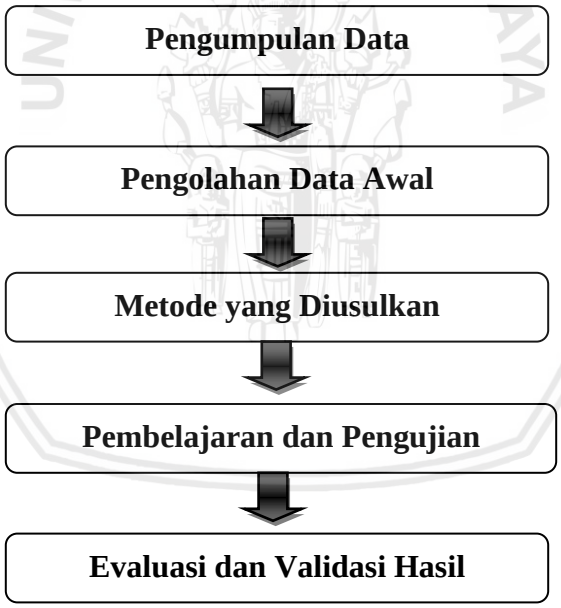
BAB IV
METODE PENELITIAN

Penelitian ini dilakukan dengan menggunakan laptop HP 2700 series dengan *processor* Intel Core i5-7200U @ 2,5 GHz 2.71 GHz, memory 4 GB, harddisk 1 TB, dan menggunakan *system* operasi Windows 10 64-bit. Pengujian ini menggunakan *microsoft excel* 2010 untuk memproses *dataset* dan perangkat lunak *Rapid miner* 5.3.0 untuk merancang dan menganalisa hasil penghitungan metode selain itu MATLAB digunakan sebagai validasi hasil. Adapun tahapan penelitian sebagai berikut:

4.1 Tahapan Penelitian

Metode penelitian memberikan gambaran, arah dan tahapan-tahapan yang akan dilakukan pada proses penelitian ini, sehingga proses yang dilakukan sesuai dengan tujuan dari penelitian.



Gambar 4.1 Kerangka solusi masalah
Sumber: Hasil Penelitian (2013)

4.2 Data Penelitian

Data yang digunakan dalam penelitian ini berasal dari data *open source* yaitu UCI *Machine Leaming Repository* dengan link http://mlr.cs.umass.edu/ml/. Data penelitian yang

digunakan antara lain: *Glass*, *Lympografi*, *Vehicle*, *Thyroid* dan *Wine*. Data klasifikasi tersebut banyak digunakan oleh peneliti untuk melakukan pengujian terhadap keandalan metode klasifikasi pada bidang kesehatan dan *manufacturin*g. Selain itu data dipilih karena data bersifat *imbalanced* dan *multiclass*. *Dataset* pada penelitian ini juga memiliki tipe nilai karakteristik atribut yang berbeda sehingga baik digunakan untuk pengujian hipotesis. Tabel 4.1 merupakan deskripsi mengenai *dataset* yang digunakan dalam penelitian.

Tabel 4.1 *Dataset* Penelitian

No.	***Dataset***	**Jumlah Data**	**Jumlah *Atribut***	**Jumlah *Class***	**Jumlah *Class Mayor***	**Jumlah *Class Minor***
1	Glass	214	10	6	2	4
2	Lympografi	148	18	4	2	2
3	vehicle	946	18	4	3	1
4	Thyroid	215	5	3	1	2
5	Wine	178	13	3	1	2

Tabel 4.2 Karakteristik *Dataset* Penelitian

No.	***Dataset***	***Class***						***Type Atribut***
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	
1	Glass	87	70	17	13	9	29	Real, Interger, Numeric
2	Lympografi	2	2	61	81	-	-	Interger
3	vehicle	240	240	240	226	-	-	Interger
4	Thyroid	150	35	30	-	-	-	Interger, Numeric
5	Wine	59	71	48	-	-	-	Numeric, Real, Interger

4.2.1 Karakteristik Data Penelitian

Setiap data memiliki karakteristik tipe dan sifat yang berbeda. Hal tersebut berpengaruh dalam menentukan metode yang digunakan dalam menemukan pola data. Pada Tabel 4.3 menjelaskan tentang macam tipe dan sifat pada data, antara lain:

Tabel 4.3 Tipe dan sifat karakteristik data

Type Atribut	Keterangan
Nominal	Tipe data berupa simbol atau nama-nama dari suatu benda
Numerik	Tipe data yang digunakan pada variabel atau konstanta untuk menyimpan nilai dalam bentuk bilangan atau angka (*Interger, Real*)
Interger	Tipe data yang merepresentasikan bilangan bulat
Real	Tipe data yang digunakan merepresentasikan angka desimal
Text	Data yang bersifat *string* atau karakter
Binomial	Data yang hanya memiliki dua jenis karakter atau variabel
Polinomial	Data yang memiliki banyak jenis karakter atau variabel
Date_time	Data yang berkaitan dengan tanggal dan waktu
Sifat Data	**Keterangan**
Kontinue	Data yang sifatnya sinambung atau kontinyu, nilainya bisa berupa pecahan
Diskrit	Data yang sifatnya terputus-putus, nilainya bukan merupakan pecahan (angka utuh)

4.2.2 Identifikasi Variabel

Data jenis klasifikasi memiliki dua jenis variabel yaitu variabel *atribut* dan variabel *class* prediksi. Variabel *atribut* dinotasikan X_n mendiskripsikan tentang karakteristik sebuah data sedangkan variabel *class* prediksi dinotasikan Y_n mendiskripsikan tentang hasil prediksi *class* data. Berikut ini penjelasan tentang karakteristik *dataset* yang digunakan pada penelitian ini.

