BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2020 Vol. 14 No. 2 Page 267 – 278
P-ISSN: 1978-7227 E-ISSN: 2615-3017
: https://doi.org/10.30598/barekengvol14iss2pp267-278
National Accredited in SINTA 3, Decree No.: 28/E/KPT/2019

PERBANDINGAN KLASIFIKASI ALGORITMA C5.0 DENGAN *CLASSIFICATION AND REGRESSION TREE* (STUDI KASUS: DATA SOSIAL KEPALA KELUARGA MASYARAKAT DESA TELUK BARU KECAMATAN MUARA ANCALONG TAHUN 2019)

Comparison of C5.0 Algorithm Classification with Classification and Regression Tree (Case Study: Social Data of Family Head of Teluk Baru Village, Muara Ancalong District in 2019)

Reni Pratiwi[1*], Memi Nor Hayati[2], Surya Prangga[3]

[1,2,3] *Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman*
Jl Gn. Kelua, Samarinda, 75119, Indonesia

e-mail: [1] renip1725@gmail.com*
*Corresponding author**

Abstrak

Decision tree adalah pohon keputusan yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan. Banyak metode yang dapat digunakan pada *decision tree*, diantaranya adalah algoritma C5.0 dan *Classification and Regression Tree* (CART). Penelitian ini bertujuan untuk mengetahui hasil klasifikasi dari algoritma C5.0 dan CART serta untuk mengetahui perbandingan ketepatan hasil klasifikasi dari kedua metode tersebut. Adapun variabel yang digunakan dalam penelitian kali ini adalah rata-rata pendapatan perbulan (Y), pekerjaan (X_1), jumlah anggota keluarga (X_2), pendidikan terakhir (X_3) dan jenis kelamin (X_4). Setelah dilakukan analisis didapatkan hasil bahwa rata-rata tingkat akurasi algoritma C5.0 sebesar 79,17% sedangkan tingkat akurasi CART 84,63%. Sehingga dapat dikatakan bahwa metode CART merupakan metode yang lebih baik dalam pengklasifikasian data rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 dibandingkan dengan metode algoritma C5.0.

Kata Kunci : *Algoritma C5.0, CART, Klasifikasi, Pohon Keputusan.*

Abstract

Decision tree is a algorithm used as a reasoning procedure to get answers from problems are entered. Many methods can be used in decision trees, including the C5.0 algorithm and Classification and Regression Tree (CART). This research aims to determine the classification results of the C5.0 and CART algorithms and to determine the comparison of the accuracy classification results from these two methods. The variables used in this research are the average monthly income (Y), employment (X_1), number of family members (X_2), last education (X_3) and gender (X_4). After analyzing the results obtained that the accuracy rate of C5.0 algorithm is 79,17% while the accuracy rate of CART is 84,63%. So it can be said that the CART method is a better method in classifying the average income of the people of Teluk Baru Village in Muara Ancalong District in 2019 compared to the C5.0 algorithm method.

Keywords: *C5.0 Algorithm, CART, Classification, Decision Tree.*

Submitted: 02nd April 2020 *Accepted: 27th May 2020*

1. PENDAHULUAN

Decision tree atau pohon keputusan adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan. Pohon yang dibentuk tidak selalu berupa pohon biner. Jika semua fitur dalam data set menggunakan 2 macam nilai kategorikal maka bentuk pohon yang didapatkan berupa pohon biner. Jika dalam fitur berisi lebih dari 2 macam nilai kategorikal atau menggunakan tipe numerik maka bentuk pohon yang didapatkan biasanya tidak berupa pohon biner. Banyak algoritma yang dapat dipakai dalam pembentukan *decision tree* yaitu ID3, *Classification and Regression Tree* (CART), C4.5, C5.0, dan lain-lain [1].

Seiring dengan perkembangan pengetahuan klasifikasi dalam *decision tree,* maka pemakaiannya telah semakin meluas ke berbagai bidang misalnya bidang kesehatan, pertanian, asuransi, sosial dan lain-lain [2]. Aplikasi metode klasifikasi dalam bidang sosial salah satunya adalah untuk melihat tingkat kesejahteraan di suatu daerah. Menurut [3] dalam BPS Sumut (2012) salah satu cara untuk menentukan tingkat kesejahteraan secara nyata dapat diukur melalui tingkat pendapatan masyarakat.

Penelitian terdahulu mengenai metode Algoritma C5.0 pernah dilakukan oleh [4] mengenai metode CART pernah dilakukan oleh [5] dan mengenai perbandingan algoritma C5.0 dengan CART pernah dilakukan oleh [6].

Berdasarkan latar belakang tersebut maka dapat diperoleh tujuan dari penelitian ini yaitu untuk memperoleh hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma C5.0 dengan proporsi data *training* dan data *testing* sebesar 90:10, untuk memperoleh hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma CART dengan proporsi data *training* dan data *testing* sebesar 90:10 dan untuk memperoleh perbandingan ketepatan hasil klasifikasi metode algoritma C5.0 dengan metode algoritma CART menggunakan *confusion matrix.*

Algoritma C5.0 merupakan perluasan dari algoritma C4.5 yang juga perpanjangan dari ID3. Algoritma C5.0 adalah klasifikasi algoritma yang cocok untuk kumpulan data besar. Algoritma C5.0 lebih baik daripada C4.5 pada kecepatan, memori, dan efisiensi. Dalam algoritma C5.0, pemilihan atribut yang akan diproses menggunakan ukuran *gain ratio*. Ukuran *gain ratio* digunakan untuk memilih atribut uji pada setiap *node* di dalam *tree.* Ukuran ini digunakan untuk memilih atau membentuk *node* pada pohon. Atribut dengan nilai *gain ratio* tertinggi akan terpilih sebagai *parent* bagi *node* selanjutnya [7]. Langkah kerja pembuatan *tree* pada algoritma C5.0 mirip dengan pembuatan *tree* pada algoritma C4.5. Kemiripan tersebut meliputi perhitungan *entropy* dan *gain.* Jika pada algoritma C4.5 berhenti sampai perhitungan *gain,* maka pada algoritma C5.0 akan melanjutkannya dengan menghitung *gain ratio* dengan menggunakan *gain* dan *entropy* yang telah ada. Adapun rumus untuk mencari nilai *entropy* adalah sebagai berikut:

$$Entropy(S) = - \sum_{j=1}^{k} p_j \log_2 p_j \qquad (1)$$

dengan *S* merupakan himpunan kasus, *k* adalah jumlah kelas pada variabel A dan p_j adalah proporsi dari S_j.

