PERBANDINGAN KLASIFIKASI ALGORITMA C5.0 DAN *CLASSIFICATION AND REGRESSION TREE*
(Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)

SKRIPSI

RENI PRATIWI
NIM. 1607015031

PROGRAM STUDI STATISTIKA
JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS MULAWARMAN
SAMARINDA
2020

PERBANDINGAN KLASIFIKASI ALGORITMA C5.0 DAN *CLASSIFICATION AND REGRESSION TREE* (Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)

SKRIPSI

Diajukan kepada
Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Mulawarman untuk memenuhi sebagian persyaratan
memperoleh gelar Sarjana Sains bidang Ilmu Statistika

Oleh:

Reni Pratiwi
NIM. 1607015031

PROGRAM STUDI STATISTIKA
JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS MULAWARMAN
SAMARINDA
2020

HALAMAN PENGESAHAN

Skripsi/Tesis Sarjana berjudul **Perbandingan Klasifikasi Algoritma C5.0 dan *Classification and Regression Tree* (Studi Kasus: Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)** Oleh **Reni Pratiwi** telah dipertahankan di depan Dewan Penguji pada tanggal 10 Maret 2020.

SUSUNAN TIM PEMBIMBING

Menyetujui,

Pembimbing I,	Pembimbing II,
Memi Nor Hayati, S.Si, M.Si	**Surya Prangga, S.Si, M.Si**
NIP. 19880503 201404 2 001	NIP. 19920926 201903 1 008

Mengetahui,
Dekan FMIPA Universitas Mulawarman

Dr. Eng. Iris Mandang, M.Si
NIP. 19711008 199802 1 001

PERNYATAAN KEASLIAN SKRIPSI/TESIS

Dengan ini saya menyatakan bahwa dalam Skripsi/Tesis yang berjudul "Perbandingan Klasifikasi Algoritma C5.0 dan *Classification and Regression Tree* (Studi Kasus: Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Sarjana/Magister di suatu perguruan tinggi manapun. Sepanjang pengetahuan saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademik dikemudian hari apabila pernyataan yang saya buat ini tidak benar.

Samarinda, Maret 2020

Reni Pratiwi

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). Algoritma C5.0 merupakan pohon keputusan *non biner* di mana cabang pohon bisa lebih dari dua sedangkan algoritma CART merupakan pohon keputusan *biner* di mana cabang pohon hanya terdiri dari dua cabang. 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). C5.0 algorithm is a non-binary decision tree where the branch of tree can be more than two, while the CART algorithm is a binary decision tree where the branch of tree consists of only two branches. 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.

KATA PENGANTAR

Puji dan syukur kepada Allah SWT yang telah melimpahkan rahmat serta hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini. Skripsi ini disusun untuk memenuhi persyaratan guna memperoleh gelar Sarjana Statistika (S.Stat) di Fakultas Matematika dan Ilmu Pengetahuan Alam Program Studi Statistika Jurusan Matematika Universitas Mulawarman, dengan judul "**Perbandingan Klasifikasi Algoritma C5.0 dan *Classification and Regression Tree* (Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)**"

Penulis menyadari bahwa dalam penulisan skripsi ini terdapat beberapa pihak yang turut berkontribusi memberikan bantuan, dorongan serta masukan-masukan yang sangat berarti. Oleh karena itu, dalam kesempatan ini penulis mengucapkan terima kasih kepada pihak-pihak yang telah membantu dan membimbing dalam penyusunan laporan ini, diantaranya:

1. Allah SWT Sang Maha Pencipta yang telah memberikan kemudahan hingga penulis dapat menyelesaikan segala sesuatunya dengan lancar.
2. Ibu Memi Nor Hayati, S.Si., M.Si selaku Dosen Pembimbing I dan Bapak Surya Prangga, S.Si., M.Si selaku Dosen Pembimbing II yang dengan sangat sabar memberikan bimbingan, masukan, nasehat, dan motivasi dari awal sampai akhir penulisan skripsi.
3. Bapak Fidia Deny Tisna Amijaya, S.Si, M.Si selaku Dosen Penguji I dan Ibu Meiliyani Siringoringo, S.Si, M.Si selaku Dosen Penguji II yang telah banyak memberikan saran dan arahan dalam penyelesaian skripsi ini.
4. Kedua orang tua Bapak Zainal Aripin dan Lili Herlina serta adik Indriani Ramadhan, Fitri Handayani, Wafa Salsabila dan M Hafidz Najwan yang senantiasa memberikan do'a, semangat, dukungan moril dan materil serta kasih sayang yang tiada henti kepada penulis.
5. Sahabat seperjuangan Aprianti Boma Padatuan, Galuh Batul Nabilah, Widya Suerni, Pratiwi Dwi Yuliasari, Sarah Pandan Arum, Ulfah Resti Inayah, Nana Nirwana, Amanah, Riska Veronika, Febriana Syafitri, Era Tri Cahyani, Novia Felysia, Fatmawati serta seluruh Mahasiswa Angkatan

2016 yang telah memberikan dukungan selama proses pembuatan skripsi ini.

6. Kepada Tobi Agustian yang telah menemani, membantu dan terus memberikan motivasi sehingga selesainya skripsi ini.

Penulis sangat menyadari bahwa dalam penulisan skripsi ini masih banyak kekurangan dan jauh dari kesempurnaan. Oleh karena itu kritik dan saran yang mengarah pada kebaikan dan kesempurnaan skripsi ini sangat diharapkan.

Akhirnya dengan segala hormat dan kerendahan hati penulis berharap semoga penulisan skripsi ini dapat menambah wawasan dan pengetahuan serta membawa manfaat bagi para pembaca.

Samarinda, Maret 2020

Penulis

DAFTAR ISI

Halaman

HALAMAN JUDUL ..... **i**

HALAMAN PENGESAHAN ..... **ii**

PERNYATAAN KEASLIAN SKRIPSI ..... **iii**

ABSTRAK ..... **iv**

ABSTRACT ..... **v**

KATA PENGANTAR ..... **vi**

DAFTAR ISI ..... **viii**

DAFTAR TABEL…… ..... **x**

DAFTAR GAMBAR ..... **xii**

DAFTAR SIMBOL ..... **xiv**

DAFTAR LAMPIRAN ..... **xvi**

BAB 1 PENDAHULUAN **1**

1.1 Latar Belakang 1

1.2 Batasan Masalah 4

1.3 Rumusan Masalah 4

1.4 Tujuan Penelitian 5

1.5 Manfaat Penelitian…….. 5

BAB 2 TINJAUAN PUSTAKA **6**

2.1 *Data Mining* 6

2.2 *Decision Tree* 9

2.3 Algoritma C5.0 14

2.4 *Classification and Regression Tree* (CART) 15

2.5 Data *Training* dan *Testing* 18

2.6 *Confusion Matrix.* 19

2.7 Pendapatan 20

BAB 3 METODOLOGI PENELITIAN **22**

3.1 Waktu dan Tempat Penelitian 22

3.2 Rancangan Penelitian 22

3.3 Populasi dan Sampel 22
3.4 Teknik *Sampling*. 22
3.5 Teknik Pengumpulan Data 23
3.6 Variabel Penelitian 23
3.7 Teknik Analisis Data 24
3.8 Kerangka Penelitian 28
BAB 4 HASIL DAN PEMBAHASAN 29
4.1 Analisis Statistika Deskriptif 29
4.2 Pembagian Data *Training* dan *Testing* 32
4.3 Algoritma C5.0 32
4.4 *Classification and Regression Tree*. 49
4.4.1 Pembentukan Pohon Klasifikasi 49
4.4.1.1 Pemilihan Pemilah 50
4.4.1.2 Penentuan *Node* Terminal 73
4.4.1.3 Penandaan Label Kelas 73
4.5 Mengukur Ketepatan Hasil Klasifikasi 76
4.5.1 Ketepatan Klasifikasi Algoritma C5.0 77
4.5.2 Ketepatan Klasifikasi Algoritma CART 77
4.5.3 Perbandingan Tingkat Akurasi Algoritma C5.0 dan CART 78
BAB 5 PENUTUP 79
5.1 Kesimpulan 79
5.2 Saran 79
DAFTAR PUSTAKA 80

DAFTAR TABEL

Halaman

Tabel 2.1 *Confusion Matrix* 19
Tabel 4.1 Tabulasi Pekerjaan Terhadap Rata-rata Pendapatan Perbulan.... 30
Tabel 4.2 Tabulasi Jumlah Anggota Keluarga Terhadap Rata-rata ..... Pendapatan Perbulan..... 30
Tabel 4.3 Tabulasi Pendidikan Terakhir Terhadap Rata-rata Pendapatan.. Perbulan 31
Tabel 4.4 Tabulasi Jenis Kelamin Terhadap Rata-rata Pendapatan ..... Perbulan 31
Tabel 4.5 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* ... Akar 34
Tabel 4.6 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 2 36
Tabel 4.7 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 3 37
Tabel 4.8 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 4 39
Tabel 4.9 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 7 40
Tabel 4.10 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 10 41
Tabel 4.11 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 12 43
Tabel 4.12 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 14 44
Tabel 4.13 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 13 45
Tabel 4.14 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 19 46
Tabel 4.15 Pemilahan Pekerjaan Kemungkinan Pertama 51
Tabel 4.16 Pemilahan Pekerjaan Kemungkinan Kedua 51
Tabel 4.17 Pemilahan Pekerjaan Kemungkinan Ketiga..... 52
Tabel 4.18 Pemilahan Pekerjaan Kemungkinan Keempat..... 52
Tabel 4.19 Pemilahan Pekerjaan Kemungkinan Kelima..... 53
Tabel 4.20 Pemilahan Pekerjaan Kemungkinan Keenam 53
Tabel 4.21 Pemilahan Pekerjaan Kemungkinan Ketujuh..... 54
Tabel 4.22 Pemilahan Pekerjaan Kemungkinan Kedelapan 54
Tabel 4.23 Pemilahan Pekerjaan Kemungkinan Kesembilan 55

Tabel 4.24 Pemilahan Pekerjaan Kemungkinan Kesepuluh 55
Tabel 4.25 Pemilahan Pekerjaan Kemungkinan Kesebelas 56
Tabel 4.26 Pemilahan Pekerjaan Kemungkinan Kedua Belas 56
Tabel 4.27 Pemilahan Pekerjaan Kemungkinan Ketiga Belas 57
Tabel 4.28 Pemilahan Pekerjaan Kemungkinan Keempat Belas 57
Tabel 4.29 Pemilahan Pekerjaan Kemungkinan Kelima Belas..... 58
Tabel 4.30 Pemilahan Jumlah Anggota Keluarga Kemungkinan Pertama ... 58
Tabel 4.31 Pemilahan Pendidikan Terakhir Kemungkinan Pertama 59
Tabel 4.32 Pemilahan Pendidikan Terakhir Kemungkinan Kedua 60
Tabel 4.33 Pemilahan Pendidikan Terakhir Kemungkinan Ketiga..... 60
Tabel 4.34 Pemilahan Jenis Kelamin Kemungkinan Pertama 61
Tabel 4.35 Hasil Perhitungan Indeks *Gini* untuk *Node* 1..... 62
Tabel 4.36 Hasil Perhitungan Indeks *Gini* untuk *Node* 2..... 63
Tabel 4.37 Hasil Perhitungan Indeks *Gini* untuk *Node* 3..... 65
Tabel 4.38 Hasil Perhitungan Indeks *Gini* untuk *Node* 4..... 66
Tabel 4.39 Hasil Perhitungan Indeks *Gini* untuk *Node* 7..... 68
Tabel 4.40 Hasil Perhitungan Indeks *Gini* untuk *Node* 8..... 69
Tabel 4.41 Hasil Perhitungan Indeks *Gini* untuk *Node* 9..... 70
Tabel 4.42 Hasil Perhitungan Indeks *Gini* untuk *Node* 12..... 71
Tabel 4.43 Hasil Perhitungan Indeks *Gini* untuk *Node* 14..... 72
Tabel 4.44 Hasil Perhitungan Indeks *Gini* untuk *Node* 18..... 73
Tabel 4.45 Perhitungan Peluang Tiap Kelas dalam *Node* Terminal. 74
Tabel 4.46 Ketepatan Klasifikasi Algoritma C5.0 77
Tabel 4.47 Ketepatan Klasifikasi Algoritma CART 77
Tabel 4.48 Perbandingan Tingkat Akurasi Kedua Metode untuk Setiap.....
Proporsi. 78

DAFTAR GAMBAR

Halaman

Gambar 2.1 Syarat Pengujian Fitur Biner 10
Gambar 2.2 Syarat Pengujian Fitur Bertipe Nominal 11
Gambar 2.3 Syarat Pengujian Fitur Bertipe Ordnial 11
Gambar 2.4 Syarat Pengujian Fitur Bertipe Numerik 12
Gambar 2.5 Struktur Pohon Klasifikasi *Decision Tree* 13
Gambar 3.1 Diargam Alir Algoritma C5.0..... 26
Gambar 3.2 Diargam Alir Algoritma CART..... 27
Gambar 3.3 Kerangka Penelitian..... 28
Gambar 4.1 Diagram Batang unutk Variabel Rata-rata Pendapatan.....
Perbulan 29
Gambar 4.2 Hasil Pembentukan Cabang di *Node* Akar 35
Gambar 4.3 Hasil Pembentukan Cabang di *Node* 2 37
Gambar 4.4 Hasil Pembentukan Cabang di *Node* 3 38
Gambar 4.5 Hasil Pembentukan Cabang di *Node* 4 40
Gambar 4.6 Hasil Pembentukan Cabang di *Node* 7 41
Gambar 4.7 Hasil Pembentukan Cabang di *Node* 10 42
Gambar 4.8 Hasil Pembentukan Cabang di *Node* 12 44
Gambar 4.9 Hasil Pembentukan Cabang di *Node* 16 46
Gambar 4.10 Hasil Pembentukan Cabang di *Node* 19 47
Gambar 4.11 Pohon Klasifikasi Algoritma C5.0 48
Gambar 4.12 Hasil Pembentukan Cabang di *Node* Akar 63
Gambar 4.13 Hasil Pembentukan Cabang di *Node* 2 65
Gambar 4.14 Hasil Pembentukan Cabang di *Node* 3 66
Gambar 4.15 Hasil Pembentukan Cabang di *Node* 4 67
Gambar 4.16 Hasil Pembentukan Cabang di *Node* 7 68
Gambar 4.17 Hasil Pembentukan Cabang di *Node* 8 69
Gambar 4.18 Hasil Pembentukan Cabang di *Node* 9 70
Gambar 4.19 Hasil Pembentukan Cabang di *Node* 12 71

Gambar 4.20 Hasil Pembentukan Cabang di *Node* 14 72
Gambar 4.21 Hasil Pembentukan Cabang di *Node* 18 73
Gambar 4.22 Pohon Klasifikasi CART. 75

DAFTAR SIMBOL

Simbol	Arti
A	Variabel
b	Banyaknya data pada suatu variabel
b_1	Banyaknya data pada subset D_1
b_2	Banyaknya data pada subset D_2
FN	Jumlah baris berlabel C1 pada *test set* namun diklasifikasikan sebagai bukan kelas C1 oleh *classifier*
FP	Jumlah baris berlabel kelas bukan C1 pada *test set* namun diklasifikasikan sebagai kelas C1 oleh *classifier*
$Gain(S, A)$	Nilai *gain* dari suatu variabel
$Gini_{pembelahan}$	Nilai indeks *gini* setiap variabel
$gini(D_1)$	Nilai indeks *gini* subset D_1 pada setiap variabel
$gini(D_2)$	Nilai indeks *gini* subset D_2 pada setiap variabel
j,k	Kelas
L	Banyaknya kategori pada suatu variabel
m	Jumlah kategori pada variabel A
$m(t)$	Jumlah pengamatan pada *node t*
$m_j(t)$	Jumlah pengamatan pada kelas j pada *node t*
N	Jumlah data yang akan digunakan sebagai sampel
p_i	Proporsi dari S_i dan S
$P(j\|t)$	Probabilitas bersyarat kelas j yang berada dalam *node t*
$P(k\|t)$	Probabilitas bersyarat kelas k yang berada dalam *node t*

S	Himpunan kasus
S_i	Himpunan kasus pada kategori ke-*i*
$\lvert S_i \rvert$	Jumlah kasus pada kategori ke-*i*
$\lvert S \rvert$	Jumlah kasus dalam *S*
TN	Jumlah baris berlabel kelas bukan C1 pada *test set* dan benar diklasifikasikan sebagai bukan kelas C1 oleh *classifier*
TP	Jumlah baris berlabel kelas C1 pada *test set* yang benar diklasifikasikan sebagai kelas C1 oleh *classifier*

DAFTAR LAMPIRAN

Halaman

Lampiran 1 Kuesioner *Social Mapping* 83

Lampiran 2 Data Sosial Kepala Keluarga (KK) di Desa Teluk Baru .....
Tahun 2019 84

Lampiran 3 Data *Testing* untuk Proporsi Data 90:10..... 88

Lampiran 4 Sintaks *Software R* untuk membuat diagram lingkaran..... 89

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Data adalah segala fakta, angka, atau teks yang dapat diproses oleh komputer. Salah satu bentuk pengolahan suatu data yaitu *data mining*. *Data mining* adalah proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam mengambil sebuah keputusan. Munculnya *data mining* didasarkan pada kenyataan bahwa jumlah data yang tersimpan dalam basis data semakin besar (Prasetyo, 2012).

Banyak fungsi yang dapat dilakukan menggunakan *data mining*, diantaranya adalah deskripsi, estimasi, prediksi, klasifikasi, pengklusteran, dan asosiasi. Klasifikasi adalah suatu metode yang digunakan untuk memberikan label dari suatu data/objek baru. Misal kita memiliki data kelayakan seorang pelanggan untuk meminjam uang di bank (keputusan: layak atau tidak layak) berdasarkan nilai *asset*, usia pelanggan, jumlah kredit dan lain-lain (Pramana dkk, 2018). Ada beberapa macam pengklasifikasian dalam *data mining* yaitu *decision tree*, *naive* Bayes, *Support Vector Machine* (SVM) dan lain-lain (Larose, 2005).

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 (Prasetyo, 2014).

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 (Kusrini dan Luthfi, 2009).

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) (Breiman dkk dalam Pratiwi dan Zain (2014)).

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

Tingkat pendapatan merupakan salah satu kriteria maju tidaknya suatu daerah. Bila pendapatan suatu daerah relatif rendah, dapat dikatakan bahwa kemajuan dan kesejahteraan daerah tersebut akan rendah pula. Begitu pula jika pendapatan masyarakat suatu daerah relatif tinggi, maka tingkat kesejahteraan dan kemajuan daerah tersebut akan tinggi juga (Soekartawi, 2002).

Tinggi rendahnya pengeluaran sangat bergantung kepada kemampuan keluarga dalam mengelola penerimaan atau pendapatannya. Selain itu pengalaman dalam bekerja juga mempengaruhi pendapatan. Semakin baik pengalaman bekerja seseorang maka semakin tinggi peluangnya dalam meningkatkan pendapatan.

Usaha meningkatkan pendapatan masyarakat dapat dilakukan dengan melakukan pemberantasan kemiskinan yaitu membina kelompok masyarakat dengan pemenuhan modal kerja, ketepatan dalam penggunaan modal kerja diharapkan dapat memberikan kontribusi terhadap pengembangan usaha sesuai dengan yang diharapkan sehingga upaya peningkatan pendapatan masyarakat dapat terwujud dengan optimal (Sudarman, 2001).

Penelitian terdahulu mengenai metode Algoritma C5.0 dan CART telah dilakukan antara lain oleh Wijaya dkk (2018) yang telah melakukan penelitian tentang implementasi algoritma C5.0 dalam klasifikasi pendapatan masyarakat di kelurahan mesjid kecamatan medan kota dengan menggunakan variabel pendapatan sebagai variabel terikat kemudian variabel umur, pendidikan, kepemilikan rumah dan jumlah anggota keluarga sebagai variabel bebas. Setelah dilakukan pengklasifikasian didapatkan hasil berupa pohon klasifikasi untuk penentuan layak atau tidak layak seseorang mendapatkan Bantuan Langsung Tunai (BLT). Sementara itu Pakpahan dkk (2018) telah melakukan penelitian tentang Penerapan algoritma CART *Decision Tree* pada penentuan penerima program bantuan pemerintah daerah kabupaten kutai kartanegara dengan menggunakan variabel status (keterangan untuk calon penerima bantuan diterima atau tidak) sebagai variabel terikat kemudian variabel umur, jenis kelamin, status perkawinan, pendidikan terakhir, pekerjaan dan keterampilan sebagai variabel bebas. Setelah dilakukan pengklasifikasian didapatkan hasil berupa pohon klasifikasi yang dapat membantu dalam pengambilan keputusan untuk seorang lansia apakah diterima atau tidak untuk menjadi penerima bantuan dengan tingkat akurasi hasil klasifikasi sebesar 98,18%. Kemudian Yusuf (2007) telah melakukan penelitian tentang perbandingan performasi algoritma *decision tree* C5.0, CART dan CHAID dengan studi kasus kasus prediksi status resiko kredit di Bank X menggunakan variabel keputusan berupa status resiko kredit kemudian variabel jenis kelamin, umur, jumlah kredit, lama pinjaman dan nilai jaminan sebagai variabel bebas. Setelah dilakukan pengklasifikasian didapatkan hasil bahwa rata-rata tingkat akurasi untuk metode algoritma C5.0 sebesar 87,72%, CART sebesar 87,27% dan CHAID sebesar 87,15%.

Berdasarkan latar belakang tersebut, maka penulis tertarik untuk membuat penelitian ilmiah dengan judul "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)".

1.2 Batasan Masalah

Berdasarkan uraian dari latar belakang, maka batasan masalah pada penelitian ini adalah :

1. Menggunakan berbagai proporsi data *training* dan data *testing.*
2. Mengetahui ketepatan hasil klasifikasi dan metode yang lebih akurat dalam melakukan pengklasifikasian dengan menggunakan bantuan *confusion matrix*.