4.2.2.1 *Dataset Glass*

Dataset glass diamati oleh vina spihler, Ph.D dari *diagnostic product corporation.* Data ini mengenai klasifikasi jenis kaca berdasarkan proses pembuatan. Data ini bersifat kontinue berjumlah 214 terdiri dari *class mayor* berjumlah dua dan *class minor* berjumlah empat. *Atribut* sebanyak 10 dinotasikan (X_n) antara lain Id *number*, RI, Na, Mg, Al, Si, K, Ca, Ba, Fe dan label *class* berjumlah 7 bertipe *Polinomial* dinotasikan (Y_n) antara lain *building_windows_float_processed, building_windows_non_float_processed,*

*vehicle_windows_float_processed, vehicle_windows_non_float _processed, containers, tableware, headla*mps. Tabel 4.4 merupakan karakteristik dari variabel data *glass*:

Tabel 4.4 Karakteristik Variabel *Dataset Glass*

Var	Nama	Tipe	Deskripsi
Y1	Building_windows_float_processed	Polinomial	Kaca bangunan menggunakan proses float (memiliki data sebanyak 87)
Y2	Building_windows_non_float_processed	Polinomial	Kaca bangunan tanpa menggunakan proses float (memiliki data sebanyak 70)
Y3	Vehicle_windows_float_processed	Polinomial	Kaca kendaraan menggunakan proses float (memiliki data sebanyak 17)
Y4	Vehicle_windows_non_float_processed	Polinomial	Kaca kendaraan tanpa menggunakan proses float (memiliki data sebanyak 0)
Y5	Containers	Polinomial	Wadah (memiliki data sebanyak 13)
Y6	Tableware	Polinomial	Jenis kaca digunakan untuk Peralatan makan (memiliki data sebanyak 9)
Y7	Headlamps	Polinomial	Jenis kaca digunakan untuk Lampu depan kendaraan (memiliki data sebanyak 29)
X1	Id Number	Interger	Nomer dataset (1 - 214)
X2	Ri	Real	refractive index (min=1.5112, max=1.5339 mean=1.5184)
X3	Na	Real	Sodium (unit measurement: weight percent in corresponding oxide, as are atributes 4-10) (min=1.5112, max=1.5339 mean=1.5184)
X4	Mg	Numeric	Magnesium (min=10.73, max=17.38, mean=13.4079)
X5	Al	Real	Aluminum (min=0.29, max=3.5, mean=1.4449)
X6	Si	Real	Silicon (min=69.81, max=75.41, mean=72.6509)
X7	K	Numeric	Potassium (min=0, max=6.21, mean=0.4971)
X8	Ca	Real	Calcium(min=5.43, max=16.19, mean=8.9570)

X9	Ba	Numeric	Barium (min=0, max=3.15, mean=0.1750)
X10	Fe	Numeric	Iron (min=0, max=0.51, mean=0.0570)

4.2.2.2 *Dataset Lympografi*

Dataset lympografi diamati oleh Dr. William H. Wolberg dari *General Surgery Dept., University of Wisconsin.* Data ini mengenai tes yang memanfaatkan teknologi x-ray untuk melihat sirkulasi limfatik dan kelenjar getah bening untuk tujuan diagnosa. Data ini bersifat diskrit berjumlah 148 terdiri dari *class mayor* berjumlah dua dan *class minor* berjumlah dua. *Atribut* sebanyak 18 dinotasikan (X_n) antara lain *lymphatics, block of affere, bl. Of lymph. c,bl. Of lymph. S, by pass, extravasates, regeneration, early uptake in, lym nodes dimin, lym nodes enlar, changes in lym, defect in node, changes in node, changes in stru, special forms, dislocation, exclusion of no, no of nodes in* dan label *class* berjumlah 4 bertipe *Polinomial* dinotasikan (Y_n) antara lain *nomal find, metastases, malign lymph, fibrosis*. Tabel 4.5 merupakan karakteristik dari variabel data *glass*:

Tabel 4.5 Variabel *Dataset Lympografi*

Var	Nama	Tipe	Deskripsi
Y1	Nomal find	Polinomial	Jaringan normal (memiliki data sebanyak 2)
Y2	Metastases	Polinomial	Penyebaran kanker awal (memiliki data sebanyak 81)
Y3	Malign lymph	Polinomial	Penyebaran di Kelenjan lymfe (memiliki data sebanyak 61)
Y4	Fibrosis	Polinomial	Penyebaran di Jaringan fibrin (memiliki data sebanyak 4)
X1	Lymphatics	Interger	1=normal, 2=arched, 3=deformed, 4=displaced
X2	Block of affere	Interger	N0=1. Yes=2
X3	Bl. Of lymph. C	Interger	Kandungan bl. Of lymph. C (no=1. Yes=2)
X4	Bl. Of lymph. S	Interger	Kandungan bl. Of lymph. S (no=1. Yes=2)

