UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 145-150.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Poor Households in West Sumatra Province using Decision Tree Algorithm C4.5

Dinda Fitriza, Atus Amadi Putra*, Dodi Vionanda dan Zilrahmi
Departemen Statistika, Universitas Negeri Padang, Kota Padang, Negara Indonesia
*Corresponding author: atusamadiputra@fmipa.unp.ac.id

Submitted : 26 Maret 2024
Revised : 30 Mei 2024
Accepted : 31 Mei 2024

ABSTRACT

The significant and increasingly complex issue of poverty poses a considerable challenge to Indonesia's development, including West Sumatra Province, with a poverty rate was 5.92% in 2022. The government has initiated programs to address poverty by focusing on the criteria of impoverished households. Data on impoverished households can be obtained through the National Socio-Economic Survey (Susenas). One method that can classify impoverished households is the decision tree. Decision tree is a flowchart that resembles a tree. The C4.5 algorithm used in this research has the ability handle discrete and continuous data, manage variables with missing values, and prune decision tree branches. The result of the analysis shows that the variables affecting the classification of poor households are the number of household members, then the age of the household head, type of house floor, type of house wall, source of drinking water, and cooking fuel. The accuracy of the test data using a confusion matrix is 69.89%, sensitivity of 71.15% for classifying regular households, and specificity of 68.72% for classifying impoverished households.

Keywords: *C4.5 Algorithm, Classification, Poverty, Decision Tree*

I. PENDAHULUAN

Masalah kemiskinan di Indonesia masih menjadi tantangan besar dan harus diperhatikan. Menurut Suryawati (2004), kemiskinan merupakan kondisi ketidakmampuan pendapatan dalam mencukupi kebutuhan pokok sehingga kurang mampu untuk menjamin kelangsungan hidup. Sehingga kemiskinan menjadi salah satu persoalan yang terus dihadapi di sejumlah daerah di Indonesia, termasuk Provinsi Sumatera Barat. Menurut Badan Pusat Statistik (BPS), Provinsi Sumatera Barat mengalami penurunan persentase penduduk yang tergolong miskin pada tahun 2022, yaitu dari 6,63% menjadi 5,92%. Namun, perbedaan persentase kemiskinan antar daerah masih tergolong tinggi. Kabupaten Kepulauan Mentawai menjadi kabupaten dengan persentase kemiskinan paling besar yaitu sebesar 13,97% dan Kota Sawahlunto dengan persentase terkecil, yaitu 2,28%.

Kemiskinan mengakibatkan banyak anak tidak dapat memperoleh pendidikan yang berkualitas, kesulitan dalam membiayai layanan kesehatan, serta kurangnya tabungan dan investasi. Selain itu, kemiskinan juga dapat menyebabkan kesulitan dalam akses ke layanan publik, kesulitan dalam mendapatkan pekerjaan yang stabil, kurangnya jaminan sosial untuk keluarga, serta meningkatnya migrasi ke kota sebagai upaya mencari kesempatan yang lebih baik. Hal ini juga dapat menghambat pertumbuhan ekonomi di Provinsi Sumatera Barat, karena menciptakan daerah-daerah tertinggal. Rumah tangga miskin (RTM) memiliki pendapatan yang lebih kecil daripada biaya yang dikeluarkan untuk memenuhi kebutuhan hidup di wilayah tempat tinggalnya. Identifikasi kriteria Rumah Tangga Miskin (RTM) dilakukan untuk meningkatkan efektivitas program penanggulangan kemiskinan di Sumatera Barat. Tujuannya adalah untuk mengurangi tingkat kemiskinan di wilayah tersebut serta melakukan pencegahan kemiskinan yang serupa di masa depan. Identifikasi kriteria RTM dapat dilakukan dengan menggunakan klasifikasi.

Klasifikasi merupakan proses penempatan objek (observasi) tertentu kedalam sekumpulan kategori berdasarkan sifat objek masing-masing (Gorunescu, 2011). Salah satu metode klasifikasi yang dapat mengidentifikasikan RTM adalah *decision tree*. Menurut Han, Kamber dan Pei (2012), d*ecision tree* adalah diagram alur yang memiliki struktur seperti pohon, dimana setiap simpul internal menunjukkan pengujian pada suatu variabel, setiap cabang mewakili hasil pengujian, dan setiap simpul daun memiliki label kelas atau distribusi kelas. Simpul teratas dalam pohon ini adalah simpul akar. Algoritma yang digunakan pada penelitian ini adalah algoritma C4.5.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Algoritma C4.5 adalah pengembangan lanjutan dari Algoritma ID3. Menurut Elisa (2007) Algoritma C4.5 memiliki kemampuan untuk mengolah data diskrit maupun kontinu dan dapat menangani variabel dengan nilai yang hilang, serta dapat memangkas (*pruning*) cabang pada pohon keputusan. Prinsip kerja algoritma C4.5 hampir sama algoritma ID3, namun untuk pemilihan atribut algoritma tersebut menggunakan *gain ratio*. Berdasarkan latar belakang tersebut, dilakukan penelitian untuk mengklasifikasikan rumah tangga miskin di Provinsi Sumatera Barat. Hasil penelitian diharapkan dapat memberikan dukungan dan bantuan kepada Pemerintah Provinsi Sumatera Barat dalam mengidentifikasi rumah tangga miskin dan tidak miskin, serta dapat membantu dalam merancang kebijakan pengurangan dan penanggulangan kemiskinan yang efektif dan terprogram.

II. METODE PENELITIAN

A. Sumber Data dan Variabel Penelitian

Penelitian ini menggunakan data sekunder yang merupakan hasil Survei Sosial Ekonomi Nasional (Susenas) Provinsi Sumatera Barat tahun 2022. Variabel respon pada penelitian ini adalah kelompok rumah tangga (Y) dengan 11 variabel prediktor yaitu umur kepala rumah tangga (X1), ijazah terakhir kepala rumah tangga (X2), status/kedudukan pekerjaan utama KRT (X3), jumlah ART (X4), luas lantai rumah (X5), jenis lantai (X6) dan jenis dinding (X7) tempat tinggal, kepemilikan fasilitas sanitasi (X8), sumber penerangan (X9), sumber air minum (X10), dan bahan bakar untuk memasak (X11).

B. Teknik Analisis Data

Teknik analisis data yang digunakan pada penelitian ini adalah algoritma C4.5 yang diimplementasikan menggunakan *RStudio*. Berikut adalah tahapan dalam membangun pohon keputusan menggunakan algoritma C4.5.

1. Melakukan analisis statistika deskriptif
2. Melakukan *cleaning data*
3. Membagi data menjadi data latih dan data uji
4. Menghitung nilai *entropy*

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

Dimana:
S : himpunan kasus
m : partisi data S
p_i : proporsi S ke-i terhadap total S

5. Menghitung nilai *information gain*.

$$Gain\ (S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy\ (S_i) \tag{2}$$

Dimana:
A : atribut/variabel
n : partisi data A
$|S_i|$: jumlah atribut/variabel A pada partisi ke-i
S : jumlah kasus S

6. Menghitung nilai *split info.*

$$SplitInfo\ (S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} \times \log_2 \frac{S_i}{S} \tag{3}$$

7. Menghitung nilai *gain ratio.*

$$Gain\ Ratio\ (S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \tag{4}$$

8. Mengulang tahapan 4 hingga 7 untuk setiap cabang sampai semua cabang memiliki keputusan.
9. Evaluasi hasil klasifikasi menggunakan *confusion matrix*.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Tabel 1. *Confusion Matrix*

Classification		***Predicted Class***	
		Class = Yes	*Class = No*
Aktual Class	*Class = Yes*	*True Positive* (TP)	*False Positive* (FP)
	Class = No	*False Negative* (FN)	*True Negative* (TN)

(Sumber: Gorunescu, F. 2011)

Evaluasi menggunakan *confusion matrix* akan menghasilkan nilai *accuracy, sensitivity,dan specificity* (Rokarch dan Maimon, 2015).

$$Accuracy\ (\%) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\% \quad (5)$$

$$Sensitifity\ (\%) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \quad (6)$$

$$Specificity\ (\%) = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{FN}} \times 100\% \quad (7)$$

III. HASIL DAN PEMBAHASAN

A. Statistik Deskriptif

Tabel 2 menyajikan gambaran karakteristik rumah tangga di Provinsi Sumatera Barat.

Tabel 2. Statistik Deskriptif

Variabel	Kategori	Jumlah
Kelompok rumah tangga (Y)	Rumah Tangga Miskin (RTM)	489
	Rumah Tangga Biasa (RTB)	11110
Umur kepala rumah tangga (X1)	1. < 26 tahun	130
	2. 26 – 46 tahun	4289
	3. ≥ 46 tahun	7180
Ijazah terakhir KRT (X2)	0. Tidak punya ijazah SD	2065
	1. SD/sederajat	2721
	2. SMP/sederajat	1800
	3. SMA/sederajat	3742
	4. D1sd Universitas	1166
	5. Tidak pernah sekolah	105
Status/kedudukan dalam pekerjaan utama KRT (X3)	0. Tidak bekerja	1296
	1. Berusaha sendiri	3226
	2. Berusaha dibantu buruh tidak tetap/tidak dibayar	1999
	3. Bekerja dibantu buruh tetap/buruh dibayar	633
	4. Buruh/karyawan/pegawai	3000
	5. Pekerja bebas	1281
	6. pekerja keluarga atau tidak dibayar	164
Jumlah ART (X4)	1. ≤ 4	10037
	2. > 4	1562
Luas lantai rumah (X5)	1. ≤ 10 m^2	33
	2. 10 – 30 m^2	980
	3. > 30 m^2	10586
Jenis lantai (X6)	1. Marmer/granit	218
	2. Keramik	3839
	3. Parket/vinil/permadani, Ubin/tegel/teraso	108
	4. Kayu/papan	1088
	5. Semen/bata merah	6289
	6. Bambu	28
	7. Tanah/Lainnya	29

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Jenis dinding (X7)	1. Tembok	8807
	2. Plesteran anyaman bambu/kawat	212
	3. Kayu, batang kayu	2537
	4. Bambu, anyaman bambu	43
Kepemilikan sanitasi (X8)	1. Sendiri	9467
	2. Bersama	822
	3. Komunal/umum	245
	4. Tidak digunakan	19
	5. Tidak ada	1046
Sumber air minum (X9)	1. Air kemasan/isi ulang	4838
	2. Leding	1552
	3. Sumur bor/ pompa	770
	4. Sumur terlindung	1692
	5. Sumur tak terlindung	1672
	6. Mata air	639
	7. Lainnya	436
Sumber penerangan (X10)	1. Listrik PLN	11319
	2. Listrik non PLN	155
	3. Bukan listrik	125
Bahan bakar memasak (X11)	0. Tidak memasak di rumah	77
	1. Listrik	29
	2. Elpigi/gas kota/ biogas	9671

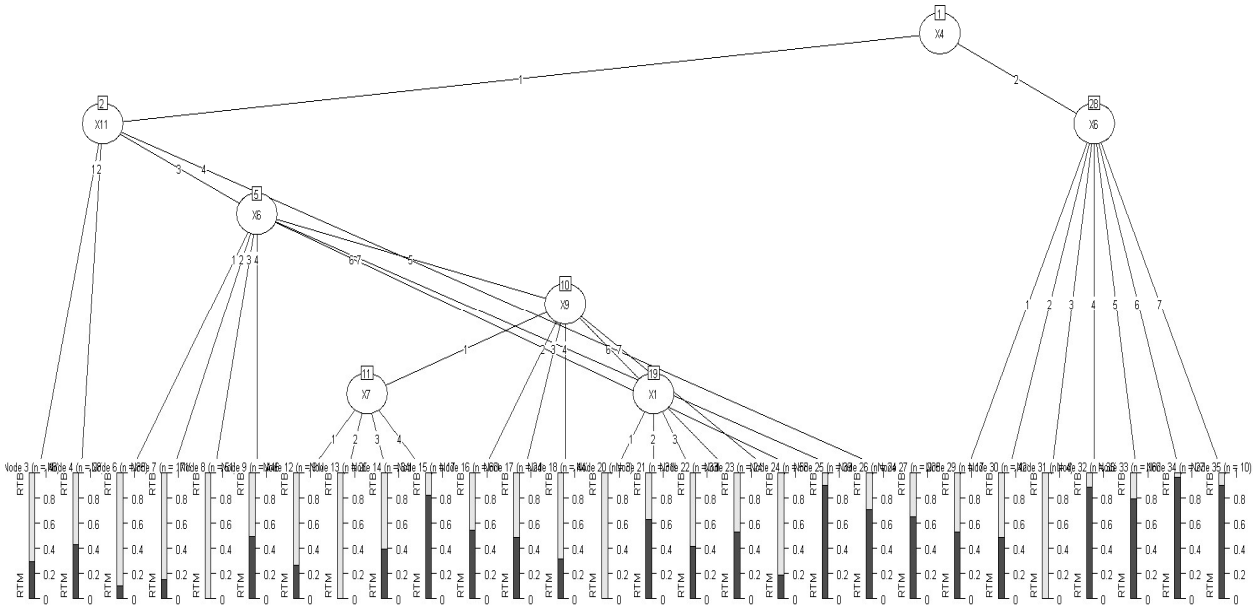
Berdasarkan Tabel 2, dapat dilihat bahwa terdapat 489 rumah tangga miskin dan 11110 rumah tangga biasa dengan mayoritas kepala rumah tangga berumur lebih dari 46 tahun, ijazah tertinggi yang dimiliki adalah SMA/sederajat dan status/kedudukan dalam pekerjaan utama yaitu berusaha sendiri. Mayoritas tempat tinggal memiliki lantai dengan luas lebih dari 30 m^2, sebanyak 6289 rumah memiliki lantai yang terbuat dari semen/bata merah, sebanyak 8807 rumah dengan jenis dinding terluas adalah tembok. Kebanyakan rumah tangga sudah memiliki fasilitas sanitasi sendiri, air minum bersumber dari air kemasan/air isi ulang, sumber penerangan berasal dari listrik PLN dan bahan bakar untuk memasak yaitu elpigi/gas kota/biogas.

B. Algoritma C4.5

Algoritma C4.5 bekerja dengan cara menghitung dan mengklasifikasikan 11599 data rumah tangga di Provinsi Sumatera Barat tahun 2022. Analisis pada penelitian ini menggunakan bantuan *software Rstudio*. Dataset dimasukkan ke dalam *Rstudio* kemudian dengan menggunakan data latih akan dibentuk model untuk algoritma C4.5. Pembentukkan model akan dimulai dengan mencari nilai *entropy*, kemuadian nilai *information gain*, *split info* dan terakhir nilai *gain ratio* yang sesuai dengan persamaan 1 sampai dengan 4, kemudian nilai yang memiliki nilai *gain ratio* paling tinggi akan dipilih simpul akar dari pohon keputusan. Proses perhitungan akan terus berlangsung hingga semua simpul memiliki keputusan. Setelah memperoleh model untuk algoritma C4.5, langkah selanjutnya adalah melakukan pengujian keputusan terhadap data uji sehingga nanti akan didapatkan hasil keputusan dan prediksi. Kemudian hasil keputusan tersebut dievaluasi menggunakan *confusion matrix* seperti yang telah dijelaskan dalam Tabel 1. Gambar 1 menampilkan hasil pohon keputusan yang dihasilkan menggunakan algoritma C4.5.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 1. Hasil pohon keputusan

Berdasarkan Gambar 1, dapat diketahui bahwa variabel X4 atau jumlah anggota rumah tangga yang menjadi simpul akar pada pohon keputusan. Hal ini menunjukkan bahwa jumlah anggota rumah tangga merupakan variabel yang paling berpengaruh dalam mengklasifikasikan rumah tanga miskin. Kemudian variabel umur KRT (X1), jenis lantai (X6), jenis dinding (X7), sumber air minum (X9) dan bahan bakar memasak (X11) menunjukkan simpul internal dan simpul daun sebagai pengelompokkan rumah tangga. Untuk mengetahui label kelas rumah tangga miskin atau tidak dilihat dari simpul daun, jika nilai simpul daun yang dimiliki mendekati 1 maka kategorinya adalah rumah tangga miskin, dan jika nilai yang dimiliki mendekati 0, maka kategorinya adalah rumah tangga biasa. Dari gambar tersebut diperoleh beberapa *rule* pengklasifikasian rumah tangga miskin. Salah satu contoh *rule* sebagai berikut, jika jumlah anggota rumah tangga lebih dari 4 orang dan jenis lantai rumah yang digunakan adalah kategori 4, 5, 6 dan 7, yaitu kayu/papan, semen/bata merah, bambu atau tanah/lainnya, maka rumah tangga dapat diklasifikasikan sebagai rumah tangga miskin, namun jika jenis lantai yang digunakan adalah kategori 1, 2, 3, yaitu marmer/granit, keramik dan parket/vinil/permadani, maka rumah tangga dapat diklasifikasikan sebagai rumah tangga biasa. Tahap selanjutnya adalah melakukan evaluasi ketepatan hasil klasifikasi menggunakan *confusion matrix*. Tabel 3 menyajikan hasil *confusion matrix*.

Tabel 3. Hasil Perhitungan *Confusion Matrix*

Classification		*Predicted Class*	
		RTB	RTM
Aktual Class	RTB	757	415
	RTM	224	924

Berdasarkan Tabel 3, dapat dilihat bahwa jumlah RTB yang juga diprediksi RTB sebanyak 757 rumah tangga sedangkan jumlah RTM yang juga diprediksi RTM sebanyak 924 rumah tangga. Berikut adalah hasil ketepatan model dengan menggunakan persamaan (5), (6), dan (7).

$$Accuracy\ (\%) = \frac{397 + 413}{397 + 413 + 188 + 161} \times 100\% = 69{,}89\%$$

$$Sensitify\ (\%) = \frac{397}{397 + 161} \times 100\% = 71{,}15\%$$

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$Specificity\ (\%) = \frac{413}{161 + 413} \times 100\% = 68{,}72\%$$

Dari perhitungan hasil ketepatan model, didapatkan nilai akurasi data prediksi sebesar 69,89% yang artinya pohon keputusan yang dilatih dapat mengklasifikasikan data baru sebesar 69,89%. Hasil perhitungan juga menunjukkan nilai sensitivitas sebesar 71,15% untuk mengukur ketepatan klasifikasi pada RTB, serta nilai spesifisitas sebesar 68,72% untuk mengukur ketepatan klasifikasi pada RTM.

IV. KESIMPULAN

Berdasarkan hasil dan pembahasan dalam mengklasifikasikan rumah tangga miskin (RTM) di Provinsi Sumatera Barat menggunakan algoritma C4.5 didapatkan variabel yang menjadi kriteria utama penentu masalah RTM adalah jumlah anggota rumah tangga dan kriteria lainnya adalah umur kepala rumah tangga, jenis lantai rumah, jenis dinding rumah, sumber air minum dan bahan bakar memasak. Ketepatan hasil klasifikasi dengan menggunakan *confusion matrix* menghasilkan nilai akurasi 69,89%, sensitivitas 71,15% untuk mengklasifikasikan rumah tangga biasa, serta spesifisitas 68,72% untuk mengklasifikasikan rumah tangga miskin.

DAFTAR PUSTAKA

Badan Pusat Statistik (BPS). 2023. *Statistik Indonesia 2023*. Jakarta: BPS

Badan Pusat Statistik (BPS). 2023. Kemiskinan dan Ketimpangan. https://sumbar.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html. Diakses pada tanggal 4 Mei 2023, pukul 20.26.

Badan Pusat Statistik (BPS). 2023. Survei Sosial Ekonomi Nasional. https://sumbar.bps.go.id/publication/2022/12/28/a5fd46b5a6bd4f1da740f54b/statistik-kesejahteraan-rakyat-provinsi-sumatera-barat-2022.html. Diakses pada tanggal 18 Mei 2023, pukul 09.15.

Badan Perencanaan Pembangunan Nasional (Bappenas). 2022. Daerah Tertinggal. https://simreg.bappenas.go.id/home/daerahtertinggal. Diakses pada tanggal 10 Juni 2023, pukul 15.38.

Elisa, E. 2007. Analisa dan Penerapan Algoritma C4.5 dalam Data Mining untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT. Arupadhatu Adisesantani. *Jurnal Sistem Informasi*, 2(1).

Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Berlin: Springer

Han, J., Kamber, M., dan Pei, J. 2012. *Data mining Concepts and Techniques Edisi ke-3*. Amerika Serikat: Elsevier.

Rokach, L., dan Maimon, O. 2015. Data Mining with Decision Trees Theory and Application 2nd Ed. Singapore: World Scientific.

Suryawati. 2004. *Teori Ekonomi Mikro*. Yogyakarta: UPP. AMP YKPN

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/157