1.3 Rumusan Masalah

Berdasarkan uraian dari latar belakang, maka rumusan masalah dalam penelitian ini adalah :

1. Bagaimana hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma C5.0?
2. Bagaimana hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma CART?
3. Bagaimana perbandingan ketepatan hasil klasifikasi metode algoritma C5.0 dengan metode algoritma CART menggunakan *confusion matrix*?

1.4 Tujuan Penelitian

Tujuan penelitian yang ingin dicapai dalam penelitian ini adalah :

1. Untuk memperoleh hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma C5.0.

2. Untuk memperoleh hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong menggunakan metode algoritma CART.
3. Untuk memperoleh perbandingan ketepatan hasil klasifikasi metode algoritma C5.0 dengan metode algoritma CART menggunakan *confusion matrix.*

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah :

1. Memberikan gambaran dan pengetahuan tentang metode algoritma C5.0 dan algoritma CART.
2. Hasil dari penelitian ini diharapkan dapat memberi masukkan untuk Kecamatan Muara Ancalong mengenai kondisi di salah satu desa yaitu Desa Teluk Baru.
3. Dapat dijadikan referensi bagi peneliti yang sedang meneliti di bidang ini.

BAB 2
TINJAUAN PUSTAKA

2.1 *Data Mining*

Menurut Nafisah dalam Mardiani (2012), kebutuhan manusia akan data dan informasi tidak dapat dipungkiri. Bahkan, sekarang melalui dunia teknologi, arus informasi dapat beredar dengan cepat dan mudah. Data perlu diorganisasikan dan dikontrol menjadi sebuah informasi agar lebih mudah dipahami. Pengolahan data menjadi informasi tersebut perlu dilakukan secara hati-hati agar informasi yang menghasilkan memiliki kualitas yang baik. Seiring dengan hal ini, maka algoritma *data mining* juga turut berkembang pada basis data yang besar. *Data mining* adalah suatu teknologi yang berguna untuk mengekstrak pengetahuan atau yang dikenal sebagai informasi dari kumpulan data, sehingga hasilnya bisa dipergunakan untuk pengambilan keputusan.

Sedangkan menurut Pramana dkk (2018), *data mining* secara umum adalah kegiatan pencarian (*discovery*) secara berulang (*iterative*) dan intensif yang bertujuan untuk mengekstrak pengetahuan dari sekumpulan data yang awalnya tidak/belum memiliki arti yang penting. Pengetahuan yang dimaksud dapat berupa *pattern*/pola, hubungan, perubahan, *anomaly*, ataupun model yang muncul dari data. Hasil yang didapatkan harus valid, berguna dan mudah dimengerti.

Menurut Gonunescu dalam Prasetyo (2014), secara sistematis ada tiga langkah utama dalam *data mining* adalah sebagai berikut:

1. Eksplorasi/Pemrosesan Awal Data

   Eksplorasi/pemrosesan awal data terdiri dari ‘pembersihan’ data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

2. Membangun Model dan Melakukan Validasi Terhadapnya

   Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis *cluster*, deteksi anomali, deteksi asosiasi,

analisis pola sekuensal, dan sebagainya. Dalam beberapa referensi, deteksi anomali juga masuk dalam langkah eksplorasi. Akan tetapi, deteksi anomali juga dapat digunakan sebagai algoritma utama, terutama untuk mencari data-data yang spesial.

3. Penerapan

Penerapan berarti menerpakan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang akan diinvestigasi.

Menurut Kusrini dan Luthfi dalam Mardiani (2012), pengelompokkan *data mining* dibagi menjadi beberapa kelompok sebagai berikut :

a. Deskripsi, yang merupakan cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data yang dimiliki.
b. Estimasi, yang hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model yang dibangun menggunakan *record* lengkap dengan menyediakan nilai dari variabel target sebagai nilai prediksi.
c. Prediksi, prediksi menerka sebuah nilai yang belum diketahui dan juga memperkirakan nilai untuk masa mendatang.
d. Klasifikasi, terdapat target variabel kategori, misal penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang dan rendah.
e. Pengklusteran, yang merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.
f. Asosiasi, yang bertugas menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut dengan analisis keranjang belanja.

Sedangkan menurut Prasetyo (2014), pekerjaan yang berkaitan dengan *data mining* dapat dibagi menjadi empat kelompok sebagai berikut :

1. Model Prediksi (*Prediction Modelling*)

Pekerjaan ini berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada

himpunan baru yang didapat. Ada 2 jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan untuk variabel target diskrit, sedangkan regresi digunakan untuk variabel target kontinu.

Contoh pekerjaan yang menggunakan jenis klasifikasi adalah melakukan deteksi jenis penyakit pasien berdasarkan sejumlah nilai-nilai parameter penyakit yang diderita masuk. Pekerjaan ini termasuk jenis klasifikasi karena target yang diharapkan adalah diskrit, hanya beberapa jenis kemungkinan nilai target yang didapatkan dan tidak ada nilai seri waktu (*time series*) yang harus didapatkan untuk mendapat target nilai akhir. Sementara melakukan prediksi jumlah penjualan yang didapatkan pada 3 bulan ke depan itu termasuk regresi karena untuk mendapatkan nilai penjualan bulan ketiga harus mendapatkan nilai penjualan bulan kedua dan untuk mendapatkan nilai penjualan kedua harus mendapatkan nilai penjualan bulan pertama. Dalam hal ini ada nilai seri waktu yang harus dihitung untuk sampai pada target akhir yang diinginkan dan ada nilai kontinu yang harus dihitung untuk mendapatkan nilai target akhir yang diinginkan.

2. Analisis *Cluster* (*Cluster Analysis*)

Contoh pekerjaan yang berkaitan dengan analisis *cluster* adalah bagaimana bisa mengetahui pola pembelian barang oleh konsumen pada waktu-waktu tertentu. Dengan mengetahui pola kelompok pembelian tersebut, maka perusahaan/*retailer* dapat menentukan jadwal promosi yang dapat diberikan sehingga dapat membantu meningkatkan omset penjualan.
Analisis kelompok melakukan pengelompokkan data ke dalam sejumlah kelompok berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, dan akan terpisah dalam kelompok yang berbeda jika keluar dari batas kesamaan kelompok tersebut.

3. Analisis Asosiasi (*Asociaton Analysis*)

Analisis asosiasi digunakan untuk menentukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Pola yang ditemukan biasanya merepresentasikan bentuk aturan implikasi atau subset fitur. Tujuannya adalah untuk menemukan pola yang menarik dengan cara yang efisien.

Penerapan yang paling dekat dengan kehidupan sehari-hari adalah analisis data keranjang belanja. Jika ibu rumah tangga akan membeli barang kebutuhan rumah tangga (misalnya beras) di sebuah supermarket, maka sangat besar kemungkinan ibu rumah tangga tersebut juga akan membeli kebutuhan rumah tangga yang lain, misalnya minyak dan telur, dan tidak mungkin (atau jarang) membeli barang lain seperti topi atau buku. Dengan mengetahui hubungan yang lebih kuat antara beras dan telur dibandingkan beras dengan topi, maka *retailer* dapat menentukan barang-barang yang sebaiknya disediakan dalam jumlah yang cukup banyak.

4. Deteksi Anomali (*Anomaly Detection*)

Pekerjaan deteksi anomali berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain. Data-data yang karakteristiknya menyimpang (berbeda) dari data yang lain disebut dengan sebagai *outlier*. Algoritma deteksi anomali yang baik harus mempunyai laju deteksi yang tinggi dan laju kesalahan yang rendah. Deteksi anomali dapat diterapkan pada sistem jaringan untuk mengetahui pola data yang memasuki jaringan sehingga dapat diketahui adanya penyusupan jika pola kerja data yang datang berbeda ataupun perilaku kondisi cuaca yang mengalami anomali juga dapat dideteksi dengan algoritma ini.

2.2 *Decision Tree*

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 (Prasetyo, 2014).

Menurut Prasetyo (2014), *decision tree* mempunyai tiga pendekatan klasik:

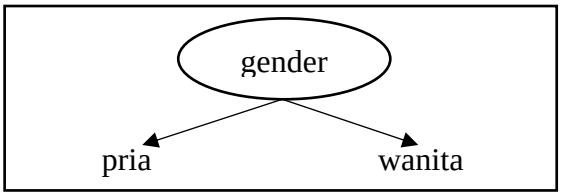
1. Pohon klasifikasi, digunakan untuk melakukan prediksi ketika ada data baru yang belum diketahui label kelasnya. Pendekatan ini yang paling banyak digunakan.
2. Pohon regresi, ketika hasil prediksi dianggap sebagai nilai nyata yang mungkin akan didapatkan. Misalnya kasus kenaikan harga rumah, prediksi inflasi tiap tahun, dan sebagainya.
3. CART (atau C&RT), ketika ada masalah klasifikasi dan regresi digunakan bersama-sama.

Dalam *decision tree,* daerah pengambilan keputusan yang sebelumnya kompleks dapat diubah menjadi lebih sederhana. Banyak algoritma yang dapat dipakai dalam pembentukan *decision tree* yaitu ID3, CART, C4.5, dan lain-lain. Algoritma adalah urutan langkah-langkah yang logis untuk menyelesaikan suatu masalah (Prasetyo, 2012).

Menurut Prasetyo (2014), terdapat hal penting dalam induksi *decision tree* yaitu bagaimana menyatakan syarat pengujian pada *node*. Ada 3 kelompok penting dalam syarat pengujian *node*:

1. Fitur Biner

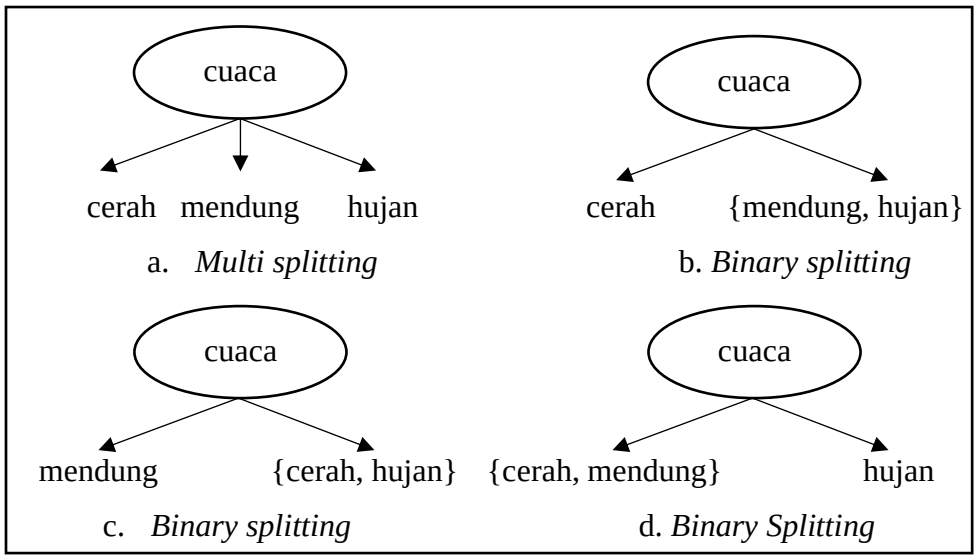
   Fitur yang hanya mempunyai dua nilai berbeda disebut dengan fitur biner. Syarat pengujian ketika fitur ini menjadi *node* (akar maupun internal) hanya punya dua pilihan cabang. Contoh pemecahannya dapat dilihat pada Gambar 2.1.



Gambar 2.1 Syarat pengujian fitur biner

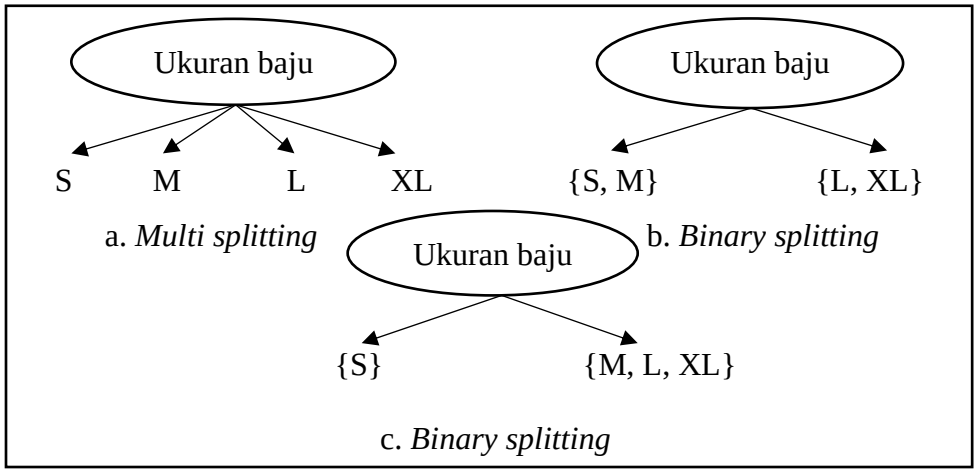
2. Fitur Bertipe Kategorikal

Fitur yang nilainya bertipe kategorikal (nominal atau ordinal) bisa mempunyai beberapa nilai berbeda. Contohnya adalah fitur ‘cuaca’ mempunyai 3 nilai berbeda, dan ini bisa mempunyai banyak kombinasi syarat pengujian pemecahan. Secara umum ada 2 pemecahan, yaitu pemecahan biner (*binary splitting*) dan (*multi splitting*) seperti pada Gambar 2.2 dan 2.3.



Gambar 2.2 Syarat Pengujian Fitur Bertipe Nominal

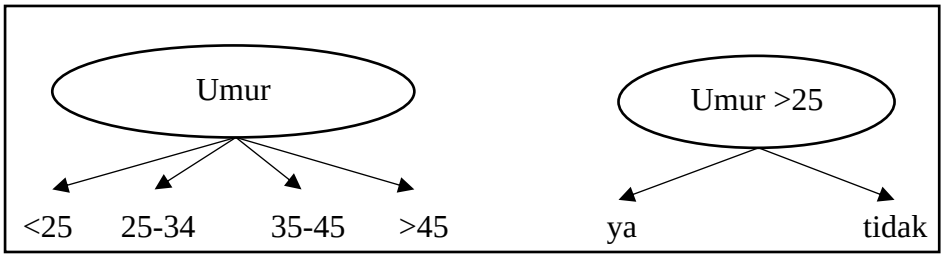
Berikut contoh fitur kategorikal ordinal ‘ukuran baju’ yang mempunyai 4 nilai berbeda:



Gambar 2.3 Syarat Pengujian Fitur Bertipe Ordinal

3. Fitur Bertipe Numerik

Fitur bertipe numerik, syarat pengujian dalam *node* (akar maupun internal). Untuk kasus pemecahan biner, maka algoritma akan memeriksa semua kemungkinan posisi pemecahan yang terbaik. Untuk cara multi, maka algoritma harus memeriksa semua kemungkinan jangkauan nilai kontinyu. Contoh pemecahan pada fitur numerik dapat dilihat pada Gambar 2.4.



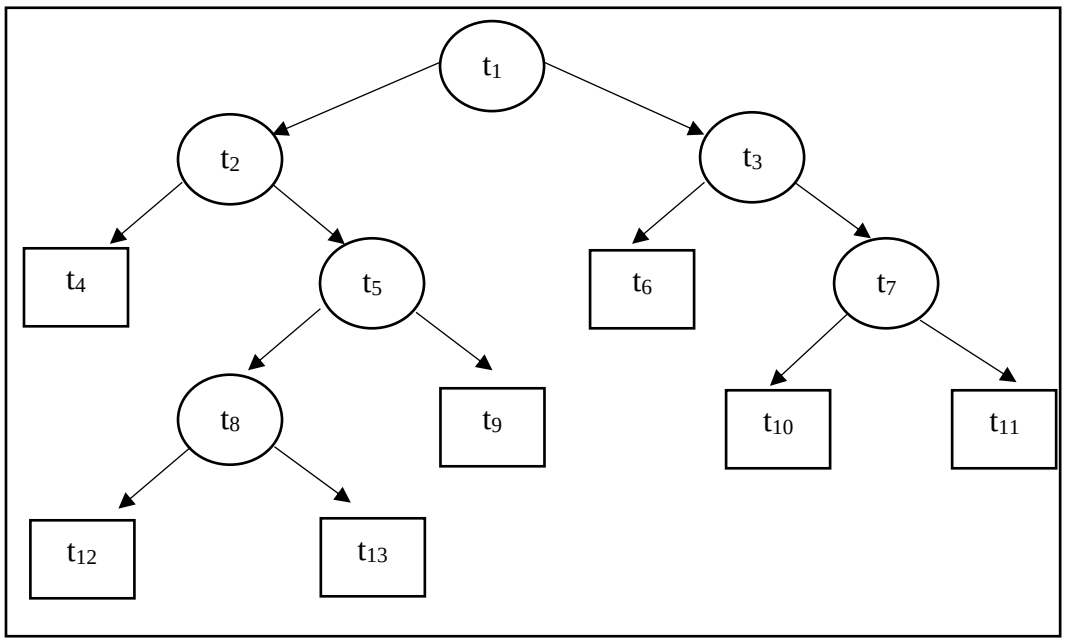
a. *Multi splitting* b. *Binary splitting*

Gambar 2.4 Syarat pengujian fitur bertipe numerik

Karakteristik dari *decision tree* dibentuk dari sejumlah elemen sebagai berikut (Prasetyo, 2012):

1. *Node*, yang menyatakan variabel. *Node* bisa berupa variabel akar, variabel cabang, dan kelas.
2. *Arm*, setiap cabang menyatakan nilai hasil pengujian di *node* bukan daun.
3. *Node* akar, tidak mempunyai *input arm* yaitu lengan masukan dan mempunyai nol atau lebih *output arm* yaitu lengan keluar.
4. *Node* internal, setiap *node* yang bukan daun (*non terminal*) yang mempunyai tepat satu *input arm* dan dua atau lebih *output arm*, *node* ini menyatakan pengujian yang didasarkan pada nilai fitur.
5. *Node* daun (*terminal*) adalah *node* yang mempunyai tepat satu *input arm* dan tidak mempunyai *output arm*. *Node* ini menyatakan label kelas (keputusan).

Berikut merupakan ilustrasi dari bentuk pohon klasifikasi yang ditunjukkan pada Gambar 2.5:



Gambar 2.5 Struktur Pohon Klasifikasi *Decision Tree*

Simpul utama (*root node*) dinotasikan sebagai t_1, sedangkan simpul t_2, t_3, t_5, t_7 dan t_8 disebut simpul dalam (*internal nodes*). Simpul akhir yang juga disebut sebagai simpul terminal (*terminal nodes*) adalah t_4, t_6, t_9, t_{10}, t_{11}, t_{12} dan t_{13} dimana tidak terjadi lagi pemilahan. Kedalaman pohon (*depth*) dihitung dimulai dari simpul utama atau t_1 yang berada pada kedalaman 1, sedangkan t_2 dan t_3 berada pada kedalam 2. Begitu seterusnya sampai pada simpul terminal t_{12} dan t_{13} yang berada pada kedalaman 5 yang ditunjukkan pada Gambar 2.5 dengan *node* yang digambarkan ◯ merupakan *node* utama dan *node* dalam sedangkan *node* yang digambarkan ▭ merupakan *node* terminal (Breiman dkk dalam Pratiwi dan Zain (2014)).

Algoritma *decision tree* secara umum bekerja secara *top-down*, dengan cara memilih atribut yang merupakan *best predictor*/prediktor terbaik sebagai *root*. Kemudian atribut berikutnya yang merupakan *best splitting attribute* menjadi cabang dari *tree* yang terbentuk, demikian seterusnya hingga *dataset* telah terbagi habis atau atribut-atribut yang ada semuanya telah menjadi cabang dari *tree*. Dalam proses pembentukan *decision tree*, terdapat beberapa metrik untuk menentukan

atribut terbaik sebagai *best predictor* seperti *entropy & information gain*, *gain ratio*, dan *gini index* (Pramana dkk, 2018).

2.3 Algoritma C5.0

Algoritma C5.0 adalah salah satu algoritma *data mining* yang khususnya diterapkan pada algoritma *decision tree*. Algoritma C5.0 ini merupakan penyempurnaan algoritma sebelumnya yang dibentuk oleh Ross Quinlan pada tahun 1987, yaitu ID3 dan C4.5. Dalam algoritma ini pemilihan atribut diproses menggunakan *gain ratio*. Algoritma ini menghasilkan *tree* dengan jumlah cabang per *node* bervariasi (Dunham dalam Putri dkk (2013)).

Algoritma C5.0 menghasilkan *tree* dengan jumlah cabang per *node* bervariasi. algoritma ini memperlakukan variabel kontinyu sama dengan yang dilakukan oleh CART, tetapi untuk variabel kategorik algoritma C5.0 memperlakukan nilai variabel kategorikal sebagai *splitter*. Sampel *subset* yang diperoleh dari percabangan yang terbentuk akan dipecah lagi setelahnya. Prosesnya akan terus berlanjut sampai sampel *subset* tidak dapat lagi dibagi. Pada akhirnya, sampel *subset* yang tidak memiliki kontribusi yang besar bagi model akan ditolak (Larose dalam Yusuf (2007)).

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:

$$Entropy(S) = -\sum_{j=1}^{k} p_j \log_2 p_j \qquad (2.1)$$

dengan :

S : Himpunan kasus

k : Jumlah kelas pada variabel A

p_j : Proporsi dari S_j dan S

Selanjutnya untuk mencari nilai *gain* digunakan persamaan berikut :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (2.2)$$

dengan :

S : Himpunan kasus

S_i : Himpunan kasus pada kategori ke-*i*

A : Variabel

m : Jumlah kategori pada variabel *A*

$|S_i|$: Jumlah kasus pada kategori ke-*i*

$|S|$: 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 :

$$Gain\ Ratio = \frac{Gain(S, A)}{\sum_{i=1}^{m} Entropy(S_i)} \quad (2.3)$$

dengan :

$Gain(S, A)$: Nilai *gain* dari suatu variabel

$\sum_{i=1}^{m} Entropy(S_i)$: Jumlah nilai *entropy* dalam suatu variabel

(Kantardzic dalam Putri dkk (2013))

Proses diulang untuk masing-masing cabang sampai semua kelas pada cabang memiliki kelasnya masing-masing.