X5	By pass	Interger	Kandungan by pass (no=1. Yes=2)
X6	Changes in lym	Interger	Kandungan changes in lym (1=bean, 2=oval, 3=round)
X7	Defect in node	Interger	Kandungan changes in lym (1=no, 2=lacunar, 3=lac. marginal, 4=lac. central)
X8	Changes in node	Interger	Kandungan changes in node (1=no, 2=lacunar, 3=lac. margin, 4=lac. central)
X9	Changes in stru	Interger	Kandungan changes in stru (1=no, 2=grainy, 3=drop-like, 4=coarse, 5=diluted, 6=reticular, 7=stripped, 8=faint)
X10	Special forms	Interger	Kandungan special forms (1=no, 2=chalices, 3=vesicles)
X11	dislocation	Interger	Kandungan dislocation (no=1. Yes=2)
X12	exclusion of no	Interger	Kandungan exclusion of no (no=1. Yes=2)
X13	no of nodes in	Interger	Kandungan no of nodes in (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70) (min=181, max=211, mean=195.63)
X14	extravasates	Interger	Kandungan extravasates (no=1. Yes=2)
X15	regeneration	Interger	Kandungan regeneration (no=1. Yes=2)
X16	early uptake in	Interger	Kandungan early uptake in (no=1. Yes=2)
X17	lym nodes dimin	Interger	Kandungan lym nodes dimin (no=1. Yes=2)

4.2.2.3 *Dataset Vehicle*

Dataset vehicle diamati oleh Drs.Pete Mowforth and Barry Shepherd dari *Turing Institute, Glasgow, Scotla*nd. Data ini mengenai klasifikasi jenis kendaraan berdasarkan sudut pandang gambar yang berbeda. Data ini bersifat kontinue berjumlah 946 terdiri dari *class mayor* berjumlah tiga dan *class minor* berjumlah satu. *Atribut* sebanyak 18 dinotasikan (X_n) antara lain *compactness, circularity, distance circularity, radius ratio, pr. Axis aspect ratio, max length aspect ratio, scatter ratio, elongatedness, pr axis rectangularity, max length rectangularity, scaled variance, scaled variance, scaled radius of gyration, skewness about,*

skewness about, kurtosis about, kurtosis about, hollows ratio dan label *class* berjumlah 4 bertipe *Polinomial* dinotasikan (Y_n) antara lain *OPEL, SAAB, BUS, V*AN. Tabel 4.6 merupakan karakteristik dari variabel data *vehicle*:

Tabel 4.6 Variabel *Dataset Vehicle*

Var	Nama	Tipe	Deskripsi
Y1	OPEL	Polinomial	Merk kendaraan OPEL (memiliki data sebanyak 240)
Y2	SAAB	Polinomial	Merk kendaraan SAAB (memiliki data sebanyak 240)
Y3	BUS	Polinomial	Kendaraan jenis Bus (memiliki data sebanyak 240)
Y4	VAN	Polinomial	Kendaraan jenis Van (memiliki data sebanyak 226)
X1	Compactness	Interger	Nilai Keserasian komponen (average perim)**2/area (min=73, max=119, mean=93.67)
X2	Circularity	Interger	Berbentuk bundar (average radius)**2/area (min=33, max=59, mean=44.86)
X3	Distance circularity	Interger	Jarak lingkaran area/(av.distance from border)**2 (min=40, max=112, mean=82.08)
X4	Radius ratio	Interger	Nilai radius ratio (max.rad-min.rad)/av.radius (min=104, max=333, mean=168.94)
X5	Pr.axis aspect ratio	Interger	(minor axis)/(major axis) (min=47, max=138, mean=61.69)
X6	Max.length aspect ratio	Interger	(length perp. max length)/(max length) (min=2, max=55, mean=8.57)
X7	Scatter ratio	Interger	inertia about minor axis)/(inertia about major axis) (min=112, max=265, mean=168.84)
X8	Elongatedness	Interger	area/(shrink width)**2 (min=26, max=61, mean=40.93)
X9	Pr.axis rectangularity	Interger	area/(pr.axis length*pr.axis width) (min=17, max=29, mean=20.59)

X10	Max.lengthrectangularity	Interger	area/(max.length*length perp. to this) (min=118 max=188, mean=148)
X11	Scaled variance major	Interger	(2nd order moment about minor axis)/area (min=130, max=320, mean=188.63)
X12	Scaled variance minor	Interger	(2nd order moment about major axis)/area (min=184, max=1018, mean=439,91)
X13	Scaled radius of gyration	Interger	(mavar+mivar)/area (min=109, max=268, mean=174,70)
X14	Skewness about major	Interger	(3rd order moment about major axis)/sigma_min**3 (min=59, max=135, mean=72.46)
X15	Skewness about minor	Interger	(3rd order moment about minor axis)/sigma_maj**3 (min=0, max=22, mean=6.38)
X16	Kurtosis about minor	Interger	(4th order moment about major axis)/sigma_min**4 (min=0, max=41, mean=12.60)
X17	Kurtosis about major	Interger	(4th order moment about minor axis)/sigma_maj**4(min=176, max=206, mean=188.93)
X18	Hollows ratio	Interger	(area of hollows)/(area of bounding polygon) (min=181, max=211, mean=195.63)

