 http://dx.doi.org/10.29313/.v0i0.28406

Analisis Faktor-Faktor yang Mempengaruhi Jumlah Anak Lahir Hidup dengan Metode *Classification and Regression Trees*

Sintia Dewi Laksa Hartati*, Yayat Karyana

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*sintiadewi.lh@gmail.com, yayatkaryana@gmail.com

Abstract. Children born according to BPS are all children who at birth show signs of life, even for a moment such as heartbeat, crying, breathing and other signs of life. In the 2012 Indonesian Demographic and Health Survey (IDHS) report, the expected total fertility rate is 2 children per woman, which is not in line with the actual fertility rate (ie 2.6 children per woman). Therefore, this study will discuss the factors that influence the number of children born alive in the 2012 IDHS data. The method used is Classification and Regression Trees (CART) with the response variable of the number of children born alive in the form of binary data. As for the predictor variables, namely education level, wealth level, working status, age at first marriage and location of residence. The CART method was developed by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Stone in 1984 which is a tree-structured classification method. In this study, all independent variables affect the number of children born alive. Respondents with a controlled number of children born alive were respondents who had secondary and higher education levels, very rich wealth levels, did not have formal jobs, with age at first marriage >35 years and resided in urban areas. Respondents with an uncontrolled number of children born alive were respondents with a low level of education, very poor wealth level, having a formal job, having a first marriage age ≤35 years, and residing in a rural area. In this study, the classification of the number of live born children resulted in the classification accuracy of the testing data of 65.19%.

Keywords: number of children born alive, IDHS, Classification and Regression Trees.

Abstrak. Anak lahir hidup menurut BPS yaitu semua anak yang waktu lahir menunjukkan tanda-tanda kehidupan, walau sesaat, seperti adanya detak jantung, menangis, bernafas serta tanda-tanda kehidupan lainnya. Dalam laporan Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2012, disebutkan angka fertilitas total yang diharapkan sebesar 2 anak per wanita, yang berarti tidak sesuai dengan angka fertilitas sebenarnya (yakni 2,6 anak per wanita). Oleh sebab itu, penelitian ini akan membahas mengenai faktor-faktor yang mempengaruhi jumlah anak lahir hidup pada data SDKI tahun 2012. Metode yang akan digunakan adalah *Classification and Regression Trees* (CART) dengan variabel respon jumlah anak lahir hidup berbentuk data biner. Sedangkan untuk variabel prediktor yaitu tingkat pendidikan, tingkat kekayaan, status bekerja, umur kawin pertama dan lokasi tempat tinggal. Metode CART mulai dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Stone pada tahun 1984 yang merupakan metode klasifikasi berstruktur pohon. Dalam penelitian ini, semua variabel bebas berpengaruh terhadap jumlah anak lahir hidup. Responden dengan jumlah anak lahir hidup terkontrol terdapat pada responden yang memiliki tingkat pendidikan menengah dan tinggi, tingkat kekayaan sangat kaya, tidak mempunyai pekerjaan formal, dengan umur kawin pertama >35 tahun dan bertempat tinggal di daerah perkotaan. Untuk responden dengan jumlah anak lahir hidup tidak terkontrol terdapat pada responden dengan tingkat pendidikan rendah, tingkat kekayaan

sangat miskin, memiliki pekerjaan formal, dengan umur kawin pertama ≤35 tahun, serta bertempat tinggal di daerah perdesaan. Dalam penelitian ini pula, pengklasifikasian jumlah anak lahir hidup menghasilkan ketepatan klasifikasi pada data *testing* sebesar 65,19%.

Kata Kunci: jumlah anak lahir hidup, SDKI, *Classification and Regression Trees*.

1. Pendahuluan

Indonesia menempati negara padat penduduk keempat di dunia setelah China, India dan Amerika Serikat. Penduduk Indonesia tahun 2020 mencapai lebih dari 267 juta jiwa. Yang diproyeksi pada tahun 2025 akan mencapai 284 juta jiwa. Penduduk yang banyak merupakan beban bagi suatu negara, karena banyak atau sedikitnya penduduk, harus di imbangi dengan kualitas penduduk di suatu negara tersebut. Maka dari itu, penting bagi pemerintah untuk teurs berusaha menekan laju pertumbuhan penduduk, salah satu upayanya adalah dengan membatasi jumlah kelahiran (Handayani&Najib, 2019).

Dalam laporan Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2012, disebutkan angka fertilitas total yang diharapkan sebesar 2 anak per wanita, yang berarti tidak sesuai dengan angka fertilitas sebenarnya (yakni 2,6 anak per wanita). Pemerintah menggagas program Keluarga Berencana (KB) 'Dua Anak Cukup', yang khususnya dijalankan oleh Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN).

Untuk mengetahui informasi secara rinci tentang penduduk, keluarga berencana dan kesehatan, maka diadakan Survei Demografi dan Kesehatan Indonesia (SDKI). Survei ini dilaksanakan di laksanakan di seluruh Indonesia. Salah satu yang dapat diketahui dari SDKI adalah jumlah anak lahir hidup. dimana menurut BPS anak lahir hidup yaitu semua anak yang waktu lahir menunjukkan tanda-tanda kehidupan, walau sesaat, seperti adanya detak jantung, menangis, bernafas serta tanda-tanda kehidupan lainnya. Seperti yang dijelaskan sebelumnya, angka fertilitas di Indonesia masih lebih tinggi daripada yang diharapkan. Maka terdapat faktor-faktor yang mempengaruhi ketidaksesuaian angka fertilitas yang diharapkan dengan nilai yang sesungguhnya.

Pada penelitian ini, akan dibahas mengenai model yang mempengaruhi jumlah anak lahir hidup berdasarkan data SDKI 2012, dengan mengklasifikasikan jumlah anak lahir hidup yang kurang dari sama dengan dua anak dan lebih dari dua anak. Terdapat berbagai metode yang dapat digunakan untuk variabel respon yang berbentuk kategorik, salah satunya adalah regresi logistik biner atau *Binary Logistic Regression*, Analisis Diskriminan Linear (ADL), dan *Classification and Regression Trees* (CART). Metode CART mulai dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Stone pada tahun 1984 yang merupakan metode klasifikasi berstruktur pohon. Tujuan utama dari metode CART yakni untuk mendapatkan keakuratan dari suatu kelompok data sebagai penciri dari pengklasifikasian.

Dari latar belakang diatas, maka identifikasi masalah yang dapat diambil adalah "bagaimana cara menerapkan dan menentukan klasifikasi faktor yang mempengaruhi jumlah anak lahir hidup dari wanita usia subur (WUS) kelompok umur 15-49 tahun di Indonesia yang dibentuk berdasarkan metode *Classification and Regression Trees (*CART)?". Selanjutnya, berdasarkan identifikasi masalah diatas maka tujuan yang ingin dicapai yaitu untuk mengetahui cara menerapkan dan menentukan klasifikasi faktor yang mempengaruhi jumlah anak lahir hidup dari wanita usia subur (WUS) kelompok umur 15-49 tahun di Indonesia yang dibentuk berdasarkan metode *Classification and Regression Trees (*CART).

2. Metodologi

Metode Pengumpulan Data

Pada penelitian ini, data yang digunakan adalah data sekunder yaitu data SDKI tahun 2012 dan telah digunakan oleh Karyana dkk (2020) untuk menganalisis keinginan menambah anak dengan 31.732 responden yang memenuhi syarat, dan merupakan WUS kelompok umur 15-49

tahun di Indonesia.

Tabel 1. Variabel Penelitian

Variabel	Kategori	Skala
Jumlah Anak Lahir Hidup (Y)	0 = Terkontrol ($\leq$ 2 anak) 1 = Tidak Terkontrol (> 2 anak)	Nominal
Tingkat Pendidikan (X_1)	0 = Rendah 1 = Menengah 2 = Tinggi	Ordinal
Tingkat Kekayaan (X_2)	0 = Sangat Miskin 1 = Miskin 2 = Menengah 3 = Kaya 4 = Sangat Kaya	Ordinal
Status Bekerja (X_3)	0 = Tidak Bekerja 1 = Bekerja	Nominal
Umur Kawin Pertama (X_4)	0 = kurang dari atau sama dengan 35 tahun ($\leq$ 35 tahun) 1 = lebih dari 35 tahun ($>$ 35 tahun)	Nominal
Lokasi tempat tinggal (X_5)	0 = perdesaan 1 = perkotaan	Nominal

Metode Analisis Data

Metode Classification and Regression Trees (CART)

Metode CART dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Stone pada tahun 1984. Metode CART mengasumsikan bahwa pohon keputusan adalah pohon biner (Zhang *et al*., 2018). Secara umum, penerapan metode CART ini terdiri dari 3 tahap, yaitu: pembentukkan pohon klasifikasi, pemangkasan pohon klasifikasi, serta penentuan pohon klasifikasi optimal.

1. Pembentukan Pohon Klasifikasi. Data akan dipilah menggunakan indeks gini. Indeks gini memiliki persamaan sebagai berikut:
   $i(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_j p^2(j|t)$ ... (1)
   Keterangan:
   $i(t)$ = Indeks gini
   $p(j|t)$ = proporsi kelas j pada simpul t, dengan j=1,2,…,n
   $p(i|t)$ = proporsi kelas i pada simpul t
   Tahapan berikutnya adalah menentukan kriteria *goodness of split* ($\phi(s,t)$) yang bertujuan untuk mengevaluasi pemilah-pemilah s yang terdapat pada simpul t. berikut merupakan *goodness of split* yang merupakan penurunan heterogenitas:
   $\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ … (2)
   Keterangan:
   $\phi(s,t)$ = goodness of split
   p_L = proporsi pengamatan yang menuju simpul kiri
   p_R = proporsi pengamatan yang menuju simpul kanan
   $i(t_L)$ = fungsi heterogenitas simpul anak kiri
   $i(t_R)$ = fungsi heterogenitas simpul anak kanan
   Suatu simpul t akan menjadi simpul terminal apabila hanya terdapat satu pengamatan (n=1) di dalam setiap simpul anak atau terdapat batasan minimum n pengamatan yang diinginkan oleh peneliti. Dilanjutkan dengan penandaaan label kelas yang terdapat pada simpul terminal akan ditentukan berdasarkan aturan jumlah terbanyak.
2. Pemangkasan Pohon Klasifikasi. Menurut Rokach (dikutip dalam Indah Prabawati *et al*., 2019), Pemangkasan pohon klasifikasi penting dilakukan untuk mencegah *overfitting*.
3. Penentuan Pohon Klasifikasi Optimal. Penduga *cross validation V-fold* pada penduga validasi silang lipat V yang akan digunakan.

Kegunaan dari ukuran ketepatan klasifikasi adalah untuk mengetahui apakah data yang telah diklasifikasikan sudah benar atau tidak. Berikut merupakan tabel yang digunakan untuk menghitung ketepatan klasifikasi:

Tabel 2. Ukuran Ketepatan Klasifikasi

OBSERVASI	**PREDIKSI**		**TOTAL**
	0	1	
0	n_{00}	n_{01}	N_0
1	n_{10}	n_{11}	N_1
TOTAL	N_0	N_1	N

Dengan demikian, jika menggunakan tabel diatas nilai *sensitivity*, *specificity*, dan akurasi dapat dihitung dengan cara sebagai berikut:

$$Sensitivity = \frac{n_{00}}{N_0} \quad \dots (3)$$

$$Specificity = \frac{n_{11}}{N_1} \quad \dots (4)$$

$$Akurasi = \frac{n_{00}+n_{11}}{N} \quad \dots (5)$$

Keterangan:
n_{00} = Frekuensi yang berasal dari kelas 0 yang tepat prediksi sebagai kelas 0
n_{01} = Frekuensi yang berasal dari kelas 0 yang tepat prediksi sebagai kelas 1
n_{10} = Frekuensi yang berasal dari kelas 1 yang tepat prediksi sebagai kelas 0
n_{11} = Frekuensi yang berasal dari kelas 1 yang tepat prediksi sebagai kelas 1
N_{00} = Jumlah observasi kelas 0
N_{11} = Jumlah observasi kelas 1

Tahapan Analisis Data

1. Mempersiapkan data yang akan digunakan untuk penelitian.
2. Membagi data yang akan digunakan menjadi dua bagian, yaitu data *training* dan data *testing*. Proporsi dari pembagian data tersebut ditentukan oleh peneliti, karena tidak terdapat aturan dalam pembagian data. Pada tahapan ini di tentukan data *training* sebesar 90% dan data *testing* sebesar 10%.
3. Membentuk pohon klasifikasi dengan metode CART, dengan tahapan sebagai berikut:
   a. Melakukan pemilihan pemilah pada masing-masing variabel menggunakan aturan pemilahan indeks gini.
   b. Melakukan evaluasi menggunakan kriteria *goodness of split.*
   c. Menentukan simpul terminal dengan cara menghentikan pembentukan pohon apabila mencapai batasan minimum dari pengamatan dalam simpul terminal.
   d. Melakukan penandaan label kelas pada simpul terminal berdasarkan pada aturan jumlah terbanyak.
   e. Menentukan pohon klasifikasi maksimal
4. Melakukan pemangkasan pada pohon klasifikasi. Pemangkasan pohon ini, menggunakan minimum *complexity parameter.*
5. Menentukan pohon klasifikasi optimal menggunakan metode *Cross Validation V-fold.*
6. Menghitung ketepatan pohon klasifikasi menggunakan *sensitiviity*, *specificity* serta akurasi.
7. Menginterpretasikan jumlah anak lahir hidup yang kurang dari sama dengan dua, dan yang lebih dari dua berdasarkan pohon klasifikasi yang sudah terbentuk.

3. Pembahasan dan Diskusi

Data *Training* dan Data *Testing*

Tahapan yang akan dilakukan pada bagian ini adalah memisahkan data menjadi dua bagian yaitu data *training* dan data *testing*, sebesar masing-masing 90% dan 10%. Pembagian data menggunakan software R Studio. Diketahui data yang digunakan dalam penelitian ini sebanyak

31.732, maka:

$$\boldsymbol{Jumlah\ Training = 31.732 \times 90\% = 28.558}$$
$$\boldsymbol{Jumlah\ Testing = 31.732 \times 10\% = 3.174}$$

Pemilihan Pemilah

Pada tahapan ini akan dilakukan pemilihan pemilah dengan menggunakan indeks gini yang tedapat pada persamaan 2.3, sebelum menghitung indeks gini, berikut merupakan calon pemilah yang dapat dilihat dalam Tabel 3.

Tabel 3. Calon Simpul Kiri dan Simpul Kanan

No	Calon Simpul Kanan	Calon Simpul Kiri
1	Tingkat Pendidikan = Rendah	Tingkat Pendidikan = Lainnya
2	Tingkat Pendidikan = Menengah	Tingkat Pendidikan = Lainnya
3	Tingkat Pendidikan = Tinggi	Tingkat Pendidikan = Lainnya
4	Tingkat Kekayaan = Sangat Miskin	Tingkat Kekayaan = Lainnya
5	Tingkat Kekayaan = Miskin	Tingkat Kekayaan = Lainnya
6	Tingkat Kekayaan = Menengah	Tingkat Kekayaan = Lainnya
7	Tingkat Kekayaan = Kaya	Tingkat Kekayaan = Lainnya
8	Tingkat Kekayaan = Sangat Kaya	Tingkat Kekayaan = Lainnya
9	Status Bekerja = Bekerja	Status Bekerja = Tidak Bekerja
10	Umur Kawin Pertama = ≤35 Tahun	Umur Kawin Pertama = >35 Tahun
11	Lokasi Tempat Tinggal = Perdesaan	Lokasi Tempat Tinggal = Perkotaan

Tabel 4. Perhitungan Probabilitas Setiap Calon Simpul

Calon Simpul	p_R	p_L	Kelas	$p(j\|t_R)$	$p(j\|t_L)$
1	0,398662	0,601338	≤ 2 Anak	0,489065	0,696617
			>2 Anak	0,510935	0,303383
2	0,485748	0,51425	≤ 2 Anak	0,681373	0,550116
			>2 Anak	0,318627	0,449884
⋮	⋮	⋮	⋮	⋮	⋮
11	0,53134	0,46866	≤ 2 Anak	0,585607	0,645921
			>2 Anak	0,414393	0,354079

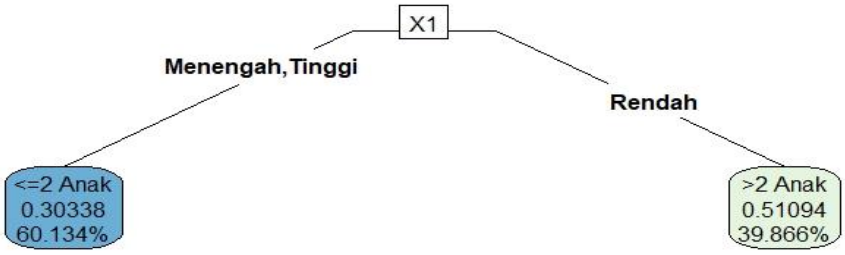
Di hitung pula nilai indeks gini dan *goodness of split* untuk setiap calon simpul, sehingga diperoleh:

Tabel 5. Indeks Gini dan Goodness of Split

Calon Simpul	Indeks Gini	$\phi(s,t)$
1	0.45341	0.02065
2	0.46546	0.00861
3	0.46843	0.00563
4	0.47045	0.00361
5	0.47407	0,00000046
6	0.4739	0.00017
7	0.47309	0.00098
8	0.47359	0.00048
9	0.47212	0.00194
10	0.47299	0.00107
11	0.47225	0.00181

Berdasarkan hasil perhitungan *goodness of split* di atas diperoleh bahwa nilai *goodness of split* tertinggi terdapat pada calon simpul pertama sebesar 0,02605.

Penentuan Simpul Terminal



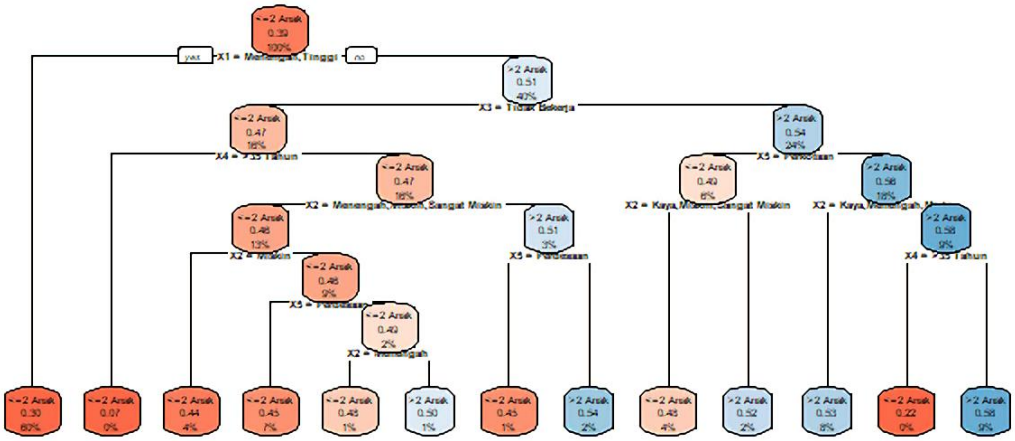
Gambar 1. Pemecahan Simpul Terminal

Penandaan Label Kelas

Penandaan label kelas ini terdapat pada simpul terminal yang ditentukan berdasarkan pada aturan jumlah terbanyak, yang dimaksudkan untuk mengetahui bawa simpul tersebut, apakah diberi label kelas "≤ 2 Anak" atau kelas ">2 Anak".

$$p(j_0|t_R) = max_j p(j|t_R) = 0,510935$$
$$p(j_0|t_L) = max_j p(j|t_L) = 0,696617$$

Pohon Klasifikasi Maksimal



Gambar 2. Struktur Pohon Klasifikasi Maksimal

Pemangkasan Pohon Klasifikasi

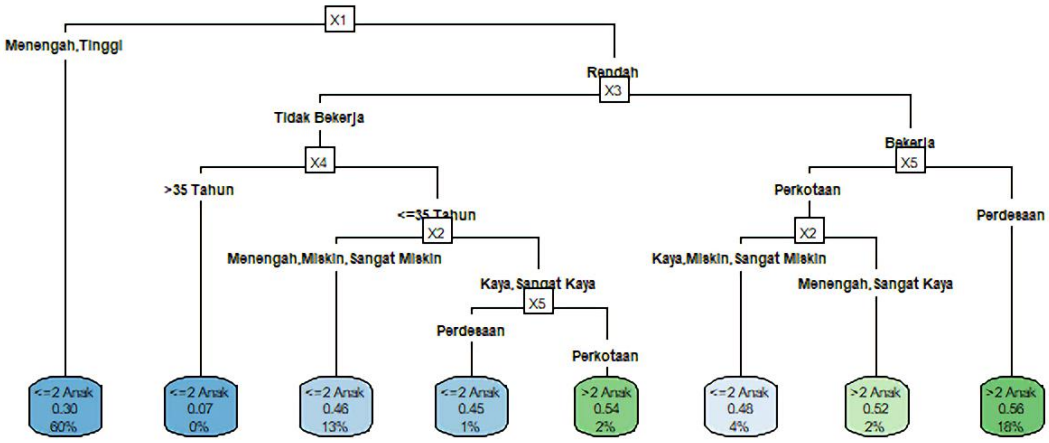
Proses pemangkasan ini dilakukan dengan melihat nilai *complexity parameter minimum*. Nilai *complexity parameter* di dapatkan berdasarkan *output* dari R studio, yang di cari dari nilai *Cross Validation V-fold*, di sajikan dalam Tabel 6 berikut ini:

Tabel 6. Nilai Complexity Parameter

No.	*Complexity Parameter*
1	0,001027780
2	0,001904416
3	0,025210846

Berdasarkan pada Tabel 6. di atas sudah di dapatkan nilai *complexity parameter*. Sehingga nilai *complexity parameter* yang terkecil atau nilai *complexity parameter minimum* sebesar 0,001027780.

Penentuan Pohon Klasifikasi Optimal



Gambar 3. Struktur Pohon Klasifikasi Optimal

Ketepatan Klasifikasi

Pohon klasifikasi optimal yang telah terbentuk, perlu dilakukan evaluasi dari hasil klasifikasinya. Dengan menggunakan *software* R-Studio, ketepatan klasifikasi untuk data *testing*, didapatkan nilai *sensitivity* yang digunakan untuk mengukur ketepatan klasifikasi pada jumlah anak lahir hidup yang terkontrol sebesar 0,8423 atau 84,23% dan nilai *specificity* yang digunakan untuk mengukur ketepatan klasifikasi pada jumlah anak lahir hidup yang tidak terkontrol sebesar 0,3339 atau 33,39%. Sedangkan nilai akurasi yang digunakan untuk mengukur tingkat ketepatan klasifikasi untuk data *testing* sebesar 0,6519 atau 65,19%, yang artinya bahwa pohon optimal yang terbentuk mampu mengklasifikasikan data baru sebesar 65,19%.

Interpretasi Pohon Klasifikasi

Tabel 7. Interpretasi Pohon Klasifikasi

No	X_1	X_2	X_3	X_4	X_5	Y
1	Menengah, tinggi	-	-	-	-	Terkontrol
2	Rendah	-	Tidak bekerja	>35 tahun	-	Terkontrol
3	Rendah	Sangat miskin, miskin dan menengah	Tidak bekerja	≤35 tahun	-	Terkontrol
4	Rendah	Kaya dan sangat kaya	Tidak bekerja	≤35 tahun	Perdesaan	Terkontrol
5	Rendah	Kaya dan sangat kaya	Tidak bekerja	≤35 tahun	Perkotaan	Tidak Terkontrol
6	Rendah	Sangat miskin, miskin dan kaya	Bekerja	-	Perkotaan	Terkontrol
7	Rendah	Menengah dan sangat kaya	Bekerja	-	Perkotaan	Tidak terkontrol

8	Rendah	-	Bekerja	-	Perdesaan	Tidak terkontrol

4. Kesimpulan

Berdasarkan hasil dan pembahasan dapat diambil kesimpulan bahwa metode CART dapat diterapkan untuk mengklasifikasikan faktor yang mempengaruhi jumlah anak lahir hidup dari Wanita Usia Subur (WUS) kelompok umur 15-49 tahun di Indonesia. Dalam penelitian ini, semua variabel bebas berpengaruh terhadap jumlah anak lahir hidup. Responden dengan jumlah anak lahir hidup terkontrol terdapat pada responden yang memiliki tingkat pendidikan menengah dan tinggi, tingkat kekayaan sangat kaya, tidak mempunyai pekerjaan formal, dengan umur kawin pertama >35 tahun dan bertempat tinggal di daerah perkotaan. Untuk responden dengan jumlah anak lahir hidup tidak terkontrol terdapat pada responden dengan tingkat pendidikan rendah, tingkat kekayaan sangat miskin, memiliki pekerjaan formal, dengan umur kawin pertama $\leq$35 tahun, serta bertempat tinggal di daerah perdesaan. Dalam penelitian ini pula, pengklasifikasian jumlah anak lahir hidup menghasilkan ketepatan klasifikasi pada data *testing* sebesar 65,19%.

5. Acknowledge

Terima kasih kepada pihak-pihak yang membantu, mendukung serta memberi masukkan sehingga penelitian ini dapat terlaksana.

Daftar Pustaka

[1] Badan Pusat Statistik, Badan Kependudukan dan Keluarga Berencana Nasional, Departemen Kesehatan, & Macro International. (2013). Survei Demografi dan Kesehatan Indonesia 2012. Jakarta: Badan Pusat Statistik.

[2] Breiman L., Friedman J.H., Olshen R.A., & Stone C.J. (1984). *Classification And Regression Trees*. Monterey: Wadsworth and Brooks.

[3] Karyana, Y., Remi, S. S., Yusuf, A. A., & Purnagunawan, M. (2020). Fertilitas Dan Keputusan Menambah Anak Di Indonesia. Dalam *Pandemi Covid-19: Saatnya Kaji Ulang Arah Penelitian Dan Pendidikan Kesehatan*, ed. Sunardhi Widyaputra dan Cissy B Kartasasmita. Buku I. Unpad Press. Jatinangor.

[4] Prabawati, I.N., Widodo, & Duskarnaen, M. F. (2019). Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta. *PINTER : Jurnal Pendidikan Teknik Informatika Dan Komputer*, **3**(2), 139–145.

[5] Zhang, B., Wei, Z., Ren, J., Cheng, Y., & Zheng, Z. (2018). An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees. *IEEE Access*, **6**, 21760.

[6] Irawadi Jody Alwin, Sunendiari Siti. (2021). *Penerapan dan Perbandingan Tiga Metode Analisis Pohon Keputusan pada Klasifikasi Penderita Kanker Payudara*. Jurnal Riset Statistika, 1(1), 19-27.