan penskalaan ulang (*rescale*) berdasarkan kategori yang digunakan oleh Badan Pusat Statistik (BPS).

Tabel 6 Peringkasan Kategori Variabel Tingkat Pendidikan dan Tingkat Pendidikan Suami

Variabel	Kategori	Keteranangan
Tingkat Pendidikan (X3) dan Tingkat Pendidikan Suami (X4)	1. Tidak pernah sekolah	Mereka yang tidak/belum pernah bersekolah sama sekali.
	2. Sekolah dasar	Mereka yang memeiliki pendidikan tidak/belum tamat SD/Ibtidaiyah, Paket A, SMP/Tsanawiyah, SMP Kejuruan dan Paket B.
	3. Sekolah menengah	Mereka yang memiliki pendidikan SMA/Aliyah, SMK dan Paket C.
	4. Sekolah tinggi	Mereka yang memiliki izajah Diploma I/II, Diploma III, Diploma IV/Sarjana dan S2/s3

3.4 Pemodelan

3.4.1 Algoritma ID3

Pohon keputusan dengan algoritma ID3 diawali dengan penentuan parameter-parameter untuk mencegah agar pohon tidak terlalu lebar. Peneliti menentukan parameter *minimal size for split* : 2, *minimal leaf size* : 40 dan *minimal gain* : 1, pohon dibuat dengan bantuan *software Rapidminer*. Algoritma ID3 menghasilkan pohon dengan 50 kondisi dengan variabel umur sebagai akar pohon keputusan, ini berarti umur adalah variabel yang paling

berpengaruh terhadap partisipasi perempuan nikah dalam kegiatan ekonomi rumah tangga di DIY.

Pengukur akurasi pohon menggunakan *10-Ford Validation*. Data dibagi kedalam 10 partisi data set,disetiap iterasi salah satu partisi data menjadi data *testing*, kemudian *confusion matrix* dari setiap iterasi dijumlahkan, total *confusion matrix* tersebut terdapat pada tabel

		Prediksi(j)	
		Bekerja	Tidak Bekerja
Aktual(i)	Bekerja	415	159
	Tidak Bekerja	56	44

Berdasarkan rumus 9,tingkat akurasi yang diperoleh sebesar 68,10% dan kesalahan klasifikasinya sebesar 31,90%, ini berarti keakuratan pohon keputusan dengan algoritma ID3 dalam menjelaskan partisipasi perempuan nikah dalam kegiatan ekonomi rumah tangga di DIY sebesar 68,10%.

3.4.2 Algoritma C4.5

Pohon keputusan dengan algoritma C4.5 diawali dengan penentuan parameter-parameter untuk mencegah agar pohon tidak terlalu besar. Peneliti menetukan parameter *maximal dept* : -1, *minimal gain* : 0,01, *minimal leaf size* : 4, *minimal siza for split 2* dan *number of prepruning alternative*: 3, pohon dibuat dengan bantuan *software Rapidminer*. Algoritma C4.5 menghasilkan pohon dengan 27 kondisi dengan variabel status pekerjaan suami sebagai akar pohon keputusan, ini berarti status pekerjaan suami adalah variabel yang paling berpengaruh terhadap partisipasi perempuan nikah dalam kegiatan ekonomi.

Pengukur akurasi pohon menggunakan *10-Ford Validation*. Data dibagi kedalam 10 partisi data set,disetiap iterasi salah satu partisi data menjadi data *testing*, kemudian *confusion matrix* dari setiap iterasi dijumlahkan, total *confusion matrix* tersebut terdapat pada tabel

Tabel 4.15 *Confusion Matrix* Algoritma C4.5

		Prediksi(j)	
		Bekerja	Tidak Bekerja
Aktual(i)	Bekerja	439	179
	Tidak Bekerja	32	24

Berdasarkan rumus 9,tingkat akurasi yang diperoleh sebesar 68,71% dan kesalahan klasifikasinya sebesar 31,29%, ini berarti keakuratan pohon keputusan dengan algoritma C4.5 dalam menjelaskan partisipasi perempuan nikah dalam kegiatan ekonomi rumah tangga di DIY sebesar 68,71%.

3.5 Perbandingan Algoritma ID3 dan AlgoritmaC4.5

Dalam membandingkan kedua pohon keputusan, ukuran yang digunakan sebagai alat pembanding adalah tingkat akirasi. Tingkat akurasi pohon keputusan dengan algoritma ID3 sebesar 68,10% dan tingkat akurasi pohon keputusan dengan algoritma C4.5 sebesar 68,10%. Berdasarkan tingkat akurasi yang diperoleh algoritma C4.5 lebih baik dari pada algoritma ID3 dengan tingkat akurasi sebesar 68,71%.

4. Kesimpulan

Dari 708 perempuan yang sudah menikah, perempuan yang bekerja memiliki persentase sekitar 68% atau sebanyak 484 orang. Sedangkan yang tidak bekerja memiliki persentase sebesar 22,24% atau sebanyak 224 orang. Hal ini menunjukkan bahwa partisipasi ekonomi perempuan nikah di DIY cukup tinggi. mayoritas perempuan nikah yang bekerja berumur 33 – 52 tahun, mayoritas perempuan nikah di DIY terdiri dari perempuan nikah yang tidak memiliki ijazah pendidikan dan berpendidikan akhir sekolah dasar, mayoritas rumah tangga memiliki jumlah anggota 3 sampai 4 orang dan mayoritas rumah tangga memiliki memiliki suami yang berpofesi sebagai buruh/karyawan/pegawai.

Algoritma C4.5 lebih baik dari algoritma ID3 dalam menjelaskan partisipasi perempuan nikah dalam kegiatan ekonomi di DIYdengan tingkat akurasi 68,71%. Dengan menggunakan algoritma C4.5,variabel status pekerjaan suami adalah variabel yang paling berpengaruh dalam partisipasi perempuan nikah dalam kegiatan ekonomi dalam kegiatan ekonomi rumah tangga dan menghasilkan 27 kondisi.

Ucapan terimakasih

Dalam penyusunan tulisan ini, banyak pihak yang telah memberikan dukungan kepada penulis. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan terima kasih kepada seluruh dosen dan pimpinan Jurusan Statistika Institut Sains & Teknologi AKPRIND Yogyakarta

Daftar Pustaka

[1]Indriyani,N., 2009, Penerapan Metode Pohon Keputusan dengan Algoritma C4.5 pada Sistem Penunjang Keputusan dalamMemperkirakan Cuaca Jangka Pendek, *Skripsi*, Fakultas Ilmu Komputer, Universitas Indonesia, Depok.

[2]Maimon, O., dan Rokach, L., 2010, *Data Mining and Knowledge Discovery Handbook*. Ed.2, Springer, New York.

[3]Nowozin,N. 2012, *Improve Information Gain for Decision Tree Induction*, https://icml.cc/2012/papers/177.pdf

[4]Prasetyo, 2014. *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*, Andi Offset, Yogyakarta.

[5]Riyanto, 2015. Analisis Faktor-faktor yang Mempengaruhi Partisipasi Kerja Perrempuan dalam Kegatan Ekonomi Rumah Tangga di Kota Semarang, *Skripsi*, Akademi Statistika Muhammadiyah, Semarang.

[5]Srivasta, M. S., dan Hui T. K., 1987, *On Assessing Multivariate Normality Based On Shapiro-Wilk W Statistic*, N0. 1, Vol. 5, 15-18.

[6]Tan P, Steinbach M, Kumar V. 2006. *Introduction to Data Mining*, Pearson Education.Inc, Boston.

[7]Williams, G., 2011. *Data Mining with Rattle and R The Art of Excavating Data for Knowledge Discovery*. Australia : Springer.