4.2.2.4 *Dataset Thyroid*

Dataset thyroid diamati oleh Danny Coomans dari *Dept. of Maths. and Stats., James Cook University*. *Thyroid* merupakan kanker kelenjar pada leher. Data ini digunakan untuk memprediksi pasien *tiroid* berdasarkan *class eutiroidisme, hipotiroidisme* atau *hipertiroidisme*. Diagnosis (prediksi *class*) didasarkan pada rekam medis lengkap, termasuk *anamnesis*, hasil scan dll. Data ini bersifat kontinue berjumlah 215 terdiri dari *class mayor* berjumlah satu dan *class minor* berjumlah dua. *Atribut* sebanyak 5 dinotasikan (X_n) antara lain *T3-resin uptake test, total serum thyroxin, total serum triiodothyronine, basal thyroid-stimulating hormone (TSH), Maximal absolute difference of* TSH dan label *class* berjumlah 3 bertipe *Polinomial*

dinotasikan (Y_n) antara lain *normal, hyper, hypo*. Tabel 4.7 merupakan karakteristik dari variabel data *thyroid*:

Tabel 4.7 Variabel *Dataset Thyroid*

Var	Nama	Tipe	Deskripsi
Y1	Normal	Polinomial	Keadaan Normal (memiliki data sebanyak 150)
Y2	Hyper	Polinomial	Kandungan lebih dari normal (memiliki data sebanyak 35)
Y3	Hypo	Polinomial	Kandungan kurang dari normal (memiliki data sebanyak 30)
X1	T3-resin uptake test	Interger	Nilai T3-resin uptake test (A percentage) (min=65 max=144, mean=110)
X2	Total serum thyroxin	Numeric	measured by the isotopic displacement method. (min=0.5 max=25.3, mean=9.8)
X3	Total serum thiiodothyronine	Numeric	measured by radioimmuno assay. (min=0.2 max=10, mean=2.1)
X4	Basal thyroid-stimulating hormone (TSH)	Numeric	measured by radioimmuno assay. (min=0.1 max=56.4, mean=2.88)
X5	Maximal absolute difference of TSH	Numeric	value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value. (min=-0.7 max=56.3, mean=4.2)

4.2.2.5 *Dataset Wine*

Dataset wine diamati oleh Stefan Aeberhard dari *Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salemo, Genoa, Italy*. Data ini adalah hasil analisis kimia dari anggur yang ditanam pada wilayah yang sama di Italia tetapi berasal dari tiga kultivar yang berbeda. Analisis menentukan jumlah 13 konstituen ditemukan di masing-masing dari tiga jenis anggur. Tujuan dari analisis ini untuk mengklasifikasikan kelas anggur berdasarkan konstituen. Data ini bersifat kontinue berjumlah 178 terdiri dari *class mayor*

berjumlah satu dan *class minor* berjumlah dua. *Atribut* sebanyak 13 dinotasikan (X_n) antara lain *alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280 of diluter wines, proline* dan label *class* berjumlah 3 bertipe *Polinomial* dinotasikan (Y_n) antara lain *class* 1, *class* 2, *class* 3. Tabel 4.8 merupakan karakteristik dari variabel data *thyroid*:

Tabel 4.8 Variabel *Dataset Wine*

Var	Nama	Tipe	Deskripsi
Y1	Class 1	Polinomial	Anggur Kelas satu (memiliki data sebanyak 59)
Y2	Class 2	Polinomial	Anggur Kelas dua (memiliki data sebanyak 71)
Y3	Class 3	Polinomial	Anggur Kelas tiga (memiliki data sebanyak 48)
X1	alcohol	Numeric	Kandungan alcohol (min=11.03 max=14.83, mean=13.00)
X2	malic acid	Real	Kandungan malic acid (min=0.74 max=5.8, mean=2.33)
X3	ash	Numeric	Kandungan ash (min=1.36 max=3.23, mean=2.36)
X4	alkalinity of ash	Numeric	Kandungan alkalinity of ash (min=10.6 max=30, mean=19.49)
X5	magnesium	Interger	Kandungan magnesium (min=70 max=162, mean=99.74)
X6	total phenols	Numeric	Kandungan total phenols (min=0.98 max=3.88, mean=2.29)
X7	flavonoids	Numeric	Kandungan flavonoids (min=0.34 max=5.08, mean=2.03)
X8	nonflavanoid phenols	Real	Kandungan nonflavanoid phenols (min=0.13 max=0.66, mean=0.36)
X9	proanthocyanins	Real	Kandungan proanthocyanins (min=0.41 max=3.58, mean=1.59)
X10	color intensity	Numeric	Kandungan *color intensity* (min=1.28 max=13, mean=148)
X11	hue	Numeric	Kandungan *hue* (min=0.48 max=1.71, mean=0.96)
X12	OD280 of	Numeric	Kandungan OD280 of diluter wines

	diluter wines		(min=1.27 max=4, mean=2.61)
X13	proline	Interger	Kandungan proline (min=278 max=1680, mean=746.89)