2.4 *Classification and Regression Tree* (CART)

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) (Breiman dkk dalam Pratiwi dan Zain (2014)).

CART mempunyai beberapa kelebihan dibandingkan metode lainnya, yaitu hasilnya lebih mudah diinterpretasikan, lebih akurat dan lebih cepat perhitungannya, selain itu CART bisa diterapkan untuk himpunan data yang mempunyai jumlah besar, variabel yang sangat banyak dengan skala variabel campuran melalui prosedur pemilahan biner. Data *training* digunakan untuk pembentukan pohon klasifikasi optimal sedangkan data *testing* digunakan untuk validasi model yaitu seberapa besar kemampuan model dalam memprediksi data baru.

Menurut Lewis dan Roger dalam Pratiwi dan Zain (2013), metode CART memiliki kelemahan sebagai berikut:

1. CART mungkin tidak stabil dalam *decision tree* (pohon keputusan) karena CART sangat sensitif dengan data baru.
2. CART sangat bergantung dengan jumlah sampel. Jika sampel data *learning* dan *testing* berubah maka pohon keputusan yang dihasilkan juga ikut berubah.
3. Tiap pemilahan bergantung pada nilai yang hanya berasal dari satu variabel penjelas.

Proses pembentukan pohon klasifikasi pada algoritma CART melalui tiga tahapan, yaitu:

a. Pemilihan Pemilah

Data yang digunakan merupakan sampel data *learning*. Himpunan bagian yang dihasilkan dari proses pemilahan harus lebih homogen dibandingkan pemilahan sebelumnya. Menurut Breiman dalam Akbar dkk (2010), rumus pemilah disajikan seperti berikut:

- Variabel bebas kontinu $= b-1$ pemilahan (2.4a)
- Variabel bebas kategori nominal $= 2^{L-1}-1$ pemilahan (2.4b)
- Variabel bebas kategori ordinal $= L-1$ pemilahan (2.4c)

dengan:

b : Banyaknya data pada suatu variabel

L : 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. Fungsi Indeks Gini ditunjukkan pada persamaan berikut:

$$j(t) = \sum_{j \neq k}^{m} P(j \mid t) P(k \mid t) \tag{2.5}$$

dengan:

j,k : Kelas

$P(j|t)$: Probabilitas bersyarat kelas *j* yang berada dalam *node t*

$P(k|t)$: Probabilitas bersyarat kelas *k* yang berada dalam *node t*

dengan *j(t)* adalah fungsi keheterogenan indeks *gini*, $P(j|t)$ adalah peluang *j* pada *node t*, dan $P(k|t)$ adalah peluang *k* pada *node t*.

Rumus indeks *gini* dapat dituliskan:

$$j(t) = 1 - \sum_{j=1}^{m} P^2(j \mid t) \tag{2.6}$$

Node t dibelah menjadi 2 subset D_1 dan D_2 dengan ukuran masing-masing b_1 dan b_2, indeks *gini* dari pembelahan tersebut didefinisikan sebagai berikut:

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2) \tag{2.7}$$

dengan :

$Gini_{pembelahan}$: Nilai indeks *gini* setiap variabel

$gini(D_1)$: Nilai indeks *gini* subset D_1 pada setiap variabel

$gini(D_2)$: Nilai indeks *gini* subset D_2 pada setiap variabel

b : Banyaknya data pada suatu variabel

b_1 : Banyaknya data pada subset D_1

b_2 : Banyaknya data pada subset D_2

b. Penentuan *Node* Terminal

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:

$$P(j_0 \mid t) = \max_j P(j \mid t) = \max_j \frac{m_j(t)}{m(t)} \tag{2.8}$$

dengan :

$P(j \mid t)$: Probabilitas bersyarat kelas *j* yang berada pada *node* t

$m_j(t)$: Jumlah pengamatan pada kelas *j* pada *node t*

$m(t)$: Jumlah pengamatan pada *node t*

Label kelas *node* terminal *t* adalah j_0 yang memberi nilai dugaan kesalahan pengklasifikasian *node t* terbesar.

(Breiman dkk dalam Pratiwi dan Zain (2013))

2.5 Data *Training* dan Data *Testing*

Menurut Prasetyo (2014), data yang akan digunakan dalam pengujian klasifikasi dibagi menjadi dua yaitu data *training* dan data *testing*. Data atau vektor yang sudah diketahui sebelumnya untuk label kelas dan digunakan untuk membangun model *classifier* disebut dengan data *training.* Data atau vektor yang belum diketahui (dianggap belum diketahui) label kelasnya menggunakan model *classifier* yang sudah dibangun disebut 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 2.9a dan 2.9b dengan N merupakan jumlah data yang akan digunakan sebagai sampel seperti berikut :

$$\text{Jumlah data } \textit{training} = \text{Proporsi data } \textit{training} \times N \quad (2.9a)$$

$$\text{Jumlah data } \textit{testing} = N - \text{Jumlah data } \textit{training} \quad (2.9b)$$

Pembagian data *training* dan data *testing* dengan metode proporsi adalah metode yang paling sederhana namun memiliki beberapa keterbatasan. Jumlah data *training* lebih sedikit yang tersedia untuk pelatihan karena sebagian harus digunakan untuk data *testing.* Akibatnya, model yang dibangun kemungkinan tidak sebagus ketika semua data digunakan sebagai data *training*. Model yang dibangun juga sangat tergantung pada komposisi pemecahan set data *training* dan data *testing* (Prasetyo, 2014).

2.6 *Confusion Matrix*

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. Tabel *confusion matrix* merupakan tabel *binary*, dimana kelas dibagi menjadi 2 label kelas saja. Jika terdapat *multiclass*, misalnya 3 kelas (A, B, C) maka *confusion matrix* dibagi menjadi 3 tabel yaitu antara kelas A dan bukan kelas A, antara kelas B dan bukan kelas B, dan antara kelas C dan bukan kelas C. Berikut adalah format umum dari *confusion matrix* :

Tabel 2.1 *Confusion Matrix*

Actual Class/Predicted Class	C1	-C1
C1	*True Positive* (TP)	*False Negative* (FN)
-C1	*False Positive* (FP)	*True Negative* (TN)

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* (Pramana dkk, 2018).

Menurut Pramana dkk (2018), akurasi adalah persentase baris *test set* yang diklasifikasikan dengan benar, berikut rumusnya :

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.10}$$

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 (Prasetyo, 2012).

2.7 Pendapatan

Pendapatan seseorang dapat didefinisikan sebagai banyaknya penerimaan yang dinilai dengan satuan mata uang yang dapat dihasilkan seseorang atau suatu bangsa dalam periode tertentu. Reksoprayitno (2004), mendefinisikan : "Pendapatan (*revenue*) dapat diartikan sebagai total penerimaan yang diperoleh pada periode tertentu".

Pendapatan masyarakat adalah penerimaan dari gaji atau balas jasa dari hasil usaha yang diperoleh individu atau kelompok rumah tangga dalam satu bulan dan digunakan untuk memenuhi kebutuhan sehari-hari. Sedangkan pendapatan dari usaha sampingan adalah pendapatan tambahan yang merupakan penerimaan lain dari luar aktifitas pokok atau pekerjaan pokok. Pendapatan sampingan yang diperoleh secara langsung dapat digunakan untuk menunjang atau menambah pendapatan pokok. Soekartawi (2002) menjelaskan pendapatan akan mempengaruhi banyak barang yang dikonsumsi, karena sering kali dijumpai bahwa

dengan bertambahnya pendapatan, maka barang yang dikonsumsi bukan saja bertambah, tapi juga kualitas barang tersebut ikut menjadi perhatian. Misalnya sebelum adanya penambahan pendapatan beras yang dikonsumsi adalah kualitas yang kurang baik, akan tetapi setelah adanya penambahan pendapatan maka konsumsi beras menjadi kualitas yang lebih baik.

Tingkat pendapatan merupakan salah satu kriteria maju tidaknya suatu daerah. Bila pendapatan suatu daerah relatif rendah, dapat dikatakan bahwa kemajuan dan kesejahteraan tersebut akan rendah pula. Kelebihan dari konsumsi maka akan disimpan pada bank yang tujuannya adalah untuk berjaga-jaga apabila baik kemajuan dibidang pendidikan, produksi dan sebagainya juga mempengaruhi tingkat tabungan masyarakat. Demikian pula apabila pendapatan masyarakat suatu daerah relatif tinggi, maka tingkat kesejahteraan dan kemajuan daerah tersebut tinggi pula (Soekartawi, 2002).

Sedangkan menurut Boediono (2002), 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.

Tingkat pendapatan mempengaruhi tingkat konsumsi masyarakat. Hubungan antara pendapatan dan konsumsi merupakan suatu hal yang sangat penting dari berbagai permasalahan ekonomi. Kenyataan menunjukkan bahwa pengeluaran konsumsi meningkat dengan naiknya pendapatan, dan sebaliknya jika pendapatan turun, pengeluaran konsumsi juga turun. Tinggi rendahnya pengeluaran sangat tergantung kepada kemampuan keluarga dalam mengelola penerimaan atau pendapatannya (Soekartawi, 2002).

BAB 3
METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada bulan Oktober tahun 2019 sampai Februari tahun 2020. Pengolahan data dilakukan di Laboratorium Statistika Komputasi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman, Jalan Barong Tongkok No. 4 Kampus Gunung Kelua, Samarinda.

3.2 Rancangan Penelitian

Penelitian yang dilakukan merupakan penelitian kuantitatif karena data yang dikumpulkan lebih mengambil bentuk yang dapat dihitung. Penelitian menggunakan rancangan kausal komparatif yang bersifat *ex post facto,* artinya data dikumpulkan setelah semua kejadian yang dipersoalkan berlangsung.

3.3 Populasi dan Sampel

Populasi dalam penelitian kali ini adalah data sosial seluruh Kepala Keluarga (KK) masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019, sedangkan yang menjadi sampel hanya 100 KK. Pengertian Kepala Keluarga (KK) dalam Kamus Besar Bahasa Indonesia (KBBI) adalah orang yang bertanggung jawab penuh terhadap anggota keluarganya (biasanya bapak), sedangkan menurut Sumiarni (2005) seorang KK tidak harus laki-laki tapi bisa juga berjenis kelamin perempuan.

3.4 Teknik *Sampling*

Pada penelitian kali ini teknik yang digunakan dalam pengambilan sampel adalah dengan menggunakan teknik *purposive sampling* atau teknik untuk menentukan sampel penelitian dengan beberapa pertimbangan tertentu yang bertujuan agar data yang diperoleh nantinya bisa lebih representatif (Sugiyono, 2010). Hal yang menjadi pertimbangan peneliti adalah kelengkapan data dalam kuisioner, apabila data tidak terisi penuh maka tidak menjadi anggota sampel.

3.5 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data primer. Pengambilan data dilakukan dengan menemui responden secara langsung dan mencatat data dengan kuisioner yang dapat dilihat pada Lampiran 1 di Desa Teluk Baru Kecamatan Muara Ancalong Kabupaten Kutai Timur mengenai data rata-rata pendapatan perbulan, pekerjaan, jumlah anggota keluarga, pendidikan terakhir, dan jenis kelamin.

3.6 Variabel Penelitian

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 Rosni (2012) 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 :

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

3.7 Teknik Analisis Data

Teknik analisis data yang digunakan pada penelitian kali ini adalah metode Algoritma C5.0 dan metode (CART) dalam mengklasifikasikan pendapatan rata-rata perbulan dengan menggunakan variabel bebas berupa pekerjaan, jumlah anggota keluarga, pendidikan terakhir, dan jenis kelamin. Adapun *software* komputer yang digunakan dalam penelitian kali ini adalah *Microsoft Excel* 2010 dan *Software R*.

Berikut langkah-langkah dalam penelitian kali ini adalah sebagai berikut :

1. Analisis Statistika Deskriptif

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna tentang suatu objek penelitian. Penyajian data dilakukan dengan membuat *cross tabulation* dan diagram lingkaran menggunakan bantuan dari *software R*.

2. Membagi Data *Training* (Data Latih) dan Data *Testing* (Data Uji)

Sebelum melakukan proses klasifikasi dengan menggunakan metode Algoritma C5.0 dan metode (CART), langkah pertama yang perlu dilakukan adalah membagi data *training* dan data *testing*, kemudian dilakukan pengacakan terlebih dahulu agar setiap data memiliki kesempatan yang sama untuk menjadi data *testing*

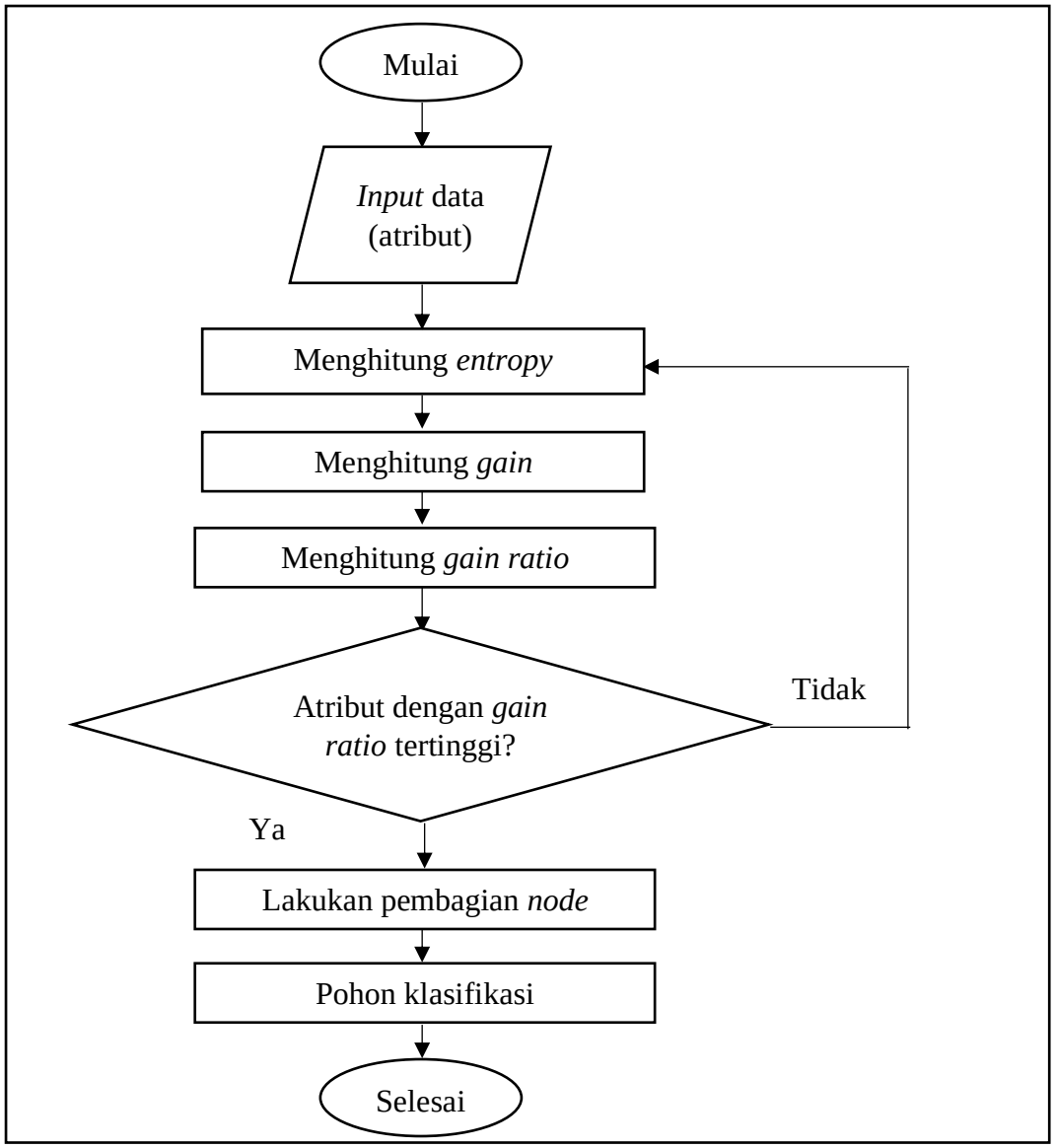
dan data *training*. Pengacakan data dilakukan dengan menggunakan bantuan *software Microsoft Excel* 2010.

3. Analisis Algoritma C5.0

Dalam analisis menggunakan Algoritma C5.0 ada beberapa langkah sebagai berikut :

a. Penentuan variabel yang akan diteliti.
b. Pemilihan *node* akar diawali dengan menghitung nilai *entropy* menggunakan persamaan (2.1). Kemudian proses dilanjutkan dengan mencari nilai *gain* menggunakan persamaan (2.2). Setelah itu mencari nilai *gain ratio* pada persamaan (2.3)
c. 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.
d. Kelas dibagi dalam cabang yang telah ditentukan.
e. Ulangi langkah c-d hingga semua kelas pada cabang memiliki kelasnya masing-masing.

Langkah analisis algoritma C5.0 untuk lebih jelasnya dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram Alir Algoritma C5.0

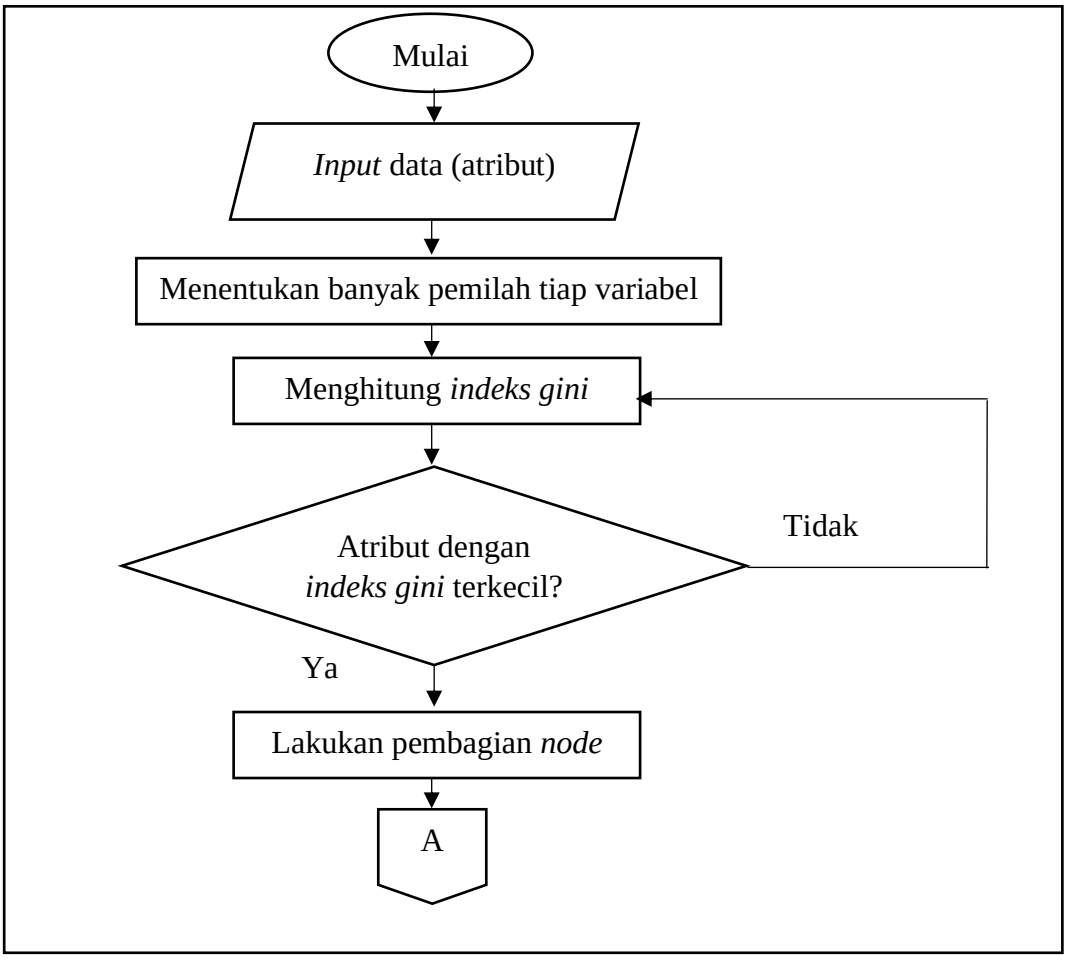
4. Analisis *Classification and Regression Tree* (CART)

Dalam analisis menggunakan Algoritma CART ada beberapa langkah sebagai berikut :

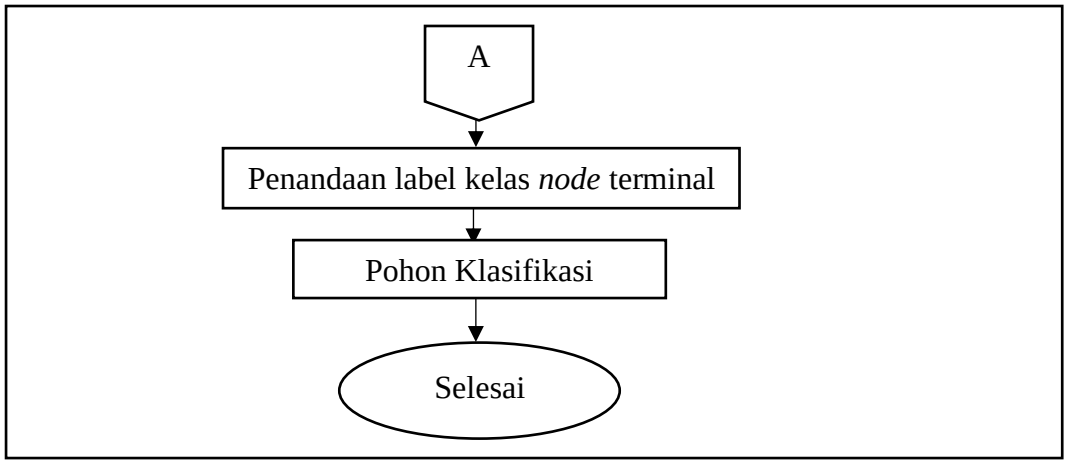
a. Penentuan variabel yang akan diteliti.
b. Penentuan banyaknya pemilah per variabel sesuai dengan jenis variabel bebasnya menggunakan persamaan (2.4a), (2.4b) dan (2.4c).

c. Menghitung nilai *indeks gini* untuk setiap pemilah sesuai dengan persamaan (2.7) kemudian pemilah yang memiliki nilai *indeks gini* terkecil akan dipilih menjadi pemilah terbaik.
d. Ulangi langkah b-c hingga tidak memungkinkan lagi untuk melakukan pemilahan.
e. Penandaan label kelas *node* terminal berdasarkan aturan jumlah anggota terbanyak menggunakan persamaan (2.8).

Langkah analisis algoritma CART untuk lebih jelasnya dapat dilihat pada Gambar 3.2.

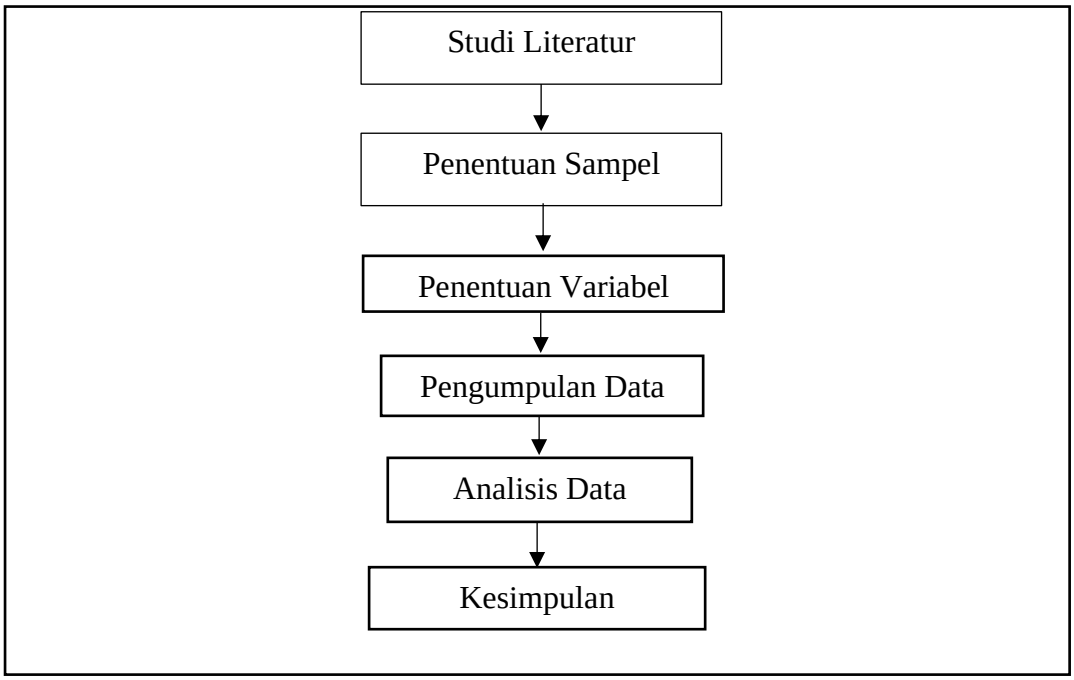


Gambar 3.2 Diagram Alir Algoritma CART



Gambar 3.2 Diagram Alir Algoritma CART (Lanjutan)

3.8 Kerangka Penelitian

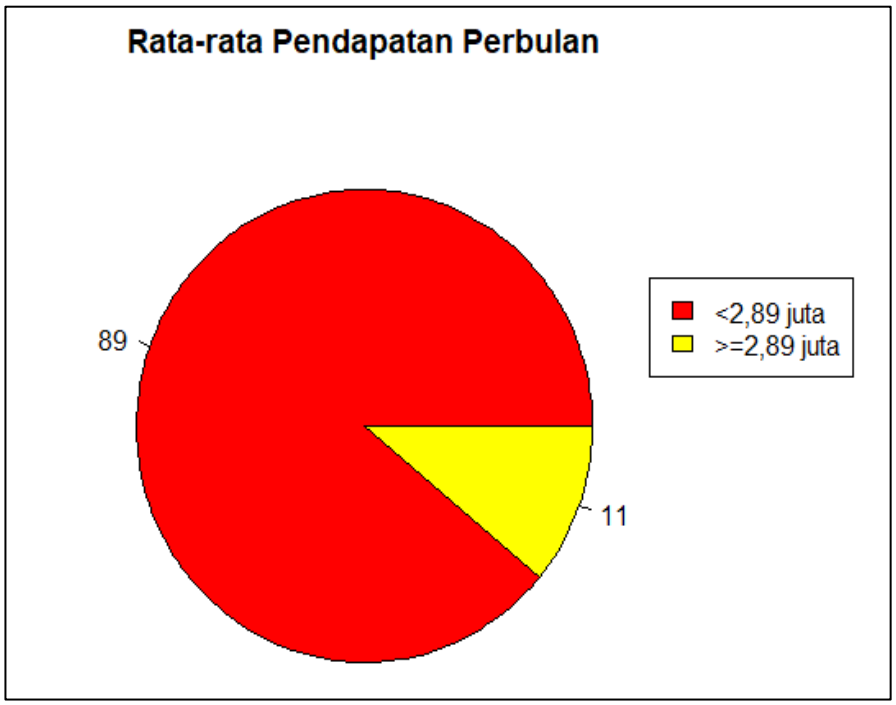


Gambar 3.3 Kerangka Penelitian

BAB 4
HASIL DAN PEMBAHASAN

4.1 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). Data lengkap dapat dilihat pada lampiran 2. Kemudian dengan bantuan *software R* menggunakan sintaks pada lampiran 4 diperoleh diagram lingkaran dari variabel rata-rata pendapatan perbulan sebagai berikut :



Gambar 4.1 Diagram Batang untuk Variabel Rata-rata Pendapatan Perbulan

Berdasarkan Gambar 4.1 dapat diketahui bahwa Kepala Keluarga (KK) yang memiliki rata-rata pendapatan < 2,89 juta ada 89 KK dan yang memiliki rata-rata pendapatan ≥2,89 juta ada 11 KK dari total sebanyak 100 KK di Desa Teluk Baru.

Berikut hasil analisis statistika deskriptif menggunakan *cross tabulation*:

Tabel 4.1 Tabulasi Pekerjaan Terhadap Rata-rata Pendapatan Perbulan

Rata-rata Pendapatan Perbulan	**Pekerjaan (X_1)**					**Jumlah**
	Petani	**Nelayan**	**PNS**	**Swasta**	**Wiraswasta**	
<2,89 juta	72 (72%)	0 (0%)	2 (2%)	10 (10%)	5 (5%)	89 (89%)
≥2,89 juta	5 (5%)	1 (1%)	1 (1%)	3 (3%)	1 (1%)	11 (11%)
Jumlah	77 (77%)	1 (1%)	3 (3%)	13 (13%)	6 (6%)	100 (100%)

Berdasarkan Tabel 4.1 mayoritas KK di desa Teluk Baru yang memiliki rata-rata pendapatan perbulan <2,89 juta bekerja sebagai petani yaitu sebanyak 72% dan tidak ada KK di desa Teluk Baru dengan rata-rata pendapatan < 2,89 juta yang bekerja sebagai nelayan. Kemudian pada rata-rata pendapatan perbulan ≥2,89 juta jumlah KK terbanyak berada pada pekerjaan Petani sebanyak 5% sedangkan yang terkecil pada pekerjaan Nelayan, PNS dan Wiraswasta sebanyak 1%.

Tabel 4.2 Tabulasi Jumlah Anggota Keluarga Terhadap Rata-Rata Pendapatan Perbulan

Rata-rata Pendapatan Perbulan	**Jumlah Anggota Keluarga (JAK) (X_2)**		**Jumlah**
	JAK >4	**JAK ≤4**	
<2,89 juta	35 (35%)	54 (54%)	89 (89%)
≥2,89 juta	5 (5%)	6 (6%)	11 (11%)
Jumlah	40 (40%)	60 (60%)	100 (100%)

Berdasarkan Tabel 4.2 mayoritas KK di desa Teluk Baru yang memiliki rata-rata pendapatan perbulan <2,89 juta dengan jumlah anggota keluarga ≤4 yaitu sebanyak 54% sedangkan yang terkecil yaitu pada jumlah anggota keluarga >4 sebanyak 35%. Kemudian pada rata-rata pendapatan perbulan ≥2,89 juta jumlah

KK terbanyak berada pada jumlah anggota keluarga ≤4 sebanyak 6% sedangkan yang terkecil pada jumlah anggota keluarga >4 sebanyak 5%.

Tabel 4.3 Tabulasi Pendidikan Terakhir Terhadap Rata-Rata Pendapatan Perbulan

Rata-rata Pendapatan Perbulan	**Pendidikan Terakhir (X_3)**				**Jumlah**
	SD	**SMP**	**SMA**	**PT**	
<2,89 juta	65 (65%)	19 (19%)	5 (5%)	0 (0%)	89 (89%)
≥2,89 juta	3 (3%)	2 (2%)	5 (5%)	1 (1%)	11 (11%)
Jumlah	68 (68%)	21 (21%)	10 (10%)	1 (1%)	100 (100%)

Berdasarkan Tabel 4.3 mayoritas KK di desa Teluk Baru yang memiliki rata-rata pendapatan perbulan <2,89 juta yaitu pada pendidikan terakhir SD sebanyak 65% dan tidak ada KK di desa Teluk Baru dengan rata-rata pendapatan <2,89 juta yang memiliki pendidikan terakhir PT. Kemudian pada rata-rata pendapatan perbulan ≥2,89 juta jumlah KK terbanyak berada pada pendidikan terakhir SMA sebanyak 5% sedangkan yang terkecil pada pendidikan terakhir PT sebanyak 1%.

Tabel 4.4 Tabulasi Jenis Kelamin Terhadap Pendapatan Rata-Rata Perbulan

Rata-rata Pendapatan Perbulan	**Jenis Kelamin (X_4)**		**Jumlah**
	Perempuan	**Laki-Laki**	
<2,89 juta	4 (4%)	85 (85%)	25 (25%)
≥2,89 juta	0 (0%)	11 (11%)	11 (11%)
Jumlah	4 (4%)	96 (96%)	100 (100%)

Berdasarkan Tabel 4.4 mayoritas KK di desa Teluk Baru yang memiliki rata-rata pendapatan perbulan <2,89 juta dengan jenis kelamin Laki-laki yaitu sebanyak 85% sedangkan yang terkecil yaitu KK dengan jenis kelamin Perempuan

sebanyak 4%. Kemudian pada rata-rata pendapatan perbulan ≥2,89 juta jumlah KK terbanyak berada pada KK dengan jenis kelamin Laki-laki sebanyak 11% dan tidak ada KK di desa Teluk Baru yang memiliki rata-rata pendapatan ≥2,89 juta dengan jenis kelamin Perempuan.

4.2 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. Berikut contoh perhitungan untuk menentukan banyaknya data yang masuk ke data *training* menggunakan proporsi 90:10:

$$\begin{aligned}\text{Jumlah data } \textit{training} &= 90\% \times 100 \\ &= \frac{90}{100} \times 100 \\ &= 90\end{aligned}$$

Berikut perhitungan untuk menentukan banyaknya data yang masuk ke data *testing*:

$$\begin{aligned}\text{Jumlah data } \textit{testing} &= 10\% \times 100 \\ &= \frac{10}{100} \times 100 \\ &= 10\end{aligned}$$

Berdasarkan hasil perhitungan di atas dapat dilihat bahwa data yang masuk ke dalam data *training* untuk porposi 90:10 sebanyak 90 dan sisanya sebanyak 10 data masuk ke dalam data *testing*.

4.3 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 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.

Langkah pertama dalam proses pembentukan pohon klasifikasi adalah menghitung nilai *entropy* menggunakan persamaan (2.1). Adapun sebagai contoh perhitungan *entropy* total dan juga *entropy* pada variabel Pekerjaan (X_1) adalah sebagai berikut:

- Menghitung *entropy total* :

$$Entropy(total) = ((-\frac{80}{90}) \times \log_2(\frac{80}{90}) + (-\frac{10}{90}) \times \log_2(\frac{10}{90})) = 0,5033$$

- Menghitung *entropy* tiap kategori dari variabel Pekerjaan

  a. Petani

  $$Entropy(\text{petani}) = ((-\frac{66}{71}) \times \log_2(\frac{66}{71}) + (-\frac{5}{71}) \times \log_2(\frac{5}{71})) = 0,3675$$

  b. Nelayan

  $$Entropy(\text{nelayan}) = ((-\frac{0}{1}) \times \log_2(\frac{0}{1}) + (-\frac{1}{1}) \times \log_2(\frac{1}{1})) = 0$$

  c. PNS

  $$Entropy(\text{PNS}) = ((-\frac{2}{3}) \times \log_2(\frac{2}{3}) + (-\frac{1}{3}) \times \log_2(\frac{1}{3})) = 0,9183$$

  d. Swasta

  $$Entropy(\text{swasta}) = ((-\frac{9}{11}) \times \log_2(\frac{9}{11}) + (-\frac{2}{11}) \times \log_2(\frac{2}{11})) = 0,6840$$

  e. Wiraswasta

  $$Entropy(\text{wiraswasta}) = ((-\frac{3}{4}) \times \log_2(\frac{3}{4}) + (-\frac{1}{4}) \times \log_2(\frac{1}{4})) = 0,8113$$

dengan langkah yang sama dilakukan pula perhitungan nilai *entropy* pada variabel bebas lainnya yaitu variabel jumlah anggota keluarga (X_2), pendidikan terakhir (X_3) dan jenis kelamin (X_4). Kemudian proses dilanjutkan dengan menghitung nilai *gain* untuk setiap variabel bebas menggunakan persamaan (2.2).

Adapun sebagai contoh perhitungan nilai *gain* pada variabel pekerjaan (X_1) adalah sebagai berikut :

$$\begin{aligned} Gain(\text{total, pekerjaan}) = Entropy(total) - (((\frac{71}{90}) \times Entropy(\text{Petani})) \\ + ((\frac{1}{90}) \times Entropy(\text{Nelayan})) + ((\frac{3}{90}) \times Entropy(\text{PNS})) \\ + ((\frac{11}{90}) \times Entropy(\text{Swasta})) + ((\frac{4}{90}) \times Entropy(\text{Wiraswasta}))) \\ = 0,5033 - (((\frac{71}{90}) \times 0,3675) + ((\frac{1}{90}) \times 0) + ((\frac{3}{90}) \times 0,9183) \\ + ((\frac{11}{90}) \times 0,6840) + ((\frac{4}{90}) \times 0,8113)) = 0,0631 \end{aligned}$$

dengan langkah yang sama dilakukan pula perhitungan nilai *gain* pada variabel bebas lainnya yaitu variabel jumlah anggota keluarga (X_2), pendidikan terakhir (X_3) dan jenis kelamin (X_4). Kemudian proses dilanjutkan dengan menghitung nilai *gain ratio* untuk setiap variabel bebas menggunakan persamaan (2.3). Adapun sebagai contoh perhitungan nilai *gain ratio* pada variabel pekerjaan adalah sebagai berikut:

$$\begin{aligned} Gain\ Ratio &= \frac{Gain(\text{total, pekerjaan})}{(E(\text{Petani}) + E(\text{Nelayan}) + E(\text{PNS}) + E(\text{Swasta}) + E(\text{Wiraswasta}))} \\ &= \frac{0,0631}{(0,3675 + 0 + 0,9183 + 0,6840 + 0,8113)} = 0,0227 \end{aligned}$$

dengan langkah yang sama dilakukan pula perhitungan nilai *gain ratio* pada variabel bebas lainnya yaitu variabel jumlah anggota keluarga (X_2), pendidikan terakhir (X_3) dan jenis kelamin (X_4).

Adapun hasil perhitungan nilai *entropy, gain* dan *gain ratio* secara lengkap disajikan pada Tabel 4.5 berikut:

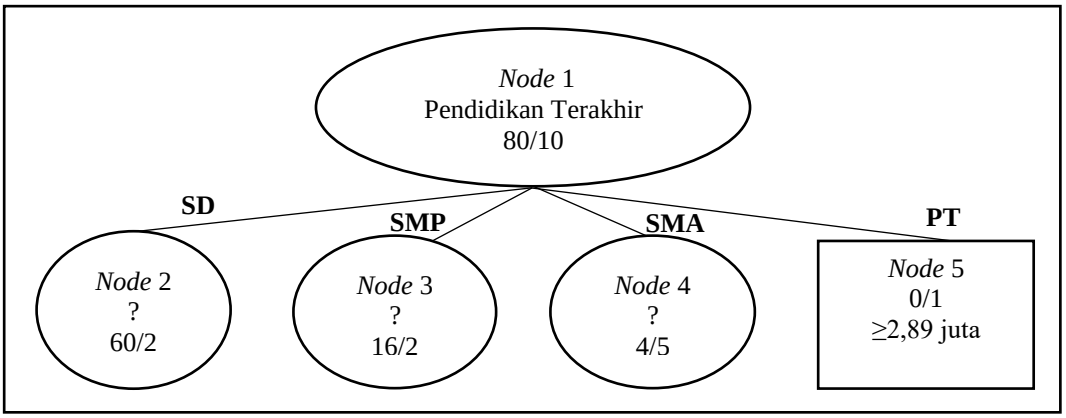
Tabel 4.5 Hasil Perhitungan *Entropy, Gain* dan *Gain Ratio* untuk *Node* Akar

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		

Tabel 4.5 Hasil Perhitungan *Entropy, Gain* dan *Gain Ratio* untuk *Node* Akar (Lanjutan)

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
1	Pekerjaan	PNS	3	2	1	0,9183	0,0631	0,0227
		Swasta	11	9	2	0,6840		
		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		

Berdasarkan Tabel 4.5 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Pendidikan Terakhir sehingga dijadikan sebagai *node* akar (*node* 1). Maka cabang untuk *node* akar ada empat, yaitu SD (*node* 2), SMP (*node* 3), SMA (*node* 4) dan PT (*node* 5) seperti ditunjukkan pada Gambar 4.2. *Node* 2, 3 dan 4 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 5 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas. Dapat dilihat pada setiap isi *node* akar dan cabang memiliki 3 elemen, sebagai contoh untuk *node* 1 elemen pertama yaitu urutan *node* tersebut, elemen kedua yaitu variabel bebas yang terpilih menjadi *node* namun apabila variabel yang terpilih belum diketahui maka diisi dengan simbol tanda tanya (?) dan elemen ketiga yaitu proporsi rata-rata pendapatan perbulan <2,89 juta dengan rata-rata pendapatan perbulan ≥2,89 juta. Kemudian untuk *node* terminal juga berisi 3 elemen sebagai contoh untuk *node* 5 elemen pertama yaitu urutan *node*, elemen kedua yaitu proporsi rata-rata pendapatan perbulan <2,89 juta dengan rata-rata pendapatan perbulan ≥2,89 juta dan elemen ketiga adalah keputusan terpilihnya kelas rata-rata pendapatan untuk *node* terminal tersebut.



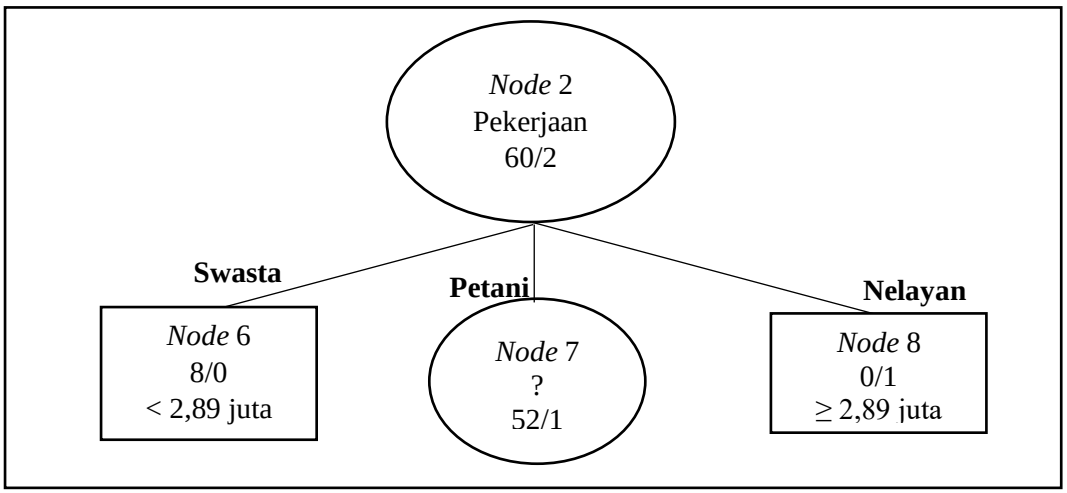
Gambar 4.2 Hasil Pembentukan Cabang di *Node* Akar

Selanjutnya untuk *node* 2, nilai *entropy, gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 4 dan 5 yakni KK dengan pendidikan terakhir SD sebanyak 62 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Pekerjaan (X_1), Jumlah Anggota Keluarga (X_2) dan Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.6 sebagai berikut:

Tabel 4.6 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 2

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
2	Pendidikan Terakhir	SD	62	60	2	0,2056		
	Pekerjaan	**Petani**	**53**	**52**	**1**	**0,1350**	**0,0902**	**0,6677**
		Nelayan	**1**	**0**	**1**	**0**		
		PNS	**0**	**0**	**0**	**0**		
		Swasta	**8**	**8**	**0**	**0**		
		Wiraswasta	**0**	**0**	**0**	**0**		
	Jumlah Anggota Keluarga	>4	22	21	1	0,2668	0,0968	0,2223
		≤4	40	39	1	0,1687		
	Jenis Kelamin	Perempuan	1	1	0	0	0,0008	0,0037
		Laki-laki	61	59	2	0,2082		

Berdasarkan Tabel 4.6 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Pekerjaan sehingga dijadikan sebagai cabang dari *node* 2. Maka cabang untuk *node* 2 ada tiga, yaitu Swasta (*node* 6), Petani (*node* 7) dan Nelayan (*node* 8) seperti ditunjukkan pada Gambar 4.3. Kategori PNS dan Wiraswasta tidak masuk ke dalam cabang karena tidak terdapat kasus pada kategori Pendidikan Terakhir SD. *Node* 7 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 6 dan 8 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas.



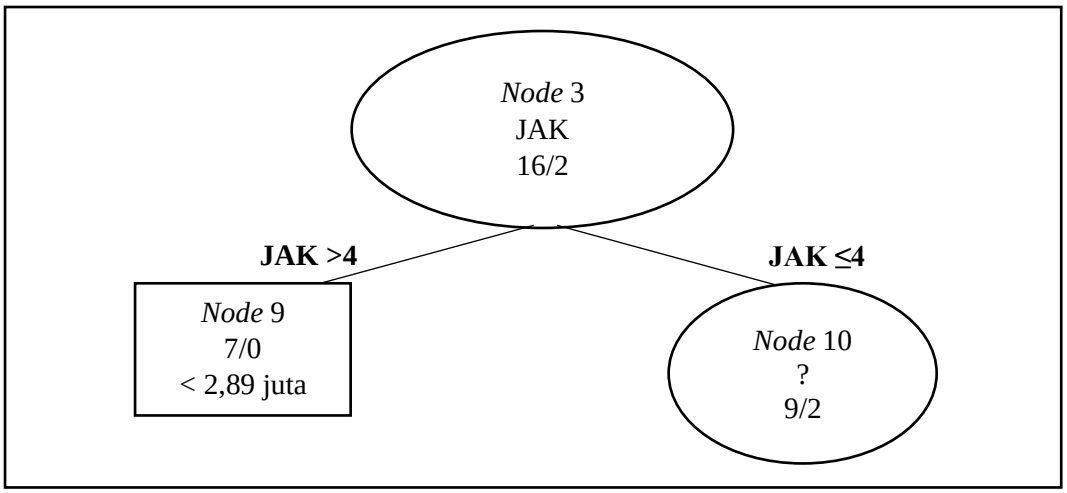
Gambar 4.3 Hasil Pembentukan Cabang di *Node* 2

Selanjutnya untuk *node* 3, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 4 dan 5 yakni KK dengan pendidikan terakhir SMP sebanyak 18 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Pekerjaan (X_1), Jumlah Anggota Keluarga (X_2) dan Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.7.

Tabel 4.7 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 3

Node	Variabel		Jumlah Kasus (S)	<2,89	≥2,89	*Entropy*	*Gain*	*Gain Ratio*
3	Pendidikan Terakhir	SMP	18	16	2	0,5033		
	Pekerjaan	Petani	13	11	2	0,6194	0,0559	0,0903
		Nelayan	0	0	0	0		
		PNS	1	1	0	0		
		Swasta	1	1	0	0		
		Wiraswasta	3	3	0	0		
	Jumlah Anggota Keluarga	**>4**	**7**	**7**	**0**	**0**	**0,0852**	**0,1246**
		≤4	**11**	**9**	**2**	**0,6840**		
	Jenis Kelamin	Perempuan	1	1	0	0	0,0097	0,0186
		Laki-laki	17	15	2	0,5226		

Berdasarkan Tabel 4.7 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Jumlah Anggota Keluarga sehingga dijadikan sebagai cabang dari *node* 3. Maka cabang untuk *node* 3 ada dua, yaitu >4 (*node* 9) dan ≤4 (*node* 10) seperti ditunjukkan pada Gambar 4.4. *Node* 10 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 9 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas.



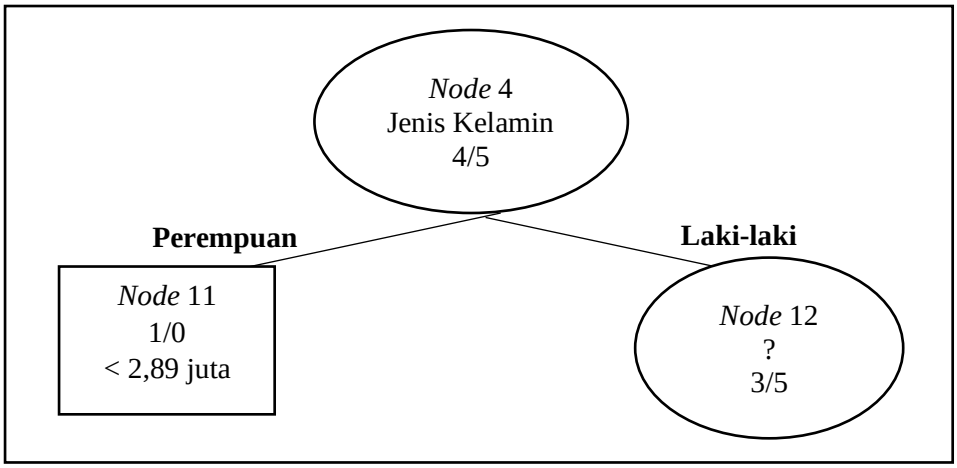
Gambar 4.4 Hasil Pembentukan Cabang di *Node* 3

Selanjutnya untuk *node* 4, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 3 dan 5 yakni KK dengan pendidikan terakhir SMA sebanyak 9 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Pekerjaan (X_1), Jumlah Anggota Keluarga (X_2) dan Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.8.

Tabel 4.8 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 4

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
4	Pendidikan Terakhir	SMA	9	4	5	0,9911		
	Pekerjaan	Petani	5	3	2	0,9710	0,2294	0,1164
		Nelayan	0	0	0	0		
		PNS	2	1	1	1		
		Swasta	2	0	2	0		
		Wiraswasta	0	0	0	0		
	Jumlah Anggota Keluarga	>4	4	1	3	0,8113	0,0911	0,0511
		≤4	11	9	2	0,6840		
	Jenis Kelamin	**Perempuan**	**1**	**1**	**0**	**0**	**0,1427**	**0,1495**
		Laki-laki	**8**	**3**	**5**	**0,9544**		

Berdasarkan Tabel 4.8 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Jenis Kelamin sehingga dijadikan sebagai cabang dari *node* 4. Maka cabang untuk *node* 4 ada dua, yaitu Perempuan (*node* 11) dan Laki-laki (*node* 12) seperti ditunjukkan pada Gambar 4.5. *Node* 12 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 11 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas.



Gambar 4.5 Hasil Pembentukan Cabang di *Node* 4

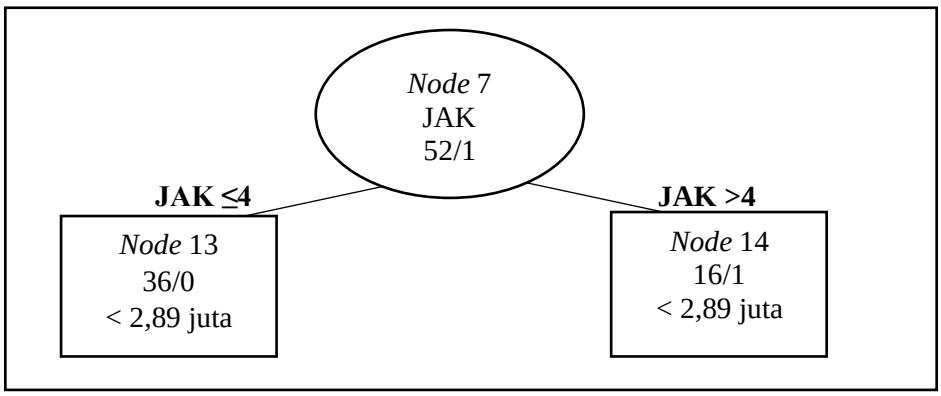
Selanjutnya untuk *node* 7, nilai *entropy, gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 3, 4 dan 11 yakni KK dengan pekerjaan Petani sebanyak 53 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Jumlah Anggota Keluarga (X_2) dan Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy, gain* dan *gain ratio* disajikan dalam Tabel 4.9.

Tabel 4.9 Hasil Perhitungan *Entropy, Gain* dan *Gain Ratio* untuk *Node* 7

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
7	Pekerjaan	Petani	53	52	1	0,1350		
	Jumlah Anggota Keluarga	**>4**	**17**	**16**	**1**	**0,3228**	**0,0315**	**0,0976**
		≤4	**36**	**36**	**0**	**0**		
	Jenis Kelamin	Perempuan	1	1	0	0	0,0005	0,0038
		Laki-laki	52	51	1	0,1371		

Berdasarkan Tabel 4.9 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Jumlah Anggota Keluarga sehingga dijadikan sebagai

cabang dari *node* 7. Maka cabang untuk *node* 7 ada dua, yaitu ≤4 (*node* 13) dan >4 (*node* 14) seperti ditunjukkan pada Gambar 4.6. *Node* 14 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 13 dan 14 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas dan tidak memungkinkan lagi untuk menghasilkan cabang baru.



Gambar 4.6 Hasil Pembentukan Cabang di *Node* 7

Selanjutnya untuk *node* 10, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 4, 5 dan 9 yakni KK dengan jumlah anggota keluarga ≤4 sebanyak 11 KK . Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Pekerjaan (X_1) dan Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.10.

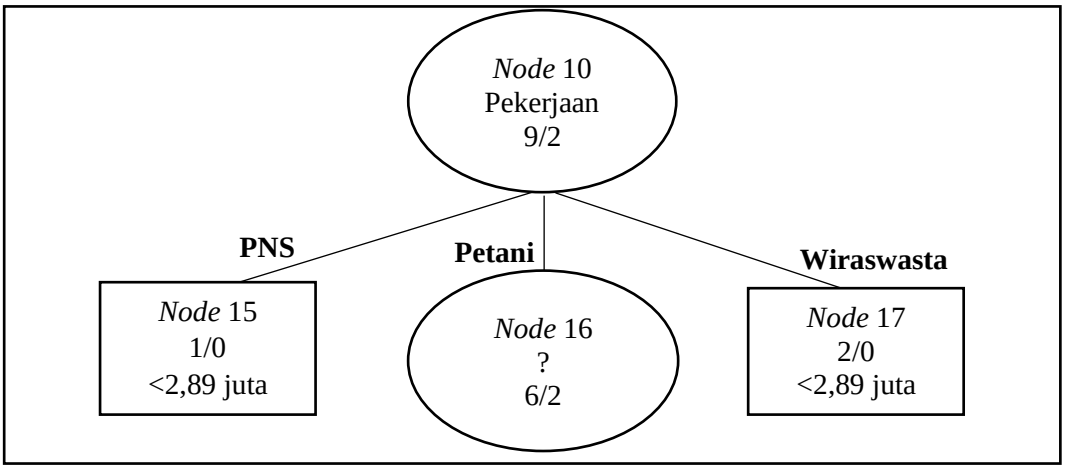
Tabel 4.10 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 10

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
10	Jumlah Anggota Keluarga	≤4	11	9	2	0,6840		

Tabel 4.10 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 10 (Lanjutan)

Node	Variabel		Jumlah Kasus (S)	<2,89	≥2,89	*Entropy*	*Gain*	*Gain Ratio*
10	**Pekerjaan**	**Petani**	**8**	**6**	**2**	**0,8113**	**0,0940**	**0,1159**
		Nelayan	**0**	**0**	**0**	**0**		
		PNS	**1**	**1**	**0**	**0**		
		Swasta	**0**	**0**	**0**	**0**		
		Wiraswasta	**2**	**2**	**0**	**0**		
	Jenis Kelamin	Perempuan	1	1	0	0	0,0277	0,0384
		Laki-laki	10	8	2	0,7219		

Berdasarkan Tabel 4.10 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Pekerjaan sehingga dijadikan sebagai cabang dari *node* 10. Maka cabang untuk *node* 10 ada tiga, yaitu PNS (*node* 15), Petani (*node* 16) dan Wiraswasta (*node* 17) seperti ditunjukkan pada Gambar 4.7. *Node* 16 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 15 dan 17 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas.



Gambar 4.7 Hasil Pembentukan Cabang di *Node* 10

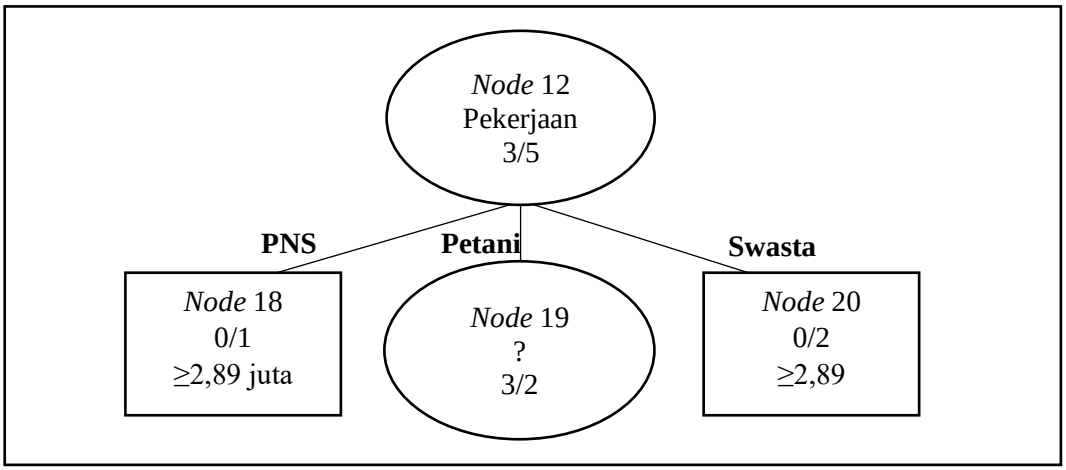
Selanjutnya untuk *node* 12, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 3, 5,

7, 10 dan 11 yakni KK dengan jenis kelamin Laki-laki sebanyak 8 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Pekerjaan (X_1) dan Jumlah Anggota Keluarga (X_2). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.11.

Tabel 4.11 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 12

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
12	Jenis Kelamin	Laki-laki	8	3	5	0,9544		
	Pekerjaan	**Petani**	**5**	**3**	**2**	**0,9710**	**0,3476**	**0,3580**
		Nelayan	**0**	**0**	**0**	**0**		
		PNS	**1**	**0**	**1**	**0**		
		Swasta	**2**	**0**	**2**	**0**		
		Wiraswasta	**0**	**0**	**0**	**0**		
	Jumlah Anggota Keluarga	>4	4	1	3	0,8113	0,0488	0,0269
		≤4	4	2	2	1,000		

Berdasarkan Tabel 4.11 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Pekerjaan sehingga dijadikan sebagai cabang dari *node* 12. Maka cabang untuk *node* 12 ada tiga, yaitu PNS (*node* 18), Petani (*node* 19) dan Swasta (*node* 20) seperti ditunjukkan pada Gambar 4.8. *Node* 19 membentuk *node* cabang karena masih terdapat data sampel pada masing-masing kelas yaitu <2,89 juta dan ≥2,89 juta. *Node* 18 dan 20 menjadi *node terminal* karena data sampel hanya ada di salah satu kelas.



Gambar 4.8 Hasil Pembentukan Cabang di *Node* 12

Selanjutnya untuk *node* 14, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 4, 5, 6, 8 dan 13 yakni KK dengan jumlah anggota keluarga >4 sebanyak 17 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.12.

Tabel 4.12 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 14

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
14	Jumlah Anggota Keluarga	>4	17	6	1	0,3228		
	Jenis Kelamin	Perempuan	0	0	0	0	0	0
		Laki-laki	17	6	1	0,3228		

Berdasarkan Tabel 4.12 dapat dilihat bahwa variabel yang memiliki nilai *gain ratio* tertinggi adalah variabel Jenis Kelamin sehingga dijadikan sebagai cabang dari *node* 12. Akan tetapi karena sampel hanya berada disalah satu kelas saja, maka

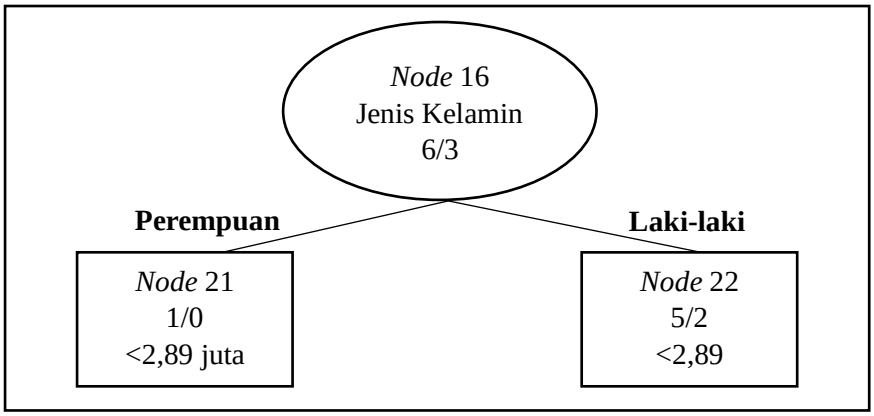
variabel Jenis Kelamin tidak bisa menjadi cabang. Oleh karena itu *node* berhenti dan *node* 12 diisi oleh variabel Jumlah Anggota Keluarga dengan kategori (JAK >4).

Selanjutnya untuk *node* 16, nilai *entropy, gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 4, 5, 9, 15 dan 17 yakni KK dengan pekerjaan petani sebanyak 8 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Jenis Kelamin (X_4). Adapun hasil perhitungan *entropy, gain* dan *gain ratio* disajikan dalam Tabel 4.13 berikut:

Tabel 4.13 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 16

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
16	Pekerjaan	Petani	8	6	3	0,8113		
	Jenis Kelamin	**Perempuan**	**1**	**1**	**0**	**0**	**0,0560**	**0,0649**
		Laki-laki	**7**	**5**	**2**	**0,8631**		

Berdasarkan Tabel 4.13 dapat dilihat bahwa hanya terdapat satu variabel saja yaitu variabel Jenis Kelamin pada *node* 16, maka tentu saja variabel Jenis Kelamin yang menjadi cabang dari *node* 16. Maka cabang untuk *node* 16 ada dua, yaitu Perempuan (*node* 21) dan Laki-laki (*node* 22) seperti ditunjukkan pada Gambar 4.9 dapat dilihat bahwa tidak ada *node* yang membentuk *node* cabang karena hanya tersisa 1 variabel dan tidak memungkinkan lagi untuk membentuk suatu cabang baru sehingga *node* 21 dan 22 menjadi *node terminal*.



Gambar 4.9 Hasil Pembentukan Cabang di *Node* 16

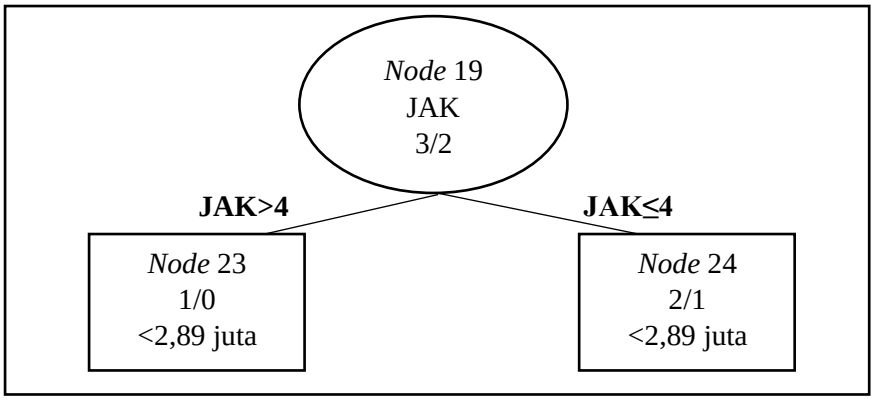
Selanjutnya untuk *node* 19, nilai *entropy*, *gain* dan *gain ratio* dihitung dahulu seperti pada langkah awal mencari *node* akar namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2, 3, 5, 11, 18 dan 20 yakni KK dengan pekerjaan Petani sebanyak 5 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.1) untuk mencari nilai *entropy*, Persamaan (2.2) untuk mencari nilai *gain* dan Persamaan (2.3) untuk mencari nilai *gain ratio*. Variabel yang digunakan untuk menentukan *node* selanjutnya adalah Jumlah Anggota Keluarga (X_2). Adapun hasil perhitungan *entropy*, *gain* dan *gain ratio* disajikan dalam Tabel 4.14.

Tabel 4.14 Hasil Perhitungan *Entropy*, *Gain* dan *Gain Ratio* untuk *Node* 19

Node	**Variabel**		**Jumlah Kasus (S)**	**<2,89**	**≥2,89**	***Entropy***	***Gain***	***Gain Ratio***
16	Pekerjaan	Petani	5	3	2	0,9710		
	Jumlah Anggota Keluarga	**>4**	**1**	**1**	**0**	**0**	**0,4200**	**0,4573**
		≤4	**3**	**2**	**1**	**0,9183**		

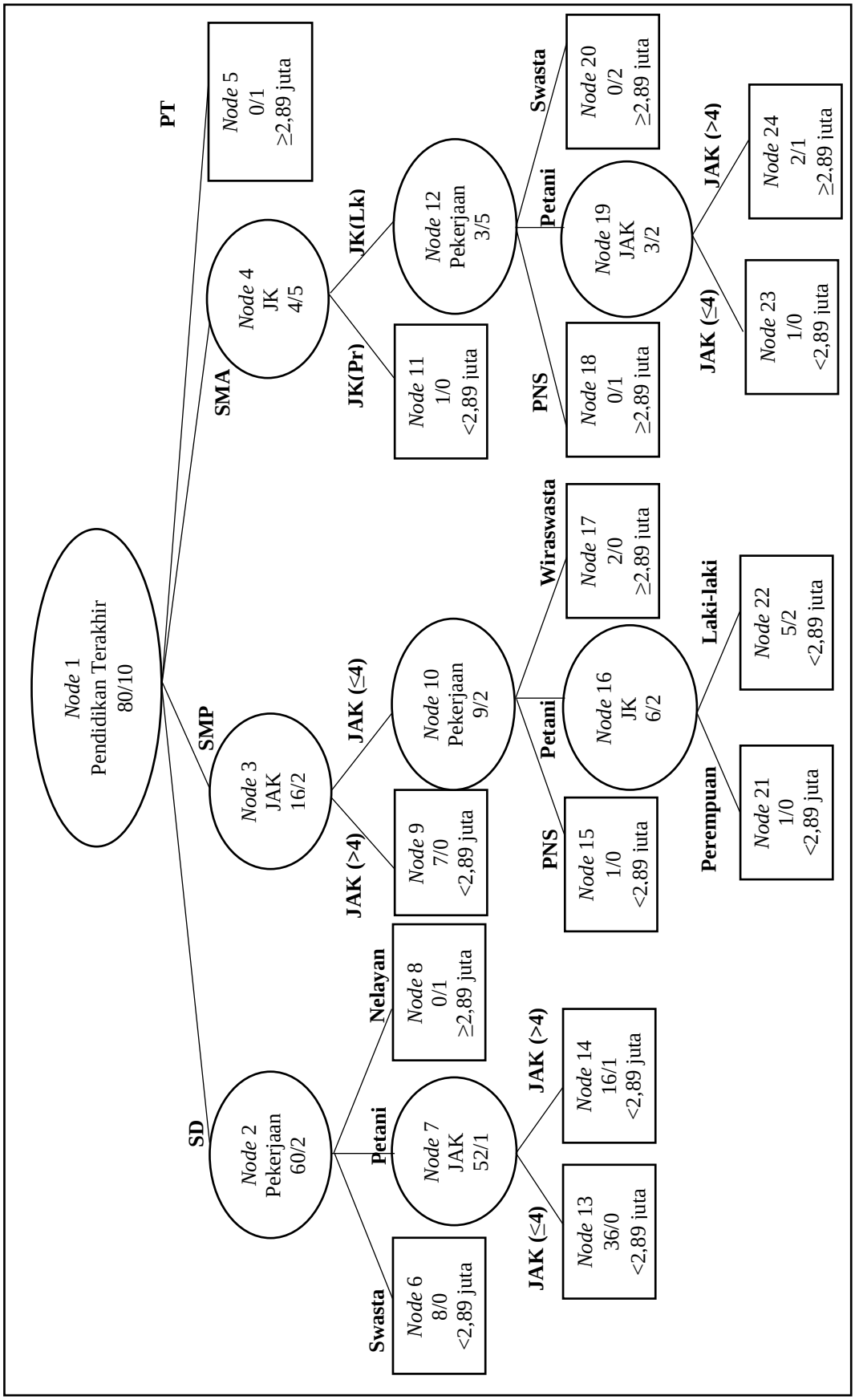
Berdasarkan Tabel 4.14 dapat dilihat bahwa hanya terdapat satu variabel saja yaitu variabel Jumlah Anggota Keluarga pada *node* 19, maka tentu saja variabel Jumlah Anggota Keluarga yang menjadi cabang dari *node* 19. Maka cabang untuk *node* 19 ada dua, yaitu >4 (*node* 23) dan ≤4 (*node* 24) seperti ditunjukkan pada Gambar

4.10 dapat dilihat bahwa tidak ada *node* yang membentuk *node* cabang karena hanya tersisa 1 variabel dan tidak memungkinkan lagi untuk membentuk suatu cabang baru sehingga *node* 23 dan 24 menjadi *node terminal*.



Gambar 4.10 Hasil Pembentukan Cabang di *Node* 19

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 C5.0 disajikan pada Gambar 4.11.



Gambar 4.11 Pohon Klasifikasi Algoritma C5.0

Pada Gambar 4.11 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.

4.4 *Classification and Regression Tree* (CART)

Berikut analisis menggunakan metode CART :

4.4.1 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.

4.4.1.1 Pemilihan Pemilah

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 (2.4b). Kemudian pemilihan pemilah untuk variabel bebas bertipe ordinal yaitu variabel Jumlah Anggota Keluarga (X_2) dan Pendidikan Terakhir (X_3).

Langkah selanjutnya yaitu menghitung nilai indeks *gini* untuk setiap pemilah sesuai dengan persamaan (2.7) 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.

Berikut proses pembentukan pemilah untuk masing-masing variabel bebas:

1. Variabel Pekerjaan mempunyai 5 kategori yaitu Petani, Nelayan, PNS, Swasta dan Wiraswasta. Maka kemungkinan pemilahan untuk variabel ini adalah 2^{5-1}-1=16-1=15 pemilah yaitu {(Petani), (Nelayan, PNS, Swasta, Wiraswasta)}, {(Nelayan), (Petani, PNS, Swasta, Wiraswasta)}, {(PNS), (Petani, Nelayan, Swasta, Wiraswasta)}, {(Swasta), (Petani, Nelayan, PNS, Wiraswasta)}, {(Wiraswasta), (Petani, Nelayan, PNS, Swasta)}, {(Petani, Nelayan), (PNS, Swasta, Wiraswasta)}, {(Petani, PNS), (Nelayan, Swasta, Wiraswasta)}, {(Petani, Swasta), (Nelayan, PNS, Wiraswasta)}, {(Petani, Wiraswasta), (Nelayan, PNS, Swasta)}, {(Petani, Nelayan, PNS), (Swasta, Wiraswasta)}, {(Petani, Nelayan, Swasta), (PNS, Wiraswasta)}, {(Petani, Nelayan, Wiraswasta), (PNS, Swasta)}, {(Nelayan, PNS), (Petani, Swasta, Wiraswasta)}, {(Nelayan, Swasta), (Petani, PNS, Wiraswasta)}, {(Nelayan, Wiraswasta), (Petani, PNS, Swasta)}.

a. Pemilahan Pekerjaan Kemungkinan Pertama

Tabel 4.15 Pemilahan Pekerjaan Kemungkinan Pertama

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani	Nelayan, PNS, Swasta, Wiraswasta	
< 2,89 juta	66	14	80
≥ 2,89 juta	5	5	10
Jumlah	71	19	90

Kemungkinan pertama memiliki dua kategori, yaitu {(Petani), (Nelayan, PNS, Swasta, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut:

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

$$= \frac{71}{90}(1-(\frac{66}{71})^2 - (\frac{5}{71})^2) + \frac{19}{90}(1-(\frac{14}{19})^2 - (\frac{5}{19})^2)$$

$$= 0{,}1852$$

b. Pemilahan Pekerjaan Kemungkinan Kedua

Tabel 4.16 Pemilahan Pekerjaan Kemungkinan Kedua

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Nelayan	Petani, PNS, Swasta, Wiraswasta	
< 2,89 juta	0	80	80
≥ 2,89 juta	1	9	10
Jumlah	1	89	90

Kemungkinan kedua memiliki dua kategori, yaitu {(Nelayan), (Petani, PNS, Swasta, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{1}{90}(1-(\frac{0}{1})^2-(\frac{1}{1})^2) + \frac{89}{90}(1-(\frac{80}{89})^2-(\frac{9}{89})^2)$$
$$= 0{,}1798$$

c. Pemilahan Pekerjaan Kemungkinan Ketiga

Tabel 4.17 Pemilahan Pekerjaan Kemungkinan Ketiga

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	PNS	Petani, Nelayan, Swasta, Wiraswasta	
< 2,89 juta	2	78	80
≥ 2,89 juta	1	9	10
Jumlah	3	87	90

Kemungkinan ketiga memiliki dua kategori, yaitu {(PNS), (Petani, Nelayan, Swasta, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut :

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{3}{90}(1-(\frac{2}{3})^2-(\frac{1}{3})^2) + \frac{87}{90}(1-(\frac{78}{87})^2-(\frac{9}{87})^2)$$
$$= 0{,}1941$$

d. Pemilahan Pekerjaan Kemungkinan Keempat

Tabel 4.18 Pemilahan Pekerjaan Kemungkinan Keempat

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Swasta	Petani, Nelayan, PNS, Wiraswasta	
< 2,89 juta	9	71	80
≥ 2,89 juta	2	8	10
Jumlah	11	79	90

Kemungkinan keempat memiliki dua kategori, yaitu {(Swasta), (Petani, Nelayan, PNS, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

e. Pemilahan Pekerjaan Kemungkinan Kelima

Tabel 4.19 Pemilahan Pekerjaan Kemungkinan Kelima

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Wiraswasta	Petani, Nelayan, PNS, Swasta	
< 2,89 juta	3	77	80
≥ 2,89 juta	1	9	10
Jumlah	4	86	90

Kemungkinan kelima memiliki dua kategori, yaitu {(Wiraswasta), (Petani, Nelayan, PNS, Swasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

f. Pemilahan Pekerjaan Kemungkinan Keenam

Tabel 4.20 Pemilahan Pekerjaan Kemungkinan Keenam

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Nelayan	PNS, Swasta, Wiraswasta	
< 2,89 juta	66	14	80
≥ 2,89 juta	6	4	10
Jumlah	72	18	90

Kemungkinan keenam memiliki dua kategori, yaitu {(Petani, Nelayan), (PNS, Wiraswasta, Swasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

g. Pemilahan Pekerjaan Kemungkinan Ketujuh

Tabel 4.21 Pemilahan Pekerjaan Kemungkinan Ketujuh

Penghasilan	Pekerjaan (X_1)		Jumlah
	Petani, PNS	Nelayan, Swasta, Wiraswasta	
< 2,89 juta	68	12	80
≥ 2,89 juta	6	4	10
Jumlah	74	16	90

Kemungkinan ketujuh memiliki dua kategori, yaitu {(Petani, PNS), (Nelayan, Wiraswasta, Swasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

h. Pemilahan Pekerjaan Kemungkinan Kedelapan

Tabel 4.22 Pemilahan Pekerjaan Kemungkinan Kedelapan

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Swasta	Nelayan, PNS, Wiraswasta	
< 2,89 juta	75	5	80
≥ 2,89 juta	7	3	10
Jumlah	82	8	90

Kemungkinan kedelapan memiliki dua kategori, yaitu {(Petani, Swasta), (Nelayan, PNS, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

i. Pemilahan Pekerjaan Kemungkinan Kesembilan

Tabel 4.23 Pemilahan Pekerjaan Kemungkinan Kesembilan

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Wiraswasta	Nelayan, PNS, Swasta	
< 2,89 juta	69	11	80
≥ 2,89 juta	6	4	10
Jumlah	75	15	90

Kemungkinan kesembilan memiliki dua kategori, yaitu {(Petani, Wiraswasta), (Nelayan, PNS, Swasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

j. Pemilahan Pekerjaan Kemungkinan Kesepuluh

Tabel 4.24 Pemilahan Pekerjaan Kemungkinan Kesepuluh

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Nelayan, PNS	Swasta, Wiraswasta	
< 2,89 juta	68	12	80
≥ 2,89 juta	7	3	10
Jumlah	75	15	90

Kemungkinan kesepuluh memiliki dua kategori, yaitu {(Petani, Nelayan, PNS), (Swasta, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

k. Pemilahan Pekerjaan Kemungkinan Kesebelas

Tabel 4.25 Pemilahan Pekerjaan Kemungkinan Kesebelas

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Nelayan, Swasta	PNS, Wiraswasta	
< 2,89 juta	75	5	80
≥ 2,89 juta	8	2	10
Jumlah	83	7	90

Kemungkinan kesebelas memiliki dua kategori, yaitu {(Petani, Nelayan, Swasta), (PNS, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

l. Pemilahan Pekerjaan Kemungkinan Kedua Belas

Tabel 4.26 Pemilahan Pekerjaan Kemungkinan Kedua Belas

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Petani, Nelayan, Wiraswasta	PNS, Swasta	
< 2,89 juta	69	11	80
≥ 2,89 juta	7	3	10
Jumlah	76	14	90

Kemungkinan kedua belas memiliki dua kategori, yaitu {(Petani, Nelayan, Swasta), (PNS, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

m. Pemilahan Pekerjaan Kemungkinan Ketiga Belas

Tabel 4.27 Pemilahan Pekerjaan Kemungkinan Ketiga Belas

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Nelayan, PNS	Petani, Swasta, Wiraswasta	
< 2,89 juta	2	78	80
≥ 2,89 juta	2	8	10
Jumlah	4	86	90

Kemungkinan ketiga belas memiliki dua kategori, yaitu {(Nelayan, PNS), (Petani, Swasta, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

n. Pemilahan Pekerjaan Kemungkinan Keempat Belas

Tabel 4.28 Pemilahan Pekerjaan Kemungkinan Keempat Belas

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Nelayan, Swasta	Petani, PNS, Wiraswasta	
< 2,89 juta	9	71	80
≥ 2,89 juta	3	7	10
Jumlah	12	78	90

Kemungkinan keempat belas memiliki dua kategori, yaitu {(Nelayan, Swasta), (Petani, PNS, Wiraswasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

o. Pemilahan Pekerjaan Kemungkinan Kelima Belas

Tabel 4.29 Pemilahan Pekerjaan Kemungkinan Kelima Belas

Rata-rata Pendapatan Perbulan	Pekerjaan (X_1)		Jumlah
	Nelayan, Wiraswasta	Petani, PNS, Swasta	
< 2,89 juta	3	77	80
≥ 2,89 juta	2	8	10
Jumlah	5	85	90

Kemungkinan kelima belas memiliki dua kategori, yaitu {(Nelayan, Wiraswasta), (Petani, PNS, Swasta)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

2. Variabel Jumlah Anggota Keluarga mempunyai 2 kategori yaitu >4 dan ≤4. Maka kemungkinan pemilahan untuk variabel ini adalah 2-1=1 pemilah yaitu {(>4), (≤4)}.

Tabel 4.30 Pemilahan Jumlah Anggota Keluarga Kemungkinan Pertama

Rata-rata Pendapatan Perbulan	Jumlah Anggota Keluarga (X_2)		Jumlah
	>4	≤4	
< 2,89 juta	29	51	80
≥ 2,89 juta	4	6	10
Jumlah	33	57	90

Kemungkinan dalam variabel jumlah anggota keluarga memiliki dua kategori, yaitu {(>4), (≤4)}.

Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut:

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{33}{90}(1-(\frac{29}{33})^2 - (\frac{4}{33})^2) + \frac{57}{90}(1-(\frac{51}{57})^2 - (\frac{6}{57})^2)$$
$$= 0{,}1974$$

3. Variabel Pendidikan Terakhir mempunyai 4 kategori yaitu SD, SMP, SMA dan PT. Maka kemungkinan pemilahan untuk variabel ini adalah 4-1=3 pemilah yaitu {(SD), (SMP, SMA, PT)}, {(SD, SMP), (SMA, PT)} dan {(SD, SMP, SMA), (PT)}

a. Pemilahan Pendidikan Terakhir Kemungkinan Pertama

Tabel 4.31 Pemilahan Pendidikan Terakhir Kemungkinan Pertama

Rata-rata Pendapatan Perbulan	Pendidikan Terakhir (X_3)		Jumlah
	SD	SMP, SMA, PT	
< 2,89 juta	60	20	80
≥ 2,89 juta	2	8	10
Jumlah	62	28	90

Kemungkinan pertama memiliki dua kategori, yaitu {(SD), (SMP, SMA, PT)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut :

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{62}{90}(1-(\frac{60}{62})^2 - (\frac{2}{62})^2) + \frac{28}{90}(1-(\frac{20}{28})^2 - (\frac{8}{28})^2)$$
$$= 0{,}1700$$

b. Pemilahan Pendidikan Terakhir Kemungkinan Kedua

Tabel 4.32 Pemilahan Pendidikan Terakhir Kemungkinan Kedua

Rata-rata Pendapatan Perbulan	Pendidikan Terakhir (X_3)		Jumlah
	SD, SMP	SMA, PT	
< 2,89 juta	76	4	80
≥ 2,89 juta	4	6	10
Jumlah	80	10	90

Kemungkinan kedua memiliki dua kategori, yaitu {(SD, SMP), (SMA, PT)}.

Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut :

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

$$= \frac{80}{90}(1-(\frac{76}{80})^2-(\frac{4}{80})^2) + \frac{10}{90}(1-(\frac{4}{10})^2-(\frac{6}{10})^2)$$

$$= 0{,}1378$$

c. Pemilahan Pendidikan Terakhir Kemungkinan Ketiga

Tabel 4.33 Pemilahan Pendidikan Terakhir Kemungkinan Ketiga

Rata-rata Pendapatan Perbulan	Pendidikan Terakhir (X_3)		Jumlah
	SD, SMP, SMA	PT	
< 2,89 juta	80	0	80
≥ 2,89 juta	9	1	10
Jumlah	89	1	90

Kemungkinan ketiga memiliki dua kategori, yaitu {(SD, SMP, SMA), (PT)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10.

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{89}{90}(1-(\frac{80}{89})^2 - (\frac{9}{89})^2) + \frac{1}{90}(1-(\frac{0}{1})^2 - (\frac{1}{1})^2)$$
$$= 0{,}1798$$

4. Variabel Jenis Kelamin mempunyai 2 kategori yaitu Perempuan dan Laki-Laki. Maka kemungkinan pemilahan untuk variabel ini adalah 2-1=1 pemilah yaitu {(Perempuan), (Laki-Laki)}.

Tabel 4.34 Pemilahan Jenis Kelamin Kemungkinan Pertama

Penghasilan	Jenis Kelamin (X_4)		Jumlah
	Perempuan	Laki-laki	
< 2,89 juta	3	77	80
≥ 2,89 juta	0	10	10
Jumlah	3	87	90

Kemungkinan pada variabel jenis kelamin memiliki dua kategori, yaitu {(Perempuan), (Laki-Laki)}. Kemudian dicari nilai indeks *gini* menggunakan persamaan 2.10 seperti berikut:

$$Gini_{pembelahan}(t) = \frac{b_1}{b} gini(D_1) + \frac{b_2}{b} gini(D_2)$$
$$= \frac{3}{90}(1-(\frac{3}{3})^2 - (\frac{0}{3})^2) + \frac{87}{90}(1-(\frac{77}{87})^2 - (\frac{10}{87})^2)$$
$$= 0{,}1967$$

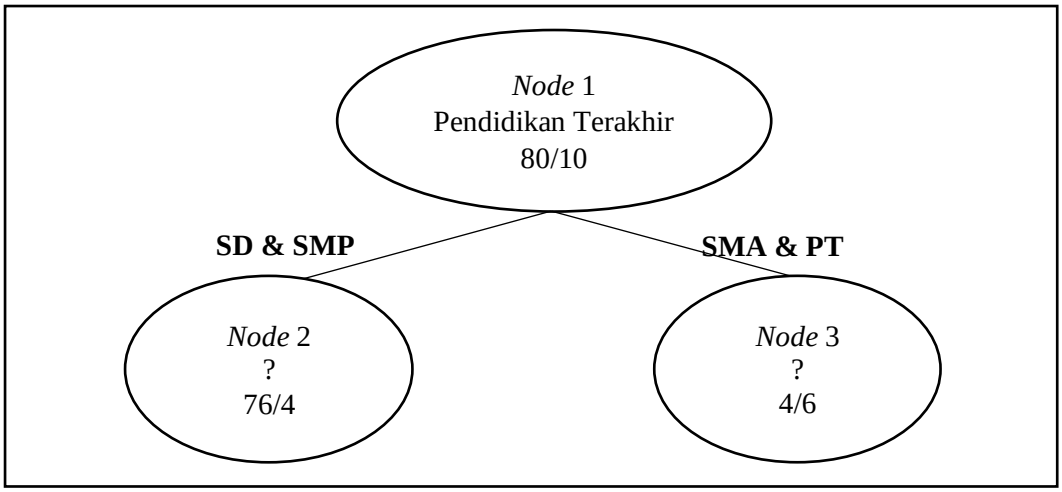
Setelah dilakukan perhitungan indeks *gini* pada masing-masing pemilah, didapat hasil sebagai berikut :

Tabel 4.35 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
	{(Swasta), (Petani, Nelayan, PNS, Wiraswasta)}	0,1961
	{(Wiraswasta), (Petani, Nelayan, PNS, Swasta)}	0,1957
	{(Petani, Nelayan), (PNS, Swasta, Wiraswasta)}	0,1914
	{(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

Berdasarkan Tabel 4.35 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Pendidikan Terakhir dengan pemilah {(SD, SMP), (SMA, PT)} sebesar 0,1378.

Maka pemilah ini terpilih menjadi pemilah pertama untuk pohon klasifikasi seperti pada Gambar 4.12 berikut :



Gambar 4.12 Hasil Pembentukan Cabang di *Node* Akar

Setelah terbentuk dan terpilih menjadi pemilah tebaik, dapat dilihat bahwa pendidikan terakhir SD dan SMP masuk ke *node* 2, sedangkan pendidikan terakhir SMA dan PT masuk ke *node* 3. Langkah selanjutnya yaitu mencari pemilah terbaik dari masing-masing cabang yaitu *node* 2 dan *node* 3.

Penentuan pemilah terbaik untuk *node* 2 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3 yakni KK dengan pendidikan terakhir SD dan SMP sebanyak 80 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.36 berikut:

Tabel 4.36 Hasil Perhitungan Indeks *Gini* untuk *Node* 2

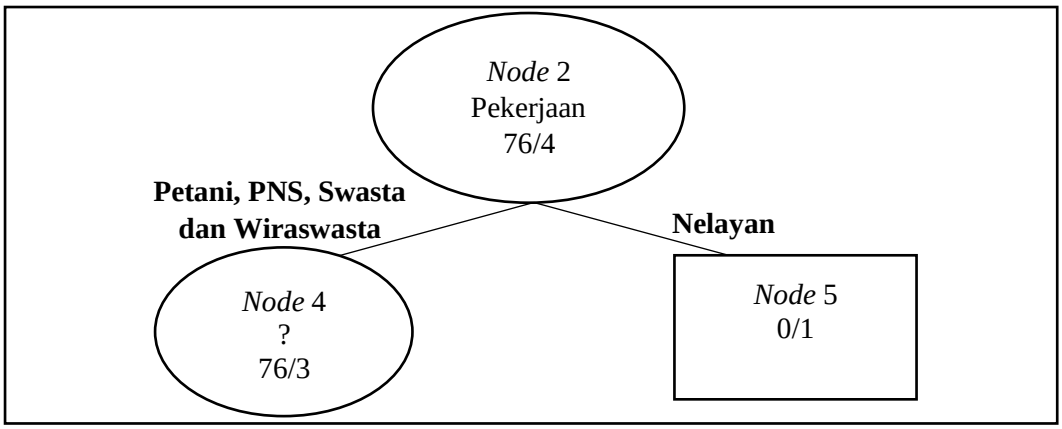
Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani), (Nelayan, PNS, Swasta, Wiraswasta)}	0,0948
	{(Nelayan), (Petani, PNS, Swasta, Wiraswasta)}	**0,0722**
	{(PNS), (Petani, Nelayan, Swasta, Wiraswasta)}	0,0949

Tabel 4.36 Hasil Perhitungan Indeks *Gini* untuk *Node* 2 (Lanjutan)

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Swasta), (Petani, Nelayan, PNS, Wiraswasta)}	0,0944
	{(Wiraswasta), (Petani, Nelayan, PNS, Swasta)}	0,0948
	{(Petani, Nelayan), (PNS, Swasta, Wiraswasta)}	0,0940
	{(Petani, PNS), (Nelayan, Swasta, Wiraswasta)}	0,0947
	{(Petani, Swasta), (Nelayan, PNS, Wiraswasta)}	0,0920
	{(Petani, Wiraswasta), (Nelayan, PNS, Swasta)}	0,0945
	{(Petani, Nelayan, PNS), (Swasta, WIraswasta)}	0,0941
	{(Petani, Nelayan, Swasta), (PNS, Wiraswasta)}	0,0947
	{(Petani, Nelayan, Wiraswasta), (PNS, Swasta)}	0,0943
	{(Nelayan, PNS), (Petani, Swasta, Wiraswasta)}	0,0846
	{(Nelayan, Swasta), (Petani, PNS, Wiraswasta)}	0,0943
	{(Nelayan, Wiraswasta), (Petani, PNS, Swasta)}	0,0908
Jumlah Anggota Keluarga	{(>4), (≤4)}	0,0947
Pendidikan Terakhir	{(SD), (SMP)}	0,0928
Jenis Kelamin	{(Perempuan), (Laki-Laki)}	0,0949

Berdasarkan Tabel 4.36 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Pekerjaan dengan pemilah {(Nelayan), (Petani, PNS, Swasta, Wiraswasta)} sebesar 0,0722. Maka pemilah ini terpilih menjadi pemilah untuk *node* 2 dan pekerjaan Nelayan menjadi *node terminal* dikarenakan sampel hanya berada di salah satu kelas.

Maka *node* 2 menjadi seperti pada Gambar 4.13 berikut :



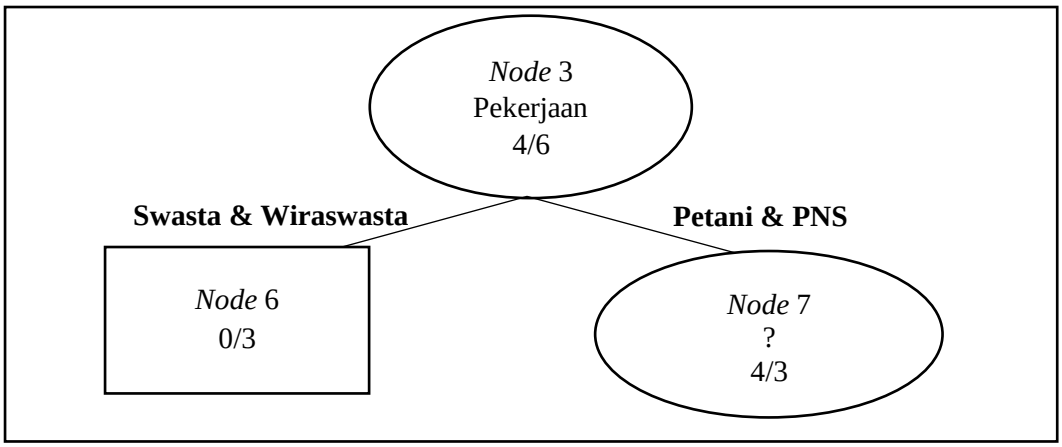
Gambar 4.13 Hasil Pembentukan Cabang di *Node* 2

Penentuan pemilah terbaik untuk *node* 3 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2 yakni KK dengan pendidikan terakhir SMA dan PT sebanyak 10 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.37.

Tabel 4.37 Hasil Perhitungan Indeks *Gini* untuk *Node* 3

Variabel	**Kategori**	**Indeks *Gini***
Pekerjaan	{(Petani), (PNS, Swasta, Wiraswasta)}	0,4000
	{(PNS), (Petani, Swasta, Wiraswasta)}	0,4750
	{(Swasta), (Petani, Swasta, Wiraswasta)}	0,4000
	{(Wiraswasta), (Petani, PNS, Swasta)}	0,4444
	{(Petani, PNS), (Swasta, Wiraswasta)}	**0,3429**
	{(Petani, Swasta), (PNS, Wiraswasta)}	0,4762
	{(Petani, Wiraswasta), (PNS, Swasta)}	0,4500
Jumlah Anggota Keluarga	{(>4), (≤4)}	0,4440
Pendidikan Terakhir	{(SMA), (PT)}	0,4500
Jenis Kelamin	{(Perempuan), (Laki-laki)}	0,4000

Berdasarkan Tabel 4.37 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Pekerjaan dengan pemilah {(Petani, PNS), (Swasta, Wiraswasta)} sebesar 0,3429. Maka pemilah ini terpilih menjadi pemilah untuk *node* 3 kemudian pekerjaan Swasta dan Wiraswasta menjadi *node terminal* dikarenakan sampel hanya berada di salah satu kelas. Maka *node* 3 menjadi seperti pada Gambar 4.14 berikut:



Gambar 4.14 Hasil Pembentukan Cabang di *Node* 3

Penentuan pemilah terbaik untuk *node* 4 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3 dan 5 yakni KK dengan pekerjaan Petani, PNS, Swasta dan Wiraswasta sebanyak 79 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.38.

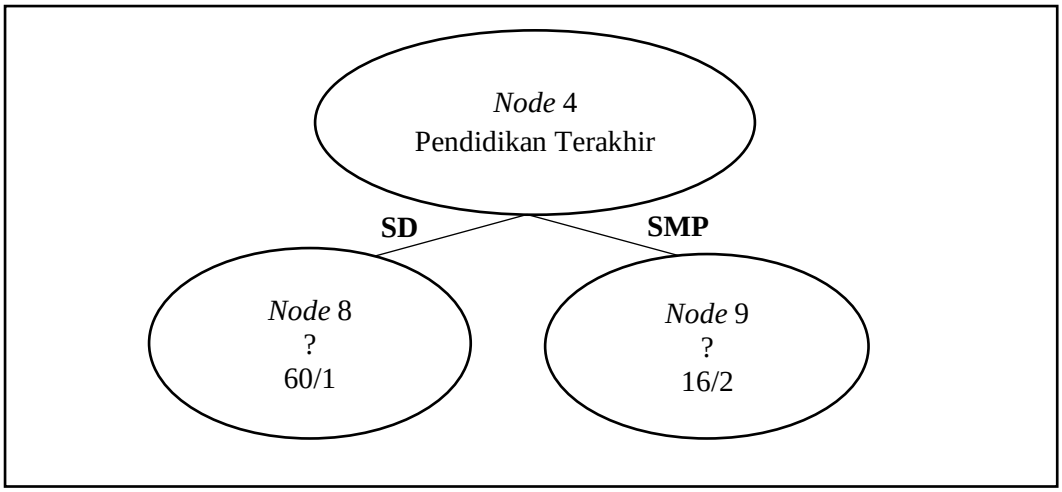
Tabel 4.38 Hasil Perhitungan Indeks *Gini* untuk *Node* 4

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani), (PNS, Swasta, Wiraswasta)}	0,0725
	{(PNS), (Petani, Swasta, Wiraswasta)}	0,0730
	{(Swasta), (Petani, Swasta, Wiraswasta)}	0,0727
	{(Wiraswasta), (Petani, PNS, Swasta)}	0,0730

Tabel 4.38 Hasil Perhitungan Indeks *Gini* untuk *Node* 4 (Lanjutan)

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani, PNS), (Swasta, Wiraswasta)}	0,0725
	{(Petani, Swasta), (PNS, Wiraswasta)}	0,0729
	{(Petani, Wiraswasta), (PNS, Swasta)}	0,0726
Jumlah Anggota Keluarga	{(>4), (≤4)}	0,0731
Pendidikan Terakhir	**{(SD), (SMP)}**	**0,0699**
Jenis Kelamin	{(Perempuan), (Laki-laki)}	0,0730

Berdasarkan Tabel 4.38 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Pendidikan Terakhir dengan pemilah {(SD), (SMP)} sebesar 0,0699. Maka *node* 4 menjadi seperti pada Gambar 4.15 berikut :



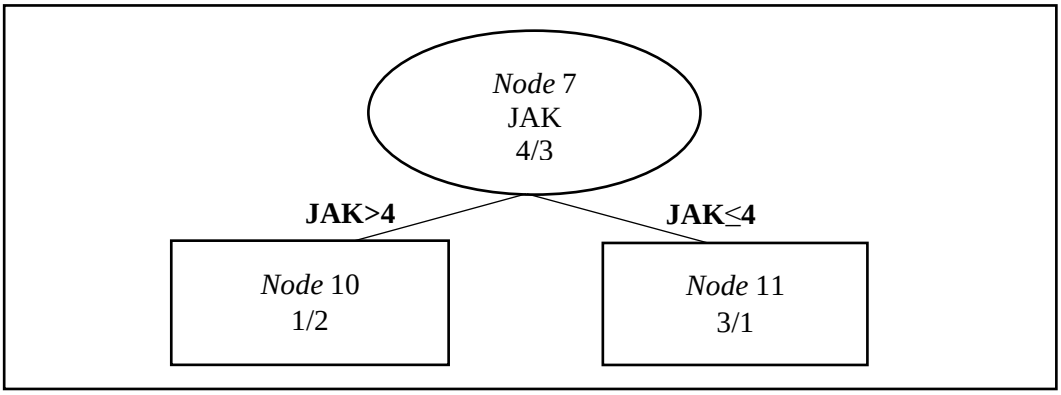
Gambar 4.15 Hasil Pembentukan Cabang di *Node* 4

Penentuan pemilah terbaik untuk *node* 7 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 2 dan 6 yakni KK dengan pekerjaan Petani dan PNS sebanyak 7 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.39.

Tabel 4.39 Hasil Perhitungan Indeks *Gini* untuk *Node* 7

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani), (PNS)}	0,4857
Jumlah Anggota Keluarga	**{(>4), (≤4)}**	**0,4048**
Jenis Kelamin	{(Perempuan), (Laki-laki)}	0,4286

Berdasarkan Tabel 4.39 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Jumlah Anggota Keluarga dengan pemilah {(>4), (≤4)} sebesar 0,4048. Karena jumlah sampel dalam tiap kelas <5 maka kedua cabang menjadi *node terminal*. Sehingga *node* 7 menjadi seperti pada Gambar 4.16 berikut:



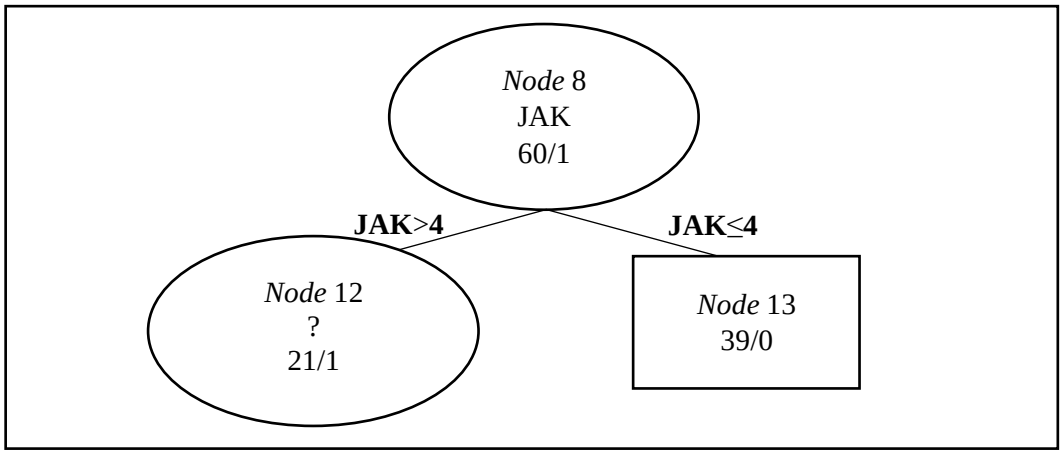
Gambar 4.16 Hasil Pembentukan Cabang di *Node* 7

Penentuan pemilah terbaik untuk *node* 8 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 5 dan 9 yakni KK dengan pendidikan terakhir SD sebanyak 61 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.40.

Tabel 4.40 Hasil Perhitungan Indeks *Gini* untuk *Node* 8

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{(Petani), (Swasta)}	0,0322
Jumlah Anggota Keluarga	**{(>4), (≤4)}**	**0,0310**
Jenis Kelamin	{(Perempuan), (Laki-laki)}	0,0322

Berdasarkan Tabel 4.40 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Jumlah Anggota Keluarga dengan pemilah {(>4), (≤4)} sebesar 0,0310. Maka pemilah ini terpilih menjadi pemilah untuk *node* 8 kemudian jumlah anggota keluarga ≤4 menjadi *node terminal* dikarenakan sampel hanya berada di salah satu kelas. Maka *node* 8 menjadi seperti pada Gambar 4.17 berikut:



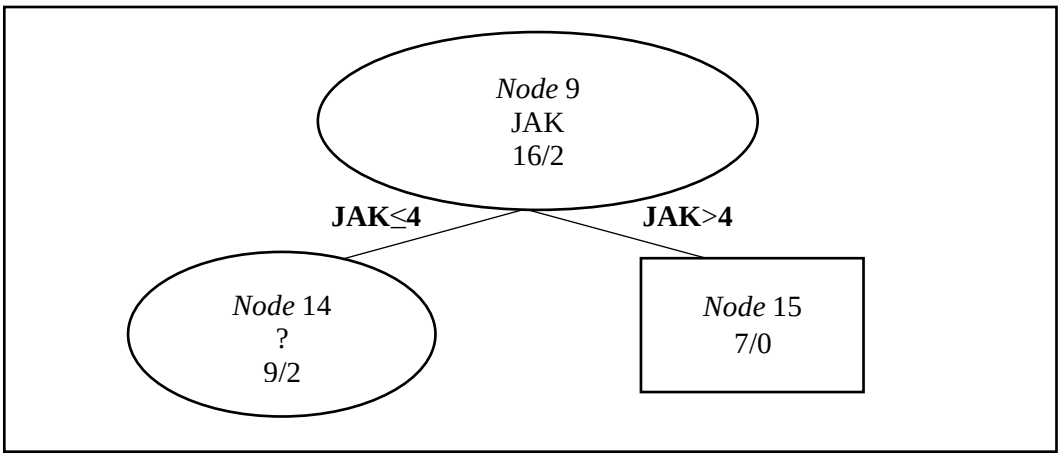
Gambar 4.17 Hasil Pembentukan Cabang di *Node* 8

Penentuan pemilah terbaik untuk *node* 9 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 5 dan 8 yakni KK dengan pendidikan terakhir SMP sebanyak 18 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.41.

Tabel 4.41 Hasil Perhitungan Indeks *Gini* untuk *Node* 9

Variabel	Kategori	Indeks *Gini*
Pekerjaan	{Petani} dan {PNS, Swasta, Wiraswasta}	0,1880
	{PNS} dan {Petani, Swasta, Wiraswasta}	0,1961
	{Swasta} dan {Petani, PNS, Wiraswasta}	0,1961
	{Wiraswasta} dan {Petani, PNS, Swasta}	0,1926
	{Petani, PNS} dan {Swasta, Wiraswasta}	0,1905
	{Petani, Swasta} dan {PNS, Wiraswasta}	0,1905
	{Petani, Wiraswasta} dan {PNS, Swasta}	0,1944
Jumlah Anggota Keluarga	**({>4}) dan ({≤4})**	**0,1818**
Jenis Kelamin	{Perempuan, Laki-laki}	0,1961

Berdasarkan Tabel 4.41 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Jumlah Anggota Keluarga dengan pemilah {(>4), (≤4)} sebesar 0,1818. Maka pemilah ini terpilih menjadi pemilah untuk *node* 8 kemudian jumlah anggota keluarga >4 menjadi *node terminal* dikarenakan sampel hanya berada di salah satu kelas. Maka *node* 9 menjadi seperti pada Gambar 4.18 berikut :



Gambar 4.18 Hasil Pembentukan Cabang di *Node* 9

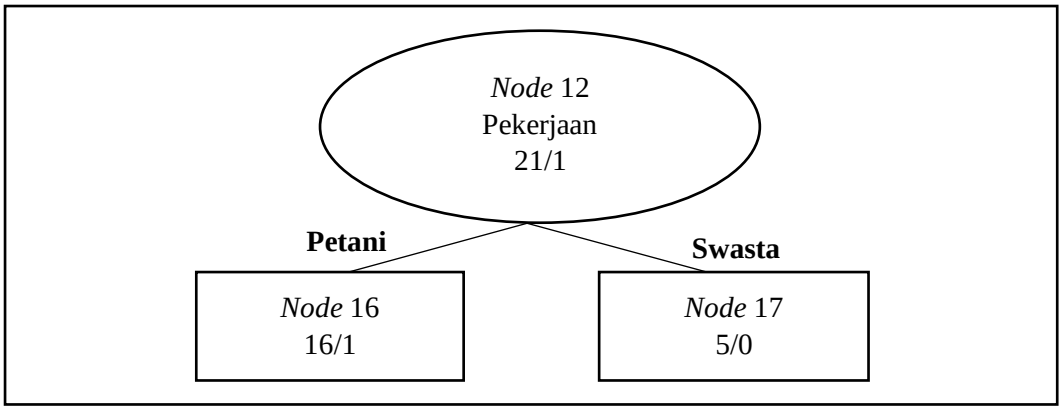
Penentuan pemilah terbaik untuk *node* 12 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 5, 9 dan 13 yakni KK dengan jumlah anggota keluarga >4 sebanyak 22 KK. Perhitungan untuk

menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.42 berikut:

Tabel 4.42 Hasil Perhitungan Indeks *Gini* untuk *Node* 12

Variabel	**Kategori**	**Indeks *Gini***
Pekerjaan	({Petani}, {Swasta})	0,1880

Berdasarkan Tabel 4.42 dapat dilihat bahwa hanya terdapat 1 pemilah saja pada variabel Pekerjaan dengan pemilah {(Petani), (Swasta)} dengan nilai *indeks gini* sebesar 0,1880. Karena hanya tersisa 1 pemilah, maka dapat dipastikan bahwa *node* cabang menjadi *node* terminal. Maka *node* 12 menjadi seperti pada Gambar 4.19 berikut:



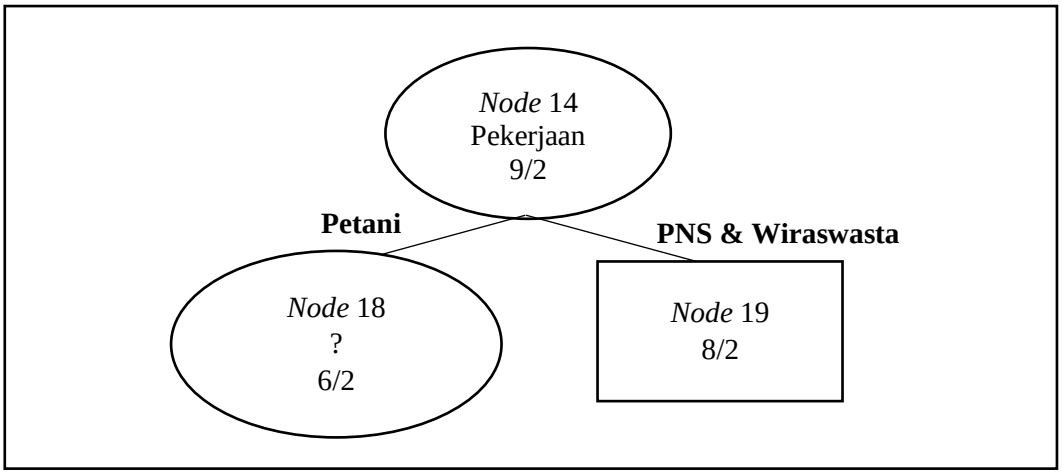
Gambar 4.19 Hasil Pembentukan Cabang di *Node* 12

Penentuan pemilah terbaik untuk *node* 14 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 5, 8 dan 15 yakni KK dengan jumlah anggota keluarga ≤4 sebanyak 11 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.43.

Tabel 4.43 Hasil Perhitungan Indeks *Gini* untuk *Node* 14

Variabel	Kategori	Indeks *Gini*
Pekerjaan	**{Petani} dan {PNS, Wiraswasta}**	**0,2727**
	{PNS} dan {Petani, Wiraswasta}	0,2909
	{Wiraswasta} dan {Petani, PNS}	0,2828
Jenis Kelamin	{Perempuan, Laki-laki}	0,2909

Berdasarkan Tabel 4.43 dapat dilihat bahwa pemilah yang memiliki nilai indeks *gini* terkecil adalah pada variabel Pekerjaan dengan pemilah {(Petani), (PNS, Wiraswasta)} sebesar 0,2727. Maka pemilah ini terpilih menjadi pemilah untuk *node* 14 kemudian pekerjaan PNS dan Wiraswasta menjadi *node terminal* dikarenakan sampel hanya berada di salah satu kelas. Maka *node* 14 menjadi seperti pada Gambar 4.20 berikut :



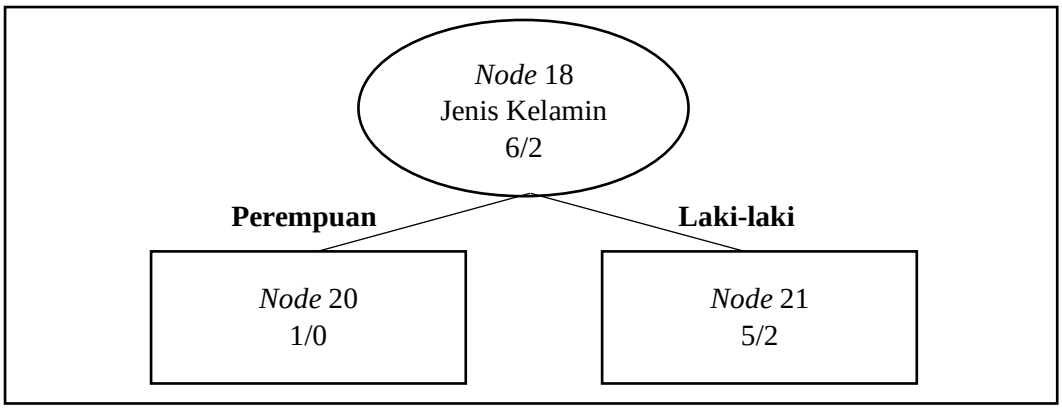
Gambar 4.20 Hasil Pembentukan Cabang di *Node* 14

Penentuan pemilah terbaik untuk *node* 18 menggunakan cara seperti mencari pemilah pada *node* 1 (*node* akar) namun data yang digunakan adalah sisa data terhadap komposisi kelas yang tidak masuk ke dalam *node* 3, 5, 8, 15 dan 19 yakni KK dengan pekerjaan Petani sebanyak 8 KK. Perhitungan untuk menentukan *node* selanjutnya menggunakan Persamaan (2.7) untuk mencari nilai indeks *gini*. Adapun hasil perhitungan indeks *gini* pada setiap kemungkinan pemilah untuk setiap variabel bebas disajikan seperti pada Tabel 4.44.

Tabel 4.44 Hasil Perhitungan Indeks *Gini* untuk *Node* 18

Variabel	Kategori	Indeks *Gini*
Jenis Kelamin	({Perempuan}, {Laki-laki})	0,3571

Berdasarkan Tabel 4.44 dapat dilihat bahwa hanya terdapat 1 pemilah saja pada variabel Jenis Kelamin dengan pemilah {(Perempuan), (Laki-laki)} dengan nilai *indeks gini* sebesar 0,3571. Karena hanya tersisa 1 pemilah, maka dapat dipastikan bahwa *node* cabang menjadi *node* terminal. Maka *node* 18 menjadi seperti pada Gambar 4.21 berikut:



Gambar 4.21 Hasil Pembentukan Cabang di *Node* 18

4.4.1.2 Penentuan *Node* Terminal

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. Suatu *node* juga akan menjadi *node* terminal apabila sampel hanya berada disatu kelas saja.

4.4.1.3 Penandaan Label Kelas

Pemberian label kelas untuk setiap *node* terminal menggunakan persamaan (2.11). Oleh karena *node* 5 merupakan *node* terminal, maka langkah selanjutnya adalah pemberian label kelas menggunakan persamaan (2.8).

Berikut contoh pemberian label kelas untuk *node* 5 :

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

$$P(\text{Rata-rata pendapatan} < 2{,}89 \text{ juta} \mid 5) = \frac{0}{1} = 0$$

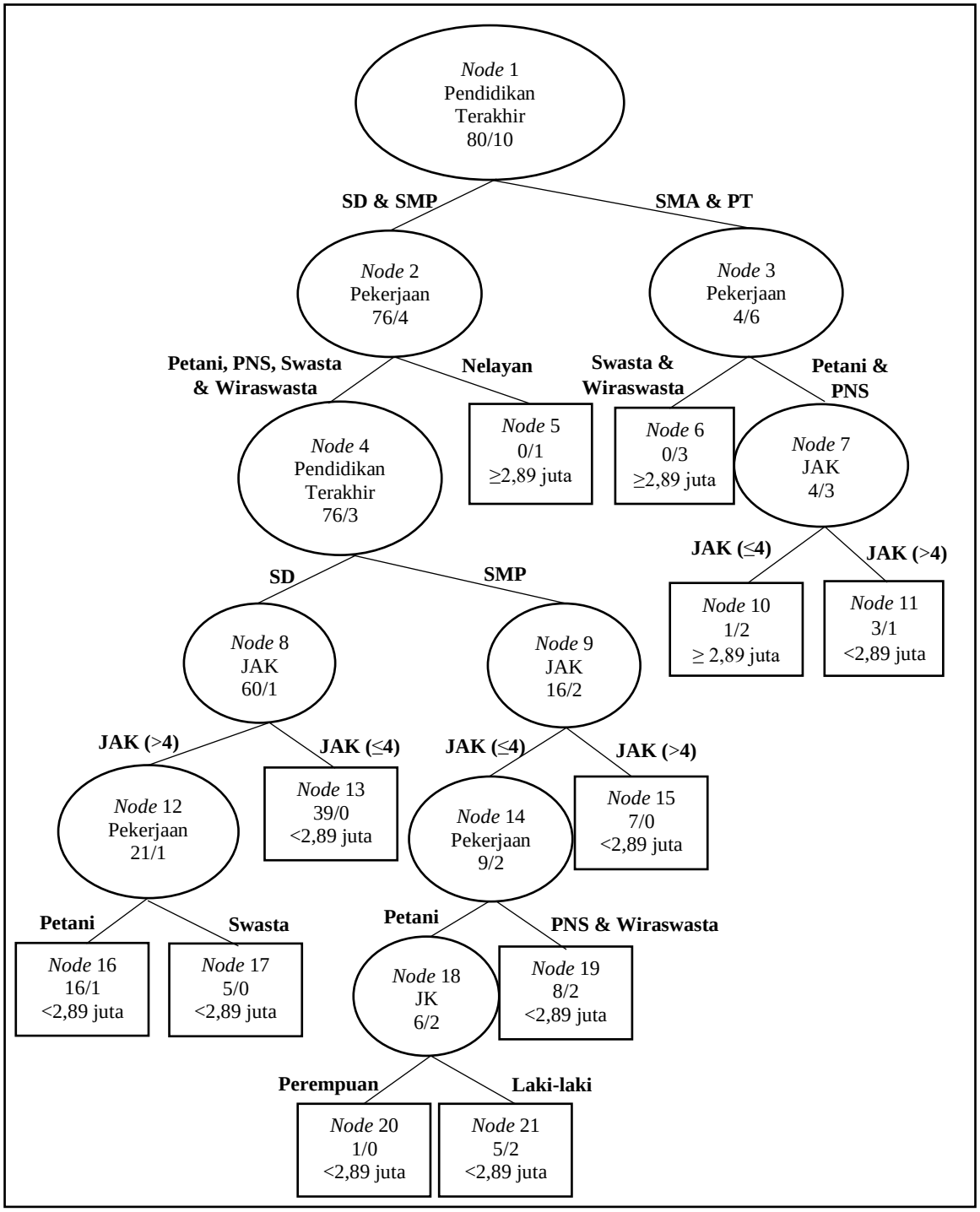
$$P(\text{Rata-rata pendapatan} \geq 2{,}89 \text{ juta} \mid 5) = \frac{1}{1} = 1$$

karena P(Rata-rata pendapatan ≥2,89 juta|5) > P(Rata-rata pendapatan <2,89 juta|5) maka *node* 5 diberi label kelas rata-rata pendapatan ≥ 2,89 juta. Adapun hasil perhitungan peluang tiap kelas rata-rata pendapatan pada masing-masing *node* terminal disajikan pada Tabel 4.45.

Tabel 4.45 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 4.22.



Gambar 4.22 Pohon Klasifikasi CART

Pada Gambar 4.22 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. 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 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.

4.5 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. Sebagai contoh menggunakan porporsi data 90:10 sehingga pengujian menggunakan data *testing* sebanyak 10 sampel yang dapat dilihat pada Lampiran 3.

4.5.1 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 4.46 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 4.46 hasil dari *confusion matrix* untuk algoritma C5.0 dengan proporsi data 90:10 sehingga dapat dihitung tingkat ketepatan akurasi menggunakan Persamaan (2.10) seperti berikut :

$$Akurasi = \frac{9+0}{9+0+1+0} \times 100\%$$
$$= \frac{9}{10} \times 100\% = 90\%$$

Tingkat akurasi sebesar 90% menyatakan bahwa dari 90 KK, terdapat 81 orang yang tepat diklasifikasikan.

4.5.2 Ketepatan Klasifikasi Algoritma CART

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

Tabel 4.47 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 4.47 hasil dari *confusion matrix* untuk algoritma CART dengan proporsi data 90:10 sehingga dapat dihitung tingkat ketepatan akurasi menggunakan Persamaan (2.10) seperti berikut :

$$Akurasi = \frac{8+0}{8+1+1+0} \times 100\%$$
$$= \frac{8}{10} \times 100\% = 80\%$$

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

4.5.3 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 4.48 berikut:

Tabel 4.48 Perbandingan Tingkat Akurasi Kedua Metode untuk Setiap Proporsi

	50:50	**60:40**	**70:30**	**80:20**	**90:10**	**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 4.48 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.

BAB 5
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang dilakukan, diperoleh kesimpulan sebagai berikut :

1. 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%.
2. Hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 menggunakan metode CART dengan porporsi 50:50 memperoleh tingkat akurasi tertinggi yaitu sebesar 94%.
3. 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.

5.2 Saran

Saran yang dapat penulis berikan pada penelitian ini yaitu:

1. Pada penelitian selanjutnya dapat menggunakan variabel dan data yang lebih banyak.
2. Dapat dibandingkan dengan metode klasifikasi yang lain, misalnya metode algoritma C5.0 dengan *Classification Rule with Unbiased Interaction Selection and Estimation* (CRUISE) dan metode *Classification and Regression Tree* (CART) dengan *Quick Unbiased Efficient Statistical Tree* (QUEST).

DAFTAR PUSTAKA

Akbar, M. S., Yuanita, D. dan Harini, S. (2010). Pendekatan CART untuk Mendapatkan Faktor yang Mempengaruhi Terjangkitnya Penyakit Demam Tifoid di Aceh Utara. *Jurnal CAUCHY.* 1(2), 2086-0382.

Boediono. (2002). *Pengantar Ekonomi*. Jakarta: Erlangga.

Kusrini dan Luthfi, E. T. (2009). *Algoritma Data Mining.* Yogyakarta: Penerbit Andi.

Larose. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Willey & Sons.

Mardiani. (2012). Perkembangan Algoritma untuk Menghitung Pola yang Sering Muncul pada Basis Data yang Besar. *Seminar Aplikasi Teknologi Informasi*, 1907-5022.

Pakpahan, H. S, Indar, F dan Wati, M. (2018). Penerapan Algoritma CART Decision Tree pada Penentuan Penerima Program Bantuan Pemerintah Daerah Kabupaten Kutai Kartanegara. *JURTI*, 2(1), 2579-8790.

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

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

_____. (2014). *Data Mining-Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit Andi.

Pratiwi, F. E. dan Zain, I. (2014). Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara. *Jurnal Sains dan Seni Pomits,* 3(1), 2337-3520.

Putri, Y. R., Mukhlash, I. dan Hidayat, N. (2013). Prediksi Pola Kecelakaan Kerja pada Perusahaan Non Ekstraktif Menggunakan Algoritma Decision Tree: C4.5 dan C5.0. *Jurnal Sains dan Seni Pomits,* 2(1), 2337-3520.

Reksoprayitno. (2004). *Sistem Ekonomi dan Demokrasi Ekonomi*. Jakarta: Bina Grafika.

Rosni. (2012). Analisis Tingkat Kesejahteraan Masyarakat Nelayan di Desa Dahari Selebar Kecamatan Talawi Kabupaten Batubara. *Jurnal Geografi*. 2549-7057.

Soekartawi. (2002). *Faktor-faktor Produksi*. Jakarta: Salemba Empat.

Sudarman, T. (2001). *Ekonomi Indonesia*. Jakarta: Raja Grafindo.

Sugiyono. (2010). *Statistika untuk Penelitian*. Bandung: Alfabeta.

Sumiarni, E. (2005). *Kajian Hukum Perkawinan Yang Berkesetaraan Jender*. Yogyakarta: Wonderful Publishing Company.

Wijaya, A. C, Hasibuan, N. A dan Ramadhani, P. (2018). Implementasi Algoritma C5.0 dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus: Kelurahan Mesjid Kecamatan Medan Kota). *Majalah Ilmiah INTI.* 13(2), 2339-210X.

Yusuf, Y. W. (2007). Perbandingan Performansi Algoritma Decision Tree C5.0, CART dan CHAID: Kasus Prediksi Status Resiko Kredit Bank X. *Seminar Nasional Aplikasi Teknologi Informasi.* 1907-5022.

LAMPIRAN

Lampiran 1. Kuesioner *Social Mapping*

KUESIONER
Social Mapping

Kami mahasiswa/i Universitas Mulawarman dalam rangka memenuhi tugas kuliah dalam kegiatan Kuliah Kerja Nyata (KKN) angkatan 45 tahun 2019 ingin melakukan penelitian mengenai data sosial Desa Teluk Baru untuk memenuhi tugas *Social Mapping* yang diberikan oleh pihak Lembaga Penelitian dan Pengabdian Kepada Masyarakat (LP2M). Berikut ini adalah kuesioner yang berkaitan dengan tugas *Social Mapping* untuk Kepala Keluarga (KK). Oleh karena itu disela-sela kesibukan anda, kami memohon dengan hormat kesediaan anda untuk dapat mengisi kuesioner ini. Atas kesediaan dan partisipasi anda kami ucapkan terimakasih.

A. Identitas Responden

Nama :
Jenis Kelamin :
No HP :

B. Daftar Pertanyaan

Petunjuk Pengisian :
Ceklislah (√) salah satu pilihan jawaban pertanyaan (nomor 1 dan 2) dan isilah titik-titik (nomor 3 dan 4) dibawah ini sesuai dengan kejadian yang sebenarnya.

1. Pendidikan Terakhir : ☐ SD ☐ SMA
☐ SMP ☐ Perguruan Tinggi

2. Pekerjaan : ☐ Petani ☐ Nelayan ☐ PNS
☐ Swasta ☐ Wiraswasta
☐ lainnya …..

3. Jumlah Anggota Keluarga : …… orang

4. Rata-rata Pendapatan Perbulan : Rp. ……

Lampiran 2. Data Sosial KK di Desa Teluk Baru Tahun 2019

No	Rata-rata Pendapatan Perbulan (Y)	Pekerjaan (X_1)	Jumlah Anggota Keluarga (X_2)	Pendidikan Terakhir (X_3)	Jenis Kelamin (X_4)
1	<2,89 juta	Petani	>4	SMP/sederajat	Laki-laki
2	≥2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
3	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
4	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
5	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
6	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
7	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
8	<2,89 juta	Petani	≤4	SD/sederajat	Perempuan
9	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
10	<2,89 juta	Petani	>4	SMP/sederajat	Laki-laki
11	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
12	<2,89 juta	Petani	≤4	SMP/sederajat	Perempuan
13	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
14	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
15	≥2,89 juta	Petani	>4	SD/sederajat	Laki-laki
16	≥2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
17	≥2,89 juta	Petani	>4	SMA/sederajat	Laki-laki
18	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
19	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
20	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
21	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
22	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
23	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
24	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
25	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
26	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki

Lampiran 2. Data Sosial KK di Desa Teluk Baru Tahun 2019 (Lanjutan)

No	Rata-rata Pendapatan Perbulan (Y)	Pekerjaan (X_1)	Jmlah Anggota Keluarga (X_2)	Pendidikan Terakhir (X_3)	Jenis Kelamin (X_4)
27	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
28	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
29	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
30	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
31	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
32	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
33	<2,89 juta	Swasta	≤4	SD/sederajat	Laki-laki
34	≥2,89 juta	PNS	>4	SMA/sederajat	Laki-laki
35	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
36	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
37	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
38	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
39	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
40	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
41	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
42	<2,89 juta	Petani	>4	SMP/sederajat	Laki-laki
43	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
44	<2,89 juta	Swasta	>4	SMP/sederajat	Laki-laki
45	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
46	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
47	<2,89 juta	Petani	>4	SMP/sederajat	Laki-laki
48	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
49	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
50	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
51	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
52	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki

Lampiran 2. Data Sosial KK di Desa Teluk Baru Tahun 2019 (Lanjutan)

No	Rata-rata Pendapatan Perbulan (Y)	Pekerjaan (X_1)	Jmlah Anggota Keluarga (X_2)	Pendidikan Terakhir (X_3)	Jenis Kelamin (X_4)
53	≥2,89 juta	Wiraswasta	≤4	Perguruan Tinggi (PT)	Laki-laki
54	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
55	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
56	≥2,89 juta	Swasta	>4	SMA/sederajat	Laki-laki
57	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
58	≥2,89 juta	Swasta	≤4	SMA/sederajat	Laki-laki
59	<2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
60	<2,89 juta	Swasta	≤4	SD/sederajat	Laki-laki
61	<2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
62	<2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
63	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
64	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
65	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
66	<2,89 juta	PNS	≤4	SMP/sederajat	Laki-laki
67	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
68	<2,89 juta	PNS	≤4	SMA/sederajat	Perempuan
69	<2,89 juta	Wiraswasta	>4	SMP/sederajat	Laki-laki
70	<2,89 juta	Petani	≤4	SMA/sederajat	Laki-laki
71	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
72	<2,89 juta	Petani	>4	SMA/sederajat	Laki-laki
73	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
74	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
75	<2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
76	<2,89 juta	Swasta	≤4	SD/sederajat	Laki-laki
77	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
78	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki

Lampiran 2. Data Sosial KK di Desa Teluk Baru Tahun 2019 (Lanjutan)

No	Rata-rata Pendapatan Perbulan (Y)	Pekerjaan (X_1)	Jmlah Anggota Keluarga (X_2)	Pendidikan Terakhir (X_3)	Jenis Kelamin (X_4)
79	≥2,89 juta	Nelayan	≤4	SD/sederajat	Laki-laki
80	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
81	<2,89 juta	Petani	>4	SMP/sederajat	Laki-laki
82	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
83	<2,89 juta	Wiraswasta	≤4	SMP/sederajat	Laki-laki
84	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki
85	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
86	<2,89 juta	Petani	≤4	SMA/sederajat	Laki-laki
87	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
88	<2,89 juta	Wiraswasta	≤4	SMP/sederajat	Laki-laki
89	<2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
90	≥2,89 juta	Petani	≤4	SMA/sederajat	Laki-laki
91	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
92	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
93	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
94	<2,89 juta	Petani	≤4	SMA/sederajat	Laki-laki
95	<2,89 juta	Swasta	>4	SMP/sederajat	Laki-laki
96	<2,89 juta	Wiraswasta	>4	SMP/sederajat	Laki-laki
97	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
98	≥2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
99	<2,89 juta	Swasta	≤4	SD/sederajat	Perempuan
100	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki

Lampiran 3. Data *Testing* untuk Proporsi Data 90:10

No	Rata-rata Pendapatan Perbulan (Y)	Pekerjaan (X_1)	Jumlah Anggota Keluarga (X_2)	Pendidikan Terakhir (X_3)	Jenis Kelamin (X_4)
1	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
2	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
3	<2,89 juta	Petani	≤4	SD/sederajat	Laki-laki
4	<2,89 juta	Petani	≤4	SMA/sederajat	Laki-laki
5	<2,89 juta	Swasta	>4	SMP/sederajat	Laki-laki
6	<2,89 juta	Wiraswasta	>4	SMP/sederajat	Laki-laki
7	<2,89 juta	Petani	>4	SD/sederajat	Laki-laki
8	≥2,89 juta	Swasta	>4	SD/sederajat	Laki-laki
9	<2,89 juta	Swasta	≤4	SD/sederajat	Perempuan
10	<2,89 juta	Petani	≤4	SMP/sederajat	Laki-laki

Lampiran 4. Sintaks *Software R* untuk membuat diagram lingkaran

```
data<-read.table("D://vary.txt",header=T)
data
pie(data$Jumlah,labels=data$Jumlah, main="Rata-rata Pendapatan   Perbulan",
    col=heat.colors(2))
colors=heat.colors(2)
legend(1,0.5, c("<2,89 juta",">=2,89 juta"),cex=1.0,fill=colors)
```

RIWAYAT HIDUP

Reni Pratiwi, lahir di Embalut, Kalimantan Timur pada tanggal 05 September 1997. Merupakan anak pertama dari lima bersaudara, dari pasangan Bapak Zainal Aripin dan Ibu Lili Herlina.

Memulai pendidikan formal pada tahun 2002 di TK YPPSB dan lulus pada tahun 2004. Pendidikan Sekolah Dasar dimulai pada tahun 2004 di SD YPPSB 1 dan lulus pada tahun 2010. Kemudian melanjutkan pendidikan di MTs Plus Darul Ulum Jombang pada tahun 2010 dan lulus tahun 2013. Pada tahun 2013 pendidikan dilanjutkan di MAN 2 Samarinda dengan bidang Ilmu Pengetahuan Alam (IPA) dan dinyatakan lulus pada tahun 2016.

Pendidikan di perguruan tinggi dimulai pada tahun 2016 di Program Studi Statistika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman. Pada tahun 2019 melaksanakan program Kuliah Kerja Nyata (KKN) Angkatan 45 di desa Teluk Baru, Kecamatan Muara Ancalong, Kabupaten Kutai Timur. Pada tahun yang sama juga melaksanakan Praktek Kerja Lapangan (PKL) di Dinas Pariwisata Kota Samarinda yang dilakukan selama 40 hari kerja yaitu tanggal 8 Oktober 2019 hingga 2 Desember 2019.