Selanjutnya untuk mencari nilai *gain* digunakan persamaan berikut :

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

dengan S_i merupakan himpunan kasus pada kategori ke-*i,* A adalah variabel yang digunakan, *m* adalah jumlah kategori pada variabel *A*, $|S_i|$ merupakan jumlah kasus pada kategori ke-*i*, dan $|S|$ merupakan jumlah kasus dalam *S.* Setelah didapat nilai *entropy* dan *gain,* selanjutnya adalah menghitung nilai *gain ratio*. Adapun rumus dasar dari perhitungan *gain ratio* adalah sebagai berikut :

$$Gain\ Ratio = \frac{Gain(S, A)}{\sum_{i=1}^{m} Entropy(S_i)} \qquad (3)$$

dengan $Gain(S, A)$ merupakan nilai *gain* dari suatu variabel dan $\sum_{i=1}^{m} Entropy(S_i)$ merupakan jumlah nilai *entropy* dalam suatu variabel.

Proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing [8]. CART merupakan salah satu metode atau algoritma dari salah satu teknik pohon keputusan. CART terbilang sederhana namun merupakan metode yang kuat. CART bertujuan untuk mendapatkan suatu kelompok data yang akurat sebagai tanda dari suatu pengklasifikasian. Selain itu CART juga dapat digunakan untuk menggambarkan hubungan antara variabel terikat dengan satu atau lebih variabel bebas. Model pohon yang dihasilkan bergantung pada skala variabel terikat, jika variabel terikat data berbentuk kontinu maka model pohon yang dihasilkan adalah *regression tree* (pohon regresi) sedangkan bila variabel terikat mempunyai skala data kategorik maka pohon yang dihasilkan adalah *classification tree* (pohon klasifikasi) [9]. Proses pembentukan pohon klasifikasi pada algoritma CART melalui tiga tahapan, yaitu:

a. Pemilihan Pemilah
   Menurut [10], rumus pemilah disajikan seperti berikut:
   - Variabel bebas kontinu = $b-1$ pemilahan (4a)
   - Variabel bebas kategori nominal = $2^{L-1}-1$ pemilahan (4b)
   - Variabel bebas kategori ordinal = $L-1$ pemilahan (4c)

   dengan *b* merupakan banyaknya data pada suatu variabel dan *L* merupakan banyaknya kategori pada suatu variabel

   Fungsi keheterogenan yang digunakan adalah Indeks Gini karena akan selalu memisahkan kelas dengan anggota paling besar/kelas terpenting dalam simpul terlebih dahulu. Indeks *gini* dari pembelahan tersebut didefinisikan sebagai berikut:

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2) \quad (5)$$

   dengan $Gini_{pembelahan}$ adalah nilai indeks *gini* setiap variabel, $gini(D_1)$ adalah nilai indeks *gini* subset D_1 pada setiap variabel, $gini(D_2)$ adalah nilai indeks *gini* subset D_2 pada setiap variabel, *b* adalah banyaknya data pada suatu variabel, b_1 adalah banyaknya data pada subset D_1 dan b_2 adalah banyaknya data pada subset D_2

b. Penentuan *Node* Terminal
   Menurut [9] suatu *node t* akan menjadi *node* terminal atau tidak, akan dipilah kembali apabila terdapat batasan minimum *n* seperti hanya terdapat satu pengamatan pada tiap *node* anak. Umumnya jumlah kasus minimum dalam suatu terminal akhir adalah 5, dan apabila hal itu terpenuhi maka pengembangan pohon akan dihentikan.

c. Penandaan Label Kelas
   Penandaan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak dengan persamaan:

$$P(j_0 \mid t) = \max_j P(j \mid t) = \max_j \frac{m_j(t)}{m(t)} \quad (6)$$

   dengan $P(j \mid t)$ merupakan probabilitas bersyarat kelas *j* yang berada pada *node* t, $m_j(t)$ adalah jumlah pengamatan pada kelas *j* pada *node t* dan $m(t)$ adalah kumlah pengamatan pada *node t.*
   Label kelas *node* terminal *t* adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian *node t* terbesar.

Menurut [1] data yang akan digunakan dalam pengujian klasifikasi dibagi menjadi dua yaitu data *training* dan data *testing*. Model klasifikasi kemudian dibangun berdasarkan data *training* dan kemudian kinerjanya diukur berdasarkan data *testing*. Proporsi pembagian data *training* dan data *testing* biasanya diskrit, misal 90:10 (artinya 90% sebagai data *training* dan 10% data *testing*) serta 50:50 (artinya 50% sebagai data *training* dan 50% data *testing*). Jumlah data *training* dan data *testing* dapat dihitung menggunakan persamaan 7a dan 7b dengan N merupakan jumlah data yang akan digunakan sebagai sampel seperti berikut:

Jumlah data *training* = Proporsi data *training* × N (7a)
Jumlah data *testing* = N - Jumlah data *training* (7b)

Salah satu alat bantu untuk menilai seberapa baik sebuah *classifier* adalah *confusion matrix.* Tabel *confusion matrix* dihasilkan dari aplikasi model pada *test set*. Dari *confusion matrix* dapat diturunkan berbagai *metric* evaluasi *classifier* seperti akurasi, *specificity, sensitivity,* dan lain-lain. *Actual class* adalah kelas yang

sebenarnya pada *test set. Predicted class* adalah kelas hasil prediksi dari model yang dihasilkan oleh *classifier*. *True positive* (TP) adalah jumlah baris kelas C1 pada *test set* yang benar diklasifikasikan sebagai kelas C1 oleh *classifier*. *False negative* (FN) adalah jumlas baris berlabel C1 pada *test set* namun diklasifikasikan sebagai bukan kelas C1 oleh *classifier*. *False positive* (FP) adalah jumlah baris berlabel kelas bukan C1 pada *test set*, namun diklasifikasikan sebagai kelas C1 oleh *classifier*. *True negative* (TN) adalah jumlah baris berlabel kelas bukan C1 pada *test set* dan benar diklasifikasikan sebagai kelas bukan C1 oleh *classifier* [11].
Menurut [11], akurasi adalah persentase baris *test set* yang diklasifikasikan dengan benar, berikut rumusnya:

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

Semua algoritma klasifikasi berusaha membentuk model yang mempunyai akurasi tinggi. Umumnya, model yang dibangun dapat diprediksi dengan benar pada semua data yang menjadi data latihnya, tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan [12].

Dalam kamus ekonomi, pendapatan (*income*) adalah uang yang diterima seseorang dalam perusahaan dalam bentuk gaji, upah, sewa, bunga, laba dan lain sebagainya Bersama dengan tunjangan pengangguran, uang pension dan lainnya [13]

Pendapatan seseorang dapat didefinisikan sebagai banyaknya penerimaan yang dinilai dengan satuan mata uang yang dapat dihasilkan seseorang atau suatu bangsa dalam periode tertentu. Sehingga dapat didefinisikan: "Pendapatan (*revenue*) dapat diartikan sebagai total penerimaan yang diperoleh pada periode tertentu" [14]. Sedangkan menurut [15], pendapatan seseorang dipengaruhi oleh beberapa faktor, antara lain sebagai berikut:

1) Jumlah faktor-faktor produksi yang dimiliki yang bersumber pada hasil-hasil tabungan tahun ini dan warisan atau pemberian.
2) Harga per unit dari masing-masing faktor produksi, harga ini ditentukan oleh penawaran dan permintaan di pasar faktor produksi.
3) Hasil kegiatan anggota keluarga sebagai pekerjaan sampingan.

2. METODE PENELITIAN

Populasi dalam penelitian kali ini adalah data sosial seluruh Kepala Keluarga (KK) masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019 yang diperoleh langsung dari desa tersebut. Sedangkan yang menjadi sampel hanya 100 KK. Variabel yang digunakan dalam penelitian ini terdiri dari variabel bebas (*X*) dan variabel terikat (*Y*) sebagai berikut:

1. Rata-rata pendapatan perbulan sebagai variabel terikat (*Y*) yang mengacu pada Upah Minimum Kabupaten (UMK) Kutai Timur tahun 2019 berdasarkan Surat Keputusan (SK) nomor 561/K.555/2018 tentang penetapan upah minimum Kabupaten Kutai Timur tahun 2019 yang menyatakan bahwa UMK Kutai Timur tahun 2019 akan ditetapkan sebesar 2,89 juta, sehingga dapat dikategorikan menjadi:

$$Y = \begin{cases} 1, \text{Jika rata-rata pendapatan perbulan} < 2{,}89 \text{ juta} \\ 2, \text{Jika rata-rata pendapatan perbulan} \geq 2{,}89 \text{ juta} \end{cases}$$

2. Pekerjaan sebagai variabel bebas (X_1) dengan kategori:

$$X_1 = \begin{cases} 1, \text{ Jika pekerjaan Petani} \\ 2, \text{ Jika pekerjaan Nelayan} \\ 3, \text{Jika pekerjaan PNS} \\ 4, \text{Jika pekerjaan Swasta} \\ 5, \text{Jika pekerjaan Wiraswasta} \end{cases}$$

3. Jumlah anggota keluarga sebagai variabel bebas (X_2) yang mengacu pada program Keluarga Berencana (KB) oleh Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) dalam [3]

menyatakan bahwa memiliki 2 anak lebih baik dan sesuai dengan slogan BKKBN yaitu "2 anak cukup" sehingga dapat dikategorikan menjadi :

$$X_2 = \begin{cases} 1, & \text{Jika jumlah anggota keluarga lebih dari 4 (>4)} \\ 2, & \text{Jika jumlah anggota keluarga kurang dari atau sama dengan 4 (} \leq 4) \end{cases}$$

4. Pendidikan terakhir sebagai variabel bebas (X_3) dengan kategori :

$$X_3 = \begin{cases} 1, & \text{Jika lulusan SD/Sederajat} \\ 2, & \text{Jika lulusan SMP/Sederajat} \\ 3, & \text{Jika lulusan SMA/Sederajat} \\ 4, & \text{Jika lulusan Perguruan Tinggi (PT)} \end{cases}$$

5. Jenis kelamin sebagai variabel bebas (X_4) dengan kategori :

$$X_4 = \begin{cases} 1, & \text{Jika jenis kelamin Perempuan} \\ 2, & \text{Jika jenis kelamin Laki-laki} \end{cases}$$

Tahap analisis data adalah sebagai berikut:

a. Algoritma C5.0
   1. Penentuan variabel yang akan diteliti.
   2. Pemilihan *node* akar diawali dengan menghitung nilai *entropy* menggunakan Persamaan (1). Kemudian proses dilanjutkan dengan mencari nilai *gain* menggunakan Persamaan (2). Setelah itu mencari nilai *gain ratio* pada Persamaan (3)
   3. Penentuan cabang untuk masing-masing *node* dengan menghitung nilai *gain ratio* tertinggi dari variabel bebas yang ada. Perhitungan untuk menentukan cabang pada metode ini dilakukan secara manual dan peneliti menggunakan bantuan *software Microsoft Excel* 2010.
   4. Kelas dibagi dalam cabang yang telah ditentukan.
   5. Ulangi langkah 1-3 hingga semua kelas pada cabang memiliki kelasnya masing-masing.

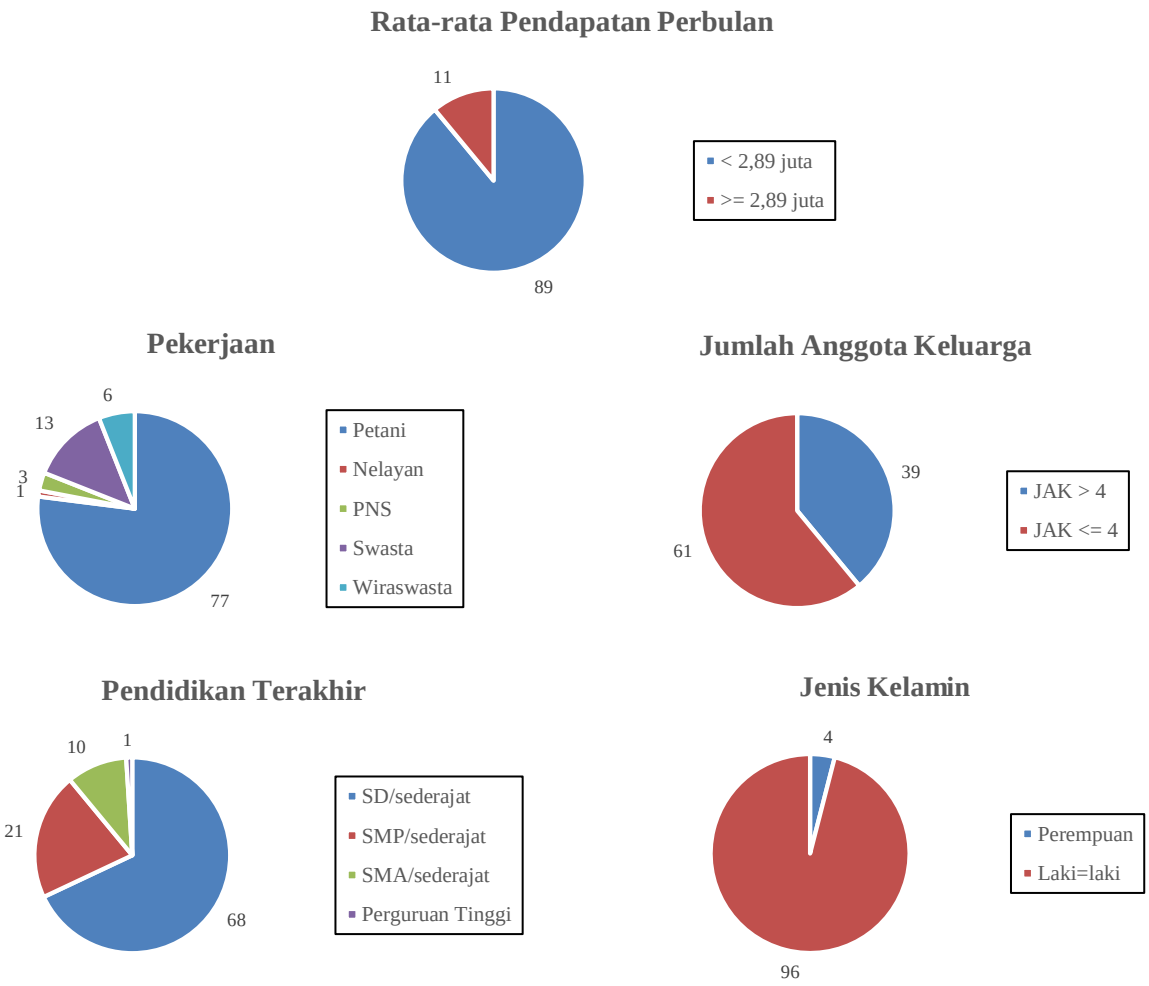
b. CART
   1. Penentuan variabel yang akan diteliti.
   2. Penentuan banyaknya pemilah pervariabel sesuai dengan jenis variabel bebasnya menggunakan persamaan (4a), (4b) dan (4c).
   3. Menghitung nilai *indeks gini* untuk setiap pemilah sesuai dengan persamaan (5) kemudian pemilah yang memiliki nilai *indeks gini* terkecil akan dipilih menjadi pemilah terbaik.
   4. Ulangi langkah 2-3 hingga tidak memungkinkan lagi untuk melakukan pemilahan.
   5. Penandaan label kelas *node* terminal berdasarkan aturan jumlah anggota terbanyak menggunakan persamaan (6).

3. HASIL DAN PEMBAHASAN

Setelah dilakukan penelitian terhadap data maka didapatkan hasil sebagai berikut:

3.1 Analisis Statistika Deskriptif

Analisis statistika deskriptif ini dilakukan untuk mengetahui karakteristik dari data yang akan diteliti. Pada bagian ini akan dibahas mengenai deskripsi untuk setiap variabel yang digunakan. Variabel-variabel tersebut adalah rata-rata pendapatan perbulan (Y), pekerjaan (X_1), jumlah anggota keluarga (X_2), pendidikan terakhir (X_3) dan jenis kelamin (X_4).



Gambar 1. Diagram Lingkaran untuk Setiap Variabel

3.2 Pembagian Data *Training* dan *Testing*

Sebelum melakukan proses klasifikasi, langkah pertama yang perlu dilakukan adalah membagi data *training* dan *testing*, kemudian dilakukan pengacakan terlebih dahulu agar setiap data memiliki kesempatan yang sama untuk menjadi data *training* dan *testing*. Pengacakan data dilakukan dengan menggunakan bantuan *software Microsoft Excel* 2010. Berdasarkan hasil perhitungan menggunakan Persamaan (7a) dan (7b) dapat diketahui bahwa data yang masuk ke dalam data *training* untuk porposi 90:10 sebanyak 90 dan sisanya sebanyak 10 data masuk ke dalam data *testing*.

3.3 Algoritma C5.0

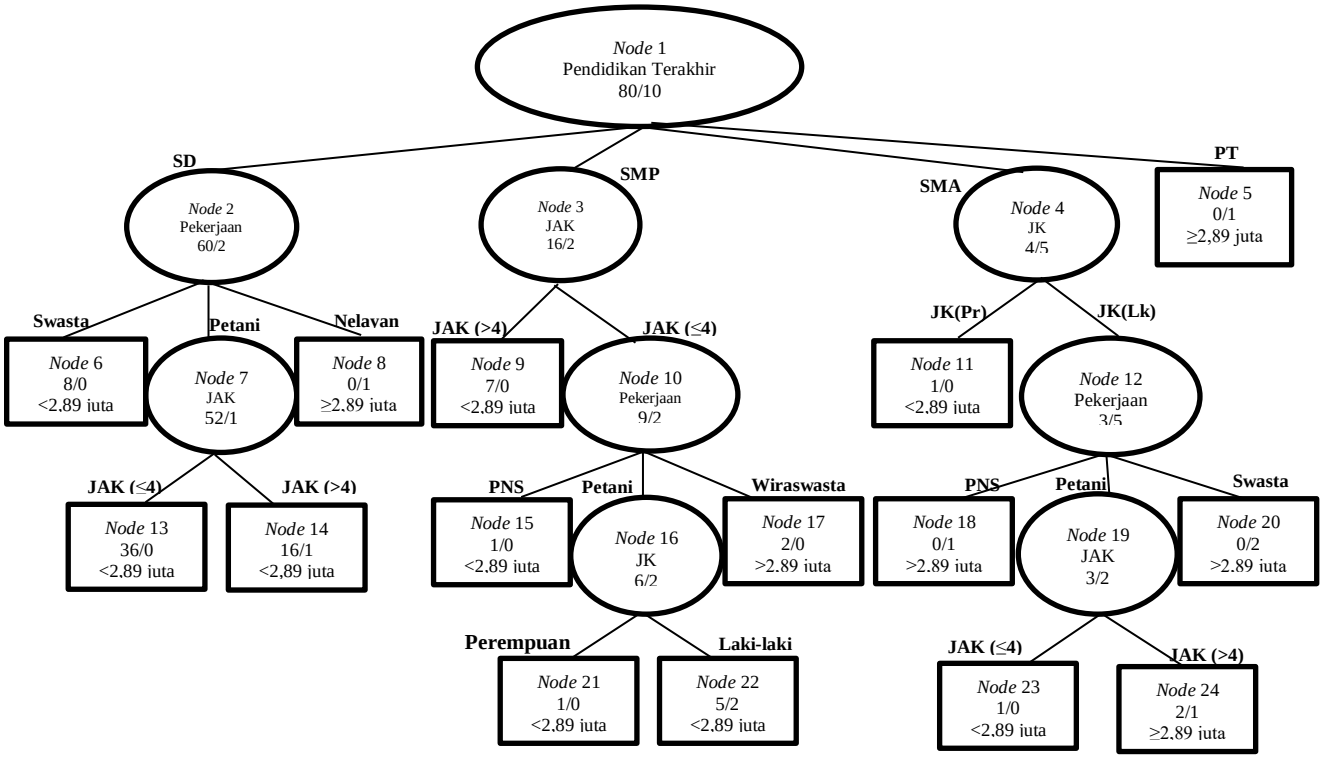
Pada proses pembentukan pohon klasifikasi algoritma C5.0 tahap pertama yaitu menentukan *node* akar, kemudian dilanjutkan dengan penentuan cabang untuk masing-masing *node*. Selanjutnya dilakukan pembagian kelas pada cabang yang telah diperoleh dan proses tersebut diulang hingga setiap cabang memiliki kelas. Adapun data yang digunakan untuk proses pembentukan pohon klasifikasi yaitu 90% dari keseluruhan data yakni sebanyak 90 sampel (data *training*), sedangkan sisanya 10% dari keseluruhan data yakni sebanyak 10 sampel digunakan sebagai data *testing* untuk pohon klasifikasi yang telah terbentuk.

Tahap pertama dalam pembentukan pohon adalah pemilihan *node* akar dengan menghitung nilai *entropy* menggunakan Persamaan (1), nilai *gain* menggunakan Persamaan (2) dan nilai *gain ratio* menggunakan Persamaan (3). Adapun hasil perhitungan nilai *entropy*, *gain* dan *gain ratio* secara lengkap disajikan pada Tabel 1, berikut:

Tabel 1. Hasil Perhitungan *Entropy, Gain* dan *Gain Ratio* untuk *Node* Akar

Node	Variabel		Jumlah Kasus (S)	< 2,89	≥ 2,89	*Entropy*	*Gain*	*Gain Ratio*
1	Total		90	80	10	0,5033		
	Pekerjaan	Petani	71	66	5	0,3675	0,0631	0,0227
		Nelayan	1	0	1	0		
		PNS	3	2	1	0,9183		
		Swasta	11	9	2	0,6840	0,0631	0,0227
		Wiraswasta	4	3	1	0,8113		
	Jumlah Anggota Keluarga	>4	33	29	4	0,5328	0,0004	0,0004
		≤4	57	51	6	0,4855		
	Pendidikan Terakhir	**SD**	**62**	**60**	**2**	**0,2056**	**0,1619**	**0,0952**
		SMP	**18**	**16**	**2**	**0,5033**		
		SMA	**9**	**4**	**5**	**0,9911**		
		PT	**1**	**0**	**1**	**0**		
	Jenis Kelamin	Perempuan	3	3	0	0	0,0058	0,0112
		Laki-laki	87	77	10	0,5146		

Perhitungan *entropy*, *gain* dan *gain ratio* dilakukan terus hingga setiap cabang memiliki kelasnya masing-masing. Sehingga terbentuklah pohon klasifikasi sebagai berikut:



Gambar 2. Pohon Klasifikasi Algoritma C5.0

Pada Gambar 2, dapat disimpulkan bahwa:

1. Apabila seseorang memiliki pendidikan terakhir SD dan memiliki pekerjaan swasta maka dapat diprediksi rata-rata pendapatan perbulan sebesar <2,89 juta sedangkan yang memiliki pekerjaan nelayan diprediksi memiliki rata-rata pendapatan sebesar ≥ 2,89 juta. Apabila memiliki pekerjaan petani dan mempunyai jumlah anggota keluarga sebanyak >4 dan ≤4 maka dapat diprediksi rata-rata pendapatan perbulan sebesar <2,89 juta.
2. Apabila seseorang memiliki pendidikan terakhir SMP dan memiliki jumlah anggota keluarga sebanyak >4 maka dapat diprediksi rata-rata pendapatan perbulan sebesar <2,89 juta sedangkan yang memiliki jumlah anggota keluarga sebanyak ≤4 dengan pekerjaan PNS, Petani dan Wiraswasta maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta.

3. Apabila seseorang memiliki pendidikan terakhir SMA dan memiliki jenis kelamin Perempuan maka dapat diprediksi rata-rata pendapatan perbulan sebesar <2,89 juta sedangkan yang memiliki jenis kelamin Laki-laki yang memiliki pekerjaan PNS dan Swasta maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta dan untuk yang memiliki pekerjaan Petani dengan jumlah anggota keluarga sebanyak >4 pun dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta sedangkan yang memiliki jumlah anggota keluarga <4 dapat diprediksi memiliki rata-rata pendapatan perbulan <2,89 juta.
4. Apabila seseorang memiliki pendidikan terakhir PT dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta. Berikut adalah interpretasi atau kesimpulan yang didapat dari pohon klasifikasi algoritma C5.0:

3.4 Algoritma CART

Berikut analisis menggunakan metode CART:

3.4.1 Pembentukan Pohon Klasifikasi

Dalam pembentukan pohon klasifikasi CART, terdapat 3 tahap yaitu pemilihan pemilah, penentuan terminal *node* dan penandaan label kelas. Adapun sebagai contoh data yang digunakan untuk proses pembentukan pohon klasifikasi yaitu 90% dari keseluruhan data yakni sebanyak 90 sampel (data *training*), sedangkan sisanya 10% dari keseluruhan data yakni sebanyak 10 sampel digunakan sebagai data *testing* untuk pohon klasifikasi yang telah terbentuk.

Tahap awal dalam pembentukan pohon klasifikasi CART adalah menentukan banyak pemilah pada setiap variabel bebas. Pemilihan pemilah untuk variabel bebas bertipe nominal yaitu variabel Pekerjaan (X_1) dan Jenis Kelamin (X_4) menggunakan Persamaan (4b). Kemudian pemilihan pemilah untuk variabel bebas bertipe ordinal yaitu variabel Jumlah Anggota Keluarga (X_2) dan Pendidikan Terakhir (X_3) menggunakan Persamaan (4c).

Langkah selanjutnya yaitu menghitung nilai indeks *gini* untuk setiap pemilah sesuai dengan Persamaan (5) kemudian pemilah yang memiliki nilai indeks *gini* terkecil akan dipilah menjadi pemilah terbaik. Klasifikasi dalam penelitian ini dibagi menjadi 2 kelas, yaitu D_1: Jika pendapatan rata-rata perbulan <2,89 juta dan D_2: Jika pendapatan rata-rata perbulan ≥2,89 juta. Adapun hasil perhitungan nilai indeks *gini* secara lengkap disajikan pada Tabel 2 berikut:

Tabel 2. Hasil Perhitungan Indeks *Gini* untuk *Node* 1

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani), (Nelayan, PNS, Swasta, Wiraswasta)}	0,1852
	{(Nelayan), (Petani, PNS, Swasta, Wiraswasta)}	0,1798
	{(PNS), (Petani, Nelayan, Swasta, Wiraswasta)}	0,1941
	{(Petani, PNS), (Nelayan, Swasta, Wiraswasta)}	0,1892
	{(Petani, Swasta), (Nelayan, PNS, Wiraswasta)}	0,1839
	{(Petani, Wiraswasta), (Nelayan, PNS, Swasta)}	0,1879
	{(Petani, Nelayan, PNS), (Swasta, Wiraswasta)}	0,1944
	{(Petani, Nelayan, Swasta), (PNS, Wiraswasta)}	0,1924
	{(Petani, Nelayan, Wiraswasta), (PNS, Swasta)}	0,1936
	{(Nelayan, PNS), (Petani, Swasta, Wiraswasta)}	0,1835
	{(Nelayan, Swasta), (Petani, PNS, Wiraswasta)}	0,1916
	{(Nelayan, Wiraswasta), (Petani, PNS, Swasta)}	0,1877
Jumlah Anggota Keluarga	{(>4), (≤4)}	0,1974
Pendidikan Terakhir	{(SD), (SMP, SMA, PT)}	0,1700
	{(SD, SMP), (SMA, PT)}	**0,1378**
	{(SD, SMP, SMA), (PT)}	0,1798
Jenis Kelamin	{(Perempuan), (Laki-laki)}	0,1967

Perhitungan indeks *gini* dilakukan terus hingga setiap cabang memiliki kelasnya masing-masing.

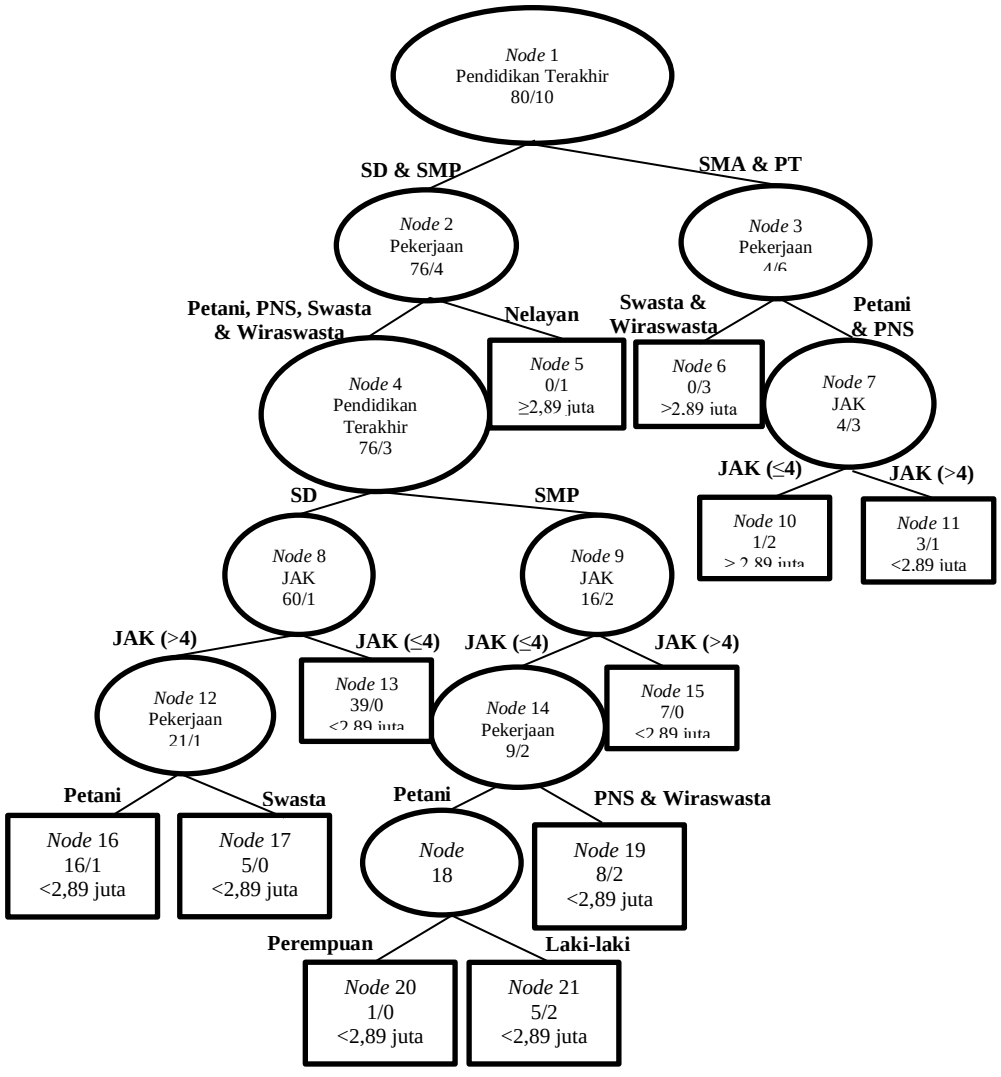
3.4.2 Pemberian Label Kelas *Node* Terminal

Pemberian label kelas untuk setiap *node* terminal menggunakan Persamaan (6). Adapun hasil perhitungan peluang tiap kelas rata-rata pendapatan pada masing-masing *node* terminal disajikan pada Tabel 3 berikut:

Tabel 3. Perhitungan Peluang Tiap Kelas dalam *Node* Terminal

Node	*P(<2,89 juta)*	*P(≥2,89 juta)*	Keputusan
5	0	1	≥2,89 juta
6	0	1	≥2,89 juta
10	0,75	0,25	<2,89 juta
11	0,3333	0,6667	≥2,89 juta
13	1	0	<2,89 juta
15	1	0	<2,89 juta
16	0,9412	0,0588	<2,89 juta
17	1	0	<2,89 juta
19	1	0	<2,89 juta
20	1	0	<2,89 juta

Karena tidak lagi memungkinkan untuk membuat cabang baru, maka proses pembuatan pohon dihentikan sehingga didapatkan sebuah pohon klasifikasi. Hasil akhir *decision tree* untuk metode algoritma CART disajikan pada Gambar 3.



Gambar 3. Pohon Klasifikasi CART

Pada Gambar 3, dapat disimpulkan bahwa:

1. Apabila seseorang memiliki pendidikan terakhir SD dan SMP dengan pekerjaan nelayan maka dapat diprediksi rata-rata pendapatan perbulan sebesar ≥2,89 juta. Sedangkan apabila seseorang memiliki pendidikan terakhir SD dengan jumlah anggota keluarga ≤4 maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar <2,89 juta sedangkan yang memiliki jumlah anggota keluarga >4 dengan pekerjaan Petani dan Swasta maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar <2,89 juta. Apabila seseorang memiliki pendidikan terakhir SMP dengan jumlah anggota keluarga >4 maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar <2,89 juta, sedangkan yang memiliki jumlah anggota ≤4 dengan pekerjaan Petani, PNS dan Wiraswasta dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar <2,89 juta,
2. Apabila seseorang memiliki pendidikan terakhir SMA dan PT dan memiliki pekerjaan Swasta dan Wiraswasta maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta. Apabila seseorang memiliki pekerjaan Petani dan PNS dengan jumlah anggota keluarga >4 maka dapat diprediksi memiliki rata-rata pendapatan perbulan sebesar <2,89 juta sedangkan yang memiliki jumlah anggota keluarga ≤4 diprediksi memiliki rata-rata pendapatan perbulan sebesar ≥2,89 juta.

3.5 Mengukur Ketepatan Hasil Klasifikasi

Setelah didapatkan hasil berupa pohon klasifikasi, maka perlu diuji tingkat ketepatannya dengan menggunakan bantuan tabel *confusion matrix* untuk masing-masing metode yaitu algoritma C5.0 dan CART.

3.5.1 Ketepatan Klasifikasi Algoritma C5.0

Berikut tabel *confusion matrix* untuk mengukur tingkat ketepatan hasil klasifikasi menggunakan algoritma C5.0 dengan proporsi data 90:10:

Tabel 4. Ketepatan Klasifikasi Algoritma C5.0

Kelas Aktual	Kelas Prediksi		Total
	< 2,89 juta	≥2,89 juta	
< 2,89 juta	9	0	9
≥ 2,89 juta	1	0	1
Total	10	0	10

Dapat dilihat pada Tabel 4, hasil dari *confusion matrix* untuk algoritma C5.0 dengan proporsi data 90:10 sehingga dapat dihitung tingkat ketepatan akurasi menggunakan Persamaan (8) sehingga didapat nilai tingkat akurasi sebesar 90%. Tingkat akurasi sebesar 90% menyatakan bahwa dari 90 KK, terdapat 81 orang yang tepat diklasifikasikan.

3.5.2 Ketepatan Klasifikasi Algoritma CART

Berikut tabel *confusion matrix* untuk mengukur tingkat ketepatan hasil klasifikasi menggunakan algoritma CART dengan porporsi data 90:10:

Tabel 5. Ketepatan Klasifikasi Algoritma CART

Kelas Aktual	Kelas Prediksi		Total
	< 2,89 juta	≥2,89 juta	
<2,89 juta	8	1	9
≥2,89 juta	1	0	1
Total	9	1	10

Dapat dilihat pada Tabel 5, hasil dari *confusion matrix* untuk algoritma CART dengan proporsi data 90:10 sehingga dapat dihitung tingkat ketepatan akurasi menggunakan Persamaan (9) sehingga didapat nilai

tingkat akurasi sebesar 80%. Tingkat akurasi sebesar 80% menyatakan bahwa dari 90 KK, terdapat 72 orang yang tepat diklasifikasikan.

3.5.3 Perbandingan Tingkat Akurasi Algoritma C5.0 dan CART

Berdasarkan analisis data yang telah dilakukan maka didapatkan tingkat akurasi dari masing-masing metode untuk setiap proporsi dapat dilihat pada Tabel 6 berikut:

Tabel 6 Perbandingan Tingkat Akurasi Kedua Metode untuk Setiap Proporsi

	50:50:00	60:40:00	70:30:00	80:20:00	90:10:00	Rata-rata
Algoritma C5.0	80%	87,50%	63,33%	75%	90%	79,17%
CART	94%	87,50%	86,67%	75%	80%	84,63%

Dapat dilihat pada Tabel 6, bahwa rata-rata tingkat akurasi CART sebesar 84,63% sedangkan tingkat akurasi algoritma C5.0 hanya 79,17%. Sehingga dapat dikatakan bahwa metode CART merupakan metode yang lebih baik dalam pengklasifikasian data rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 dibandingkan dengan metode algoritma C5.0.

4. KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang dilakukan, diperoleh kesimpulan bahwa hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 menggunakan metode algoritma C5.0 dengan porporsi 90:10 memperoleh tingkat akurasi tertinggi yaitu sebesar 90% sedangkan apabila menggunakan metode CART dengan porporsi 50:50 diperoleh tingkat akurasi tertinggi yaitu sebesar 94%. Hasil rata-rata tingkat akurasi ketepatan klasifikasi algoritma C5.0 sebesar 79,17% sedangkan metode CART sebesar 84,63%. Sehingga dapat dikatakan bahwa metode CART merupakan metode yang lebih baik dalam pengklasifikasian data rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 dibandingkan dengan metode algoritma C5.0.

DAFTAR PUSTAKA

[1] Prasetyo, E, *Data Mining-Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit Andi, 2014.

[2] Mardiani, "Perkembangan Algoritma untuk Menghitung Pola yang Sering Muncul pada Basis Data yang Besar," *Seminar Aplikasi Teknologi Informasi*, Juni, 2012.

[3] Rosni, "Analisis Tingkat Kesejahteraan Masyarakat Nelayan di Desa Dahari Selebar Kecamatan Talawi Kabupaten Batubara," *Jurnal Geografi*, pp. 53-66, 2012.

[4] Wijaya, A. C, Hasibuan, N. A dan Ramadhani, P., "Implementasi Algoritma C5.0 dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus: Kelurahan Mesjid Kecamatan Medan Kota)," *Majalah Ilmiah INTI.* vol. 13, no. 2, pp. 192-198, Mei 2018.

[5] Pakpahan, H. S, Indar, F dan Wati, M., "Penerapan Algoritma CART Decision Tree pada Penentuan Penerima Program Bantuan Pemerintah Daerah Kabupaten Kutai Kartanegara," *JURTI*, vol. 2, no. 1, pp. 27-36, Juni 2018.

[6] Yusuf, Y. W., "Perbandingan Performansi Algoritma Decision Tree C5.0, CART dan CHAID: Kasus Prediksi Status Resiko Kredit Bank X," *Seminar Nasional Aplikasi Teknologi Informasi*, pp. 59-62, Juni, 2007.

[7] Kusrini dan Luthfi, E. T., *Algoritma Data Mining.* Yogyakarta: Penerbit Andi, 2009.

[8] Putri, Y. R., Mukhlash, I. dan Hidayat, N., "Prediksi Pola Kecelakaan Kerja pada Perusahaan Non Ekstraktif Menggunakan Algoritma Decision Tree: C4.5 dan C5.0," *Jurnal Sains dan Seni Pomits*, vol. 2, no. 1, pp. 1-6, 2013.

[9] Pratiwi, F. E. dan Zain, I., "Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara," *Jurnal Sains dan Seni Pomits*, vol. 3, no. 1, pp. 54-59, 2014.

[10] Akbar, M. S., Yuanita, D. dan Harini, S., "Pendekatan CART untuk Mendapatkan Faktor yang Mempengaruhi Terjangkitnya Penyakit Demam Tifoid di Aceh Utara," *Jurnal CAUCHY.* vol. 1, no. 2, pp. 71-77, 2010.

[11] Pramana, S, Yuniarto, B, Mariyah, Siti, Santoso, Ibnu dan Nooraeni., *Data Mining dengan R. Konsep Serta Implementasi*. Bogor: Penerbit IN MEDIA. 2018.

[12] Prasetyo, E. (2012). *Data Mining-Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Penerbit Andi.

[13] Pass, Christopher dan Bryan Lowes. (1994). *Kamus Lengkap Ekonomi*, Edisi Kedua. Jakarta: Erlangga.
[14] Reksoprayitno. (2004). *Sistem Ekonomi dan Demokrasi Ekonomi*. Jakarta: Bina Grafika.
[15] Boediono. (2002). *Pengantar Ekonomi*. Jakarta: Erlangga.