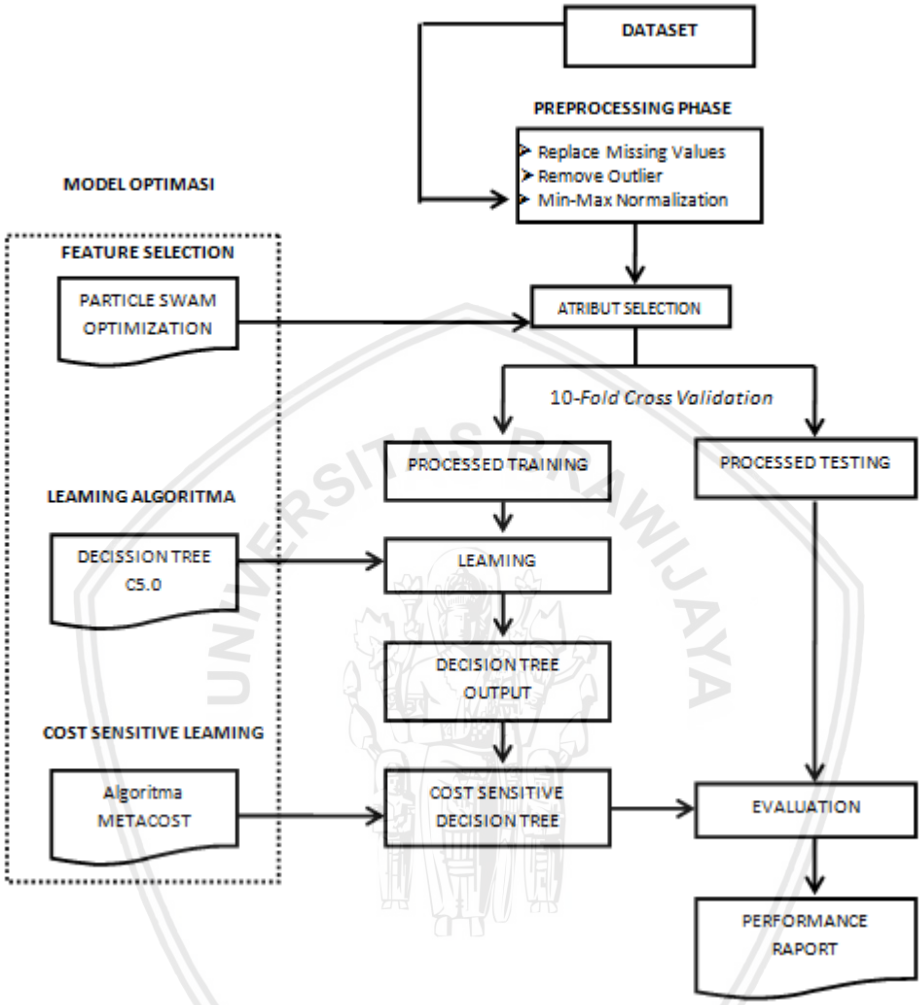
4.3 Pengolahan Data Awal

Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan adalah sebagai berikut:

a. *Selection Data*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier /noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
b. *Normalisasi Min-Max*, merupakan metode normalisasi dengan cara pengskalaan antara skala 0 sampai 1. Dengan tujuan agar setiap data memiliki nilai diantara 0 dan 1 sehingga lebih imbang dalam menentukan nilai informasi yang terkandung dalam sebuah data.
c. *Selection Atribut*, untuk memperoleh dataset dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif. Seleksi atribut menggunakan metode *particle swarm optimazation* (PSO).

4.4 Metode yang Diusulkan

Pada penelitian ini rancangan metode yang digunakan dalam menyelesaikan permasalahan data *imbalanced multiclass* ditampilkan pada gambar di bawah ini:



Gambar 4.2 Rancangan Metode Klasifikasi

Tahap awal yaitu *dataset* melalui *preprocessing* untuk menghasilkan data baru yang tidak memiliki *outlier, missing value.* selanjutnya dilakukan pemilihan atribut yang *informative* dan *relevan*. Tahap kedua *dataset* hasil dari *preprocessing* dan seleksi atribut kemudian dibagi menjadi dua yaitu data pelatihan (*training*) dan data pengujian (*testing*) dengan menggunakan metode 10 - *fold cross validation*. Kemudian data *training* dilakukan pembelajaran dengan menggunakan algoritma C5.0 untuk membuat pola model *decision tree*. Selanjutnya pola tersebut diuji menggunakan data *testing* dimana hasil dari pengujian tersebut akan

mengevaluasi nilai *cost* pada setiap label *class* dengan menggunakan algoritma *metacost*. Label yang memiliki *cost* yang besar kemudian akan di lakukan *evaluasi* hingga terbentuk pola model pohon keputusan baru dengan nilai *cost* minimum. Pola model *decision tree* kemudian dilakukan evaluasi perfomansi dengan mengukur nilai *accuracy, recall, precision, F_measure* dan total *cost*.

4.5 Konsep Algoritma PSO

Particle swarm optimation (PSO) merupakan algoritma yang digunakan pada penelitian ini untuk proses menyeleksi atribut. Atribut pada proses ini diasumsikan sebagai partikel yang memiliki dua karakteristik yaitu posisi dan kecepatan. Berikut ini merupakan formulasi matematis untuk menggambarkan kecepatan dan posisi partikel:

$$V_j^i = V_j^i + c_1.r_1 \times \left(P_{best,j} - X_j^{i-1}\right) + c_2.r_2 \times \left(G_{best} - X_j^{i-1}\right) \quad (4.1)$$

$$X_j^i = X_j^{i-1} + V_j^i \quad (4.2)$$

Setiap partikel melakukan penyesuaian terhadap posisi partikel terbaik dari partikel tersebut (*local best*) dan penyesuaian terhadap posisi partikel terbaik dari seluruh partikel (*global best*). Sedangkan c_1 dan c_2 adalah suatu konstanta yang bemilai positif yang biasanya disebut factor pembelajaran (*leaming factor*) atau factor percepatan (*acceleration factor*). Kemudian r_1 dan r_2 adalah suatu bilangan acak (*random*) yang bemilai antara 0 sampai 1. Proses dilakukan dengan jumlah iterasi tertentu sampai mendapatkan nilai kriteria maksimal pada posisi terbaik setiap partikel (*global best*).

Berikut *Pseudo code* dari algoritma PSO:

For setiap partikel

    Inisialisasi partikel

End

Repeat

    For setiap partikel

        Hitung nilai kecepatan dan posisi

If nilai kecepatan dan posisi baru lebih baik dari pada nilai kecepatan dan posisi lama

Perbarui nilai kecepatan dan posisi dari partikel tersebut

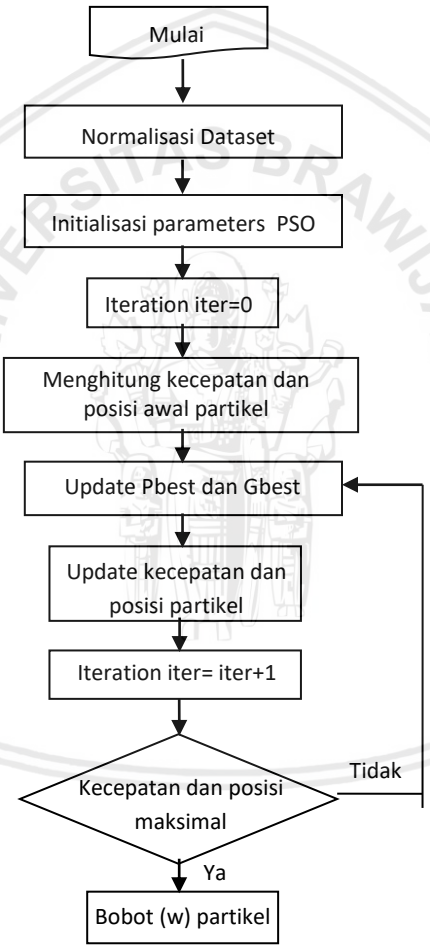
End

End

Pilih partikel dengan nilai kecepatan dan posisi terbaik diantara semua partikel

End

Berikut ini adalah diagram alir dari algoritma PSO:



Gambar 4.3 Diagram Alir *Particle Swarm Optimazation*

Tahapan dari algoritma PSO untuk memilih atribut yang relevan berdasarkan nilai bobot (w) adalah sebagai berikut:

Step 1. Inisialisasi nilai partikel secara acak (*random).*

Step 2. Inisialisasi posisi dari setiap partikel dan kecepatan dari setiap partikel.

Step 3. Hitung kecepatan partikel *i* pada *d* dimensi menggunakan persamaan 4.1

Step 4. Menghitung posisi partikel *i* pada *d* dimensi menggunakan persamaan 4.2

Step 5. Melakukan evaluasi *Pbest* dan *Gbest* untuk mendapatkan posisi dan kecepatan paling maksimal pada setiap partikel.

Step 6. Evaluasi partikel (atribut) berdasarkan nilai bobot yang mendekati 0 atau bemilai 0.

Step 7. Partikel (atribut) yang tidak relevan dapat dihilangkan.

Contoh seleksi atribut :

Tabel 4.9 Contoh *Dataset Thyroid*

Class	x_1	x_2	x_3	x_4	x_5
1	107	10.1	2.2	0.9	2.7
1	113	9.9	3.1	2	5.9
1	127	12.9	2.4	1.4	0.6
2	113	17.2	1.8	1	0
2	65	25.3	5.8	1.3	0.2
2	88	24.1	5.5	0.8	0.1
2	65	18.2	10	1.3	0.1
3	125	2.3	0.9	16.5	9.5
3	120	6.8	2.1	10.4	38.6
3	108	3.5	0.6	1.7	1.4

Tabel 4.9 merupakan sampel data yang diperoleh dari *dataset thyroid* dengan mengambil sepuluh sampel data dari setiap *class* masing-masing berjumlah *class* 1 = 3, *class* 2 = 4, *class* 3 = 3. *Dataset thyroid* memiliki lima variabel atribut antara lain x_1 =*T3-resin uptake test*, x_2 = *total serum thyroxin* , x_3 = *total serum triiodothyronine*, x_4 = *thyroid-stimulating hormone (TSH)*, x_5 = *Maximal absolute difference of* TSH. Diasumsikan dalam kasus ini nilai kecepatan awal adalah 0 dan bobot $c_1 = c_2 = 1.5$. Proses perhitungan dilakukan hingga iterasi ke 5 (*i*=1,2,3,4,5).

Langkah-langkah proses seleksi atribut sebagai berikut:

1. Pada tahap awal *dataset* dilakukan proses normalisasi menggunakan metode *normalize min-max* yang bertujuan untuk merubah nilai data menjadi skala antara 0 sampai 1. Diperoleh data seperti tabel

Tabel 4.10 Normalisasi *Dataset Thyroid*

Class	x_1	x_2	x_3	x_4	x_5
1	0.68	0.34	0.17	0.01	0.07
1	0.77	0.33	0.27	0.08	0.15
1	1.00	0.46	0.19	0.04	0.02
2	0.77	0.65	0.13	0.01	0.00
2	0.00	1.00	0.55	0.03	0.01
2	0.37	0.95	0.52	0.00	0.00
2	0.00	0.69	1.00	0.03	0.00
3	0.97	0.00	0.03	1.00	0.25
3	0.89	0.20	0.16	0.61	1.00
3	0.69	0.05	0.00	0.06	0.04

2. Inisialisasi data, dengan jumlah partikel N = 5. Didapat populasi awal secara acak (*random*) berdasarkan nilai rata-rata dari setiap partikel, misalkan diambil contoh pada baris ke 2 sebagai berikut:

$X_1^0 =$ 0.77
$X_2^0 =$ 0.33
$X_3^0 =$ 0.27
$X_4^0 =$ 0.08
$X_5^0 =$ 0.15

3. Seperti yang sudah dipaparkan diatas, ditentukan kecepatan awal $v_1^0 = v_2^0 = v_3^0 = v_4^0 = v_5^0 = 0$ dan ditetapkan iterasi pertama yaitu i = 1.

4. Selanjutnya ditemukan nilai terbaik lokal sejauh ini yang merupakan posisi partikel terbaik dari partikel tersebut ($P_{best,j}$) serta nilai terbaik global yang merupakan partikel terbaik dari seluruh partikel (G_{best}).

   $P_{best,1}$ = 0.77

   $P_{best,2}$ =0.33

   $P_{best,3}$ =0.27

   $P_{best,4}$ =0.08

   $P_{best,5}$ =0.15

   G_{best} =0.77

5. Hitung v_j^i dengan $c_1 = c_2 = 1.5$ dan misalkan nilai bilangan acak (*random*) yang didapat $r_1 = 0.2$ dan $r_2 = 0.3$ menggunakan persamaan (4.1). Diperoleh:

   $$V_1^1 = 0 + (1.5).(0.2) \times (0.77 - 0.77) + (1.5).(0.3) \times (0.77 - 0.77) = 0$$

   $$V_2^1 = 0 + (1.5).(0.2) \times (0.33 - 0.33) + (1.5).(0.3) \times (0.77 - 0.33) = 0.20$$

   $$V_3^1 = 0 + (1.5).(0.2) \times (0.27 - 0.27) + (1.5).(0.3) \times (0.77 - 0.27) = 0.23$$

   $$V_4^1 = 0 + (1.5).(0.2) \times (0.08 - 0.08) + (1.5).(0.3) \times (0.77 - 0.08) = 0.31$$

   $$V_5^1 = 0 + (1.5).(0.2) \times (0.15 - 0.15) + (1.5).(0.3) \times (0.77 - 0.15) = 0.28$$

   Sehingga nilai x_j^i sebagai posisi partikel yang baru adalah :

   $$X_1^1 = 0.77 + 0 = 0.77$$

   $$X_2^1 = 0.33 + 0.20 = 0.53$$

   $$X_3^1 = 0.27 + 0.23 = 0.50$$

   $$X_4^1 = 0.08 + 0.31 = 0.49$$

   $$X_5^1 = 0.15 + 0.28 = 0.43$$

6. Sehingga diperoleh Pbest, j baru untuk masing-masing partikel sebagai berikut:

   $P_{best,1}$ = 0.77

   $P_{best,2}$ = 0.53

   $P_{best,3}$ = 0.50

$P_{best,4} = 0.49$

$P_{best,5} = 0.43$

$G_{best} = 0.77$

7. Hitung kecepatan baru dan misalkan nilai bilangan acak (*random*) selanjutnya yang didapat adalah

$$V_1^2 = 0 + (1.5).(0.4) \times (0.77 - 0.77) + (1.5).(0.5) \times (0.77 - 0.77) = 0$$

$$V_2^2 = 0.20 + (1.5).(0.4) \times (0.53 - 0.53) + (1.5).(0.5) \times (0.77 - 0.53) = 0.38$$

$$V_3^2 = 0.23 + (1.5).(0.4) \times (0.50 - 0.50) + (1.5).(0.5) \times (0.77 - 0.50) = 0.43$$

$$V_4^2 = 0.31 + (1.5).(0.4) \times (0.40 - 0.40) + (1.5).(0.5) \times (0.77 - 0.50) = 0.60$$

$$V_5^2 = 0.28 + (1.5).(0.4) \times (0.43 - 0.43) + (1.5).(0.5) \times (0.77 - 0.43) = 0.53$$

Sehingga nilai sebagai posisi partikel yang baru adalah:

$$X_1^0 = 0.77 + 0.00 = 0.77$$

$$X_2^0 = 0.53 + 0.38 = 0.91$$

$$X_3^2 = 0.50 + 0.43 = 0.93$$

$$X_4^2 = 0.39 + 0.60 = 0.99$$

$$X_5^0 = 0.43 + 0.53 = 0.96$$

8. Proses tersebut berlanjut sampai iterasi ke 5 (i = 5 ; j = 1,2,3,4,5) sehingga diperoleh nilai partikel maksimal sebagai berikut:

Tabel 4.11 Nilai iterasi ke-1

Partikel	Nilai x_j
x_1	1.73
x_2	2.39
x_3	2.48
x_4	2.77
x_5	2.67

9. Kemudian dilakukan normalisasi menggunakan metode *normalize min-max* untuk mendapatkan nilai bobot partikel (w) dengan skala 0 sampai 1.

Tabel 4.12 Nilai Bobot *Dataset Thyroid*

Partikel	Bobot (w) x_j
x_1	0.00
x_2	0.64
x_3	0.72
x_4	1.00
x_5	0.90

10. Langkah selanjutnya dilakukan seleksi partikel (atribut) dengan cara menghapus partikel yang memiliki nilai bobot mulai paling kecil kemudian dilakukan pengujian nilai *accuracy*. Dimisalkan hasil dari pengujian sebagai berikut:

Tabel 4.13 Evaluasi Atribut *Dataset Thyroid*

Partikel	Bobot (w) x_j	Accuracy
x_1	0.00	97.21%
x_2	0.64	90.23%
x_3	0.72	60.56%
x_4	1.00	-
x_5	0.90	-

11. Berdasarkan Tabel 4.13 dengan menghilangkan partikel x_1 mampu memperoleh *accuracy* sebesar 97.21% namun ketika partikel dengan bobot > 0.5 dihilangkan nilai *accuracy* menjadi menurun. Sehingga pada contoh ini partikel (atribut) yang dihilangkan adalah x_1 =*T3-resin uptake test* dan atribut yang digunakan adalah x_2 = *total serum thyroxin* , x_3 = *total serum triiodothyronine*, x_4 = *thyroid-stimulating hormone (TSH)*, x_5 = *Maximal absolute difference of* TSH.

4.6 Konsep Metode *Decision Tree*

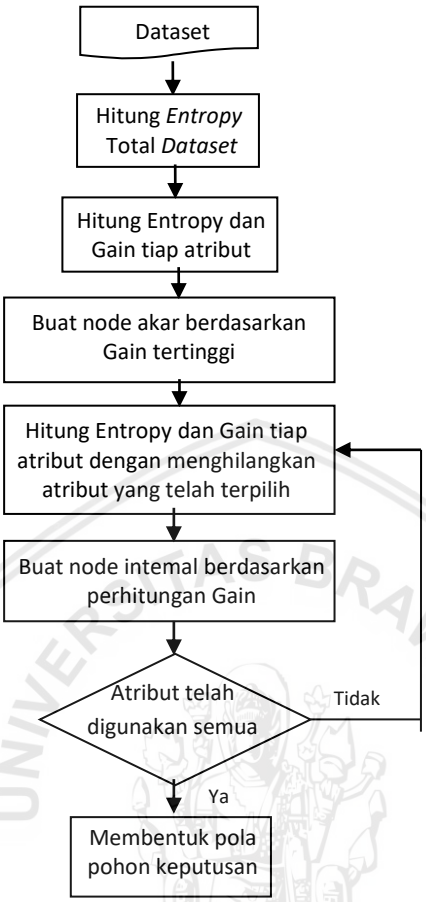
Decision tree merupakan metode klasifikasi yang digunakan untuk mengambil keputusan berdasarkan pola pembelajaran. Algoritma C5.0 digunakan untuk membuat pola model pohon keputusan berdasarkan nilai *entropy* dan *information gain*. Berikut ini merupakan formulasi matematis untuk mendapatkan nilai *entropy* dan *information gain*:

$$H(S) = \sum_{i=1}^{n} - p(s_i) \log_2 (p(s_i)) \quad (4.2)$$

$$IG(S, A_i) = H(S) - \sum_{\alpha \in A_i}^{n} \frac{|S_\alpha|}{|S|} H(S_\alpha) \quad (4.3)$$

Entropy (S) adalah banyaknya jumlah *bit* yang dibutuhkan untuk mengekstrak suatu *class* dari sekumpulan data acak pada ruang sample S. Semakin kecil nilai *entropy* maka memiliki hasil yang baik untuk digunakan dalam mengekstraksi suatu *class*. Nilai informasi yang dinyatakan sebagai panjang kode adalah p2log− *bits* yang memiliki probabilitas p. *Entropy* merupakan suatu parameter yang digunakan untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan data sampel. Semakin besar nilai entropy maka jumlah kriteria pada data sampel semakin heterogen. Setelah mendapatkan nilai *entropy* selanjutnya mencari nilai *information gain*. *Information gain* digunakan untuk mengukur efektivitas karakteristik atribut dalam mengklasifikasikan *class*.

Berikut ini adalah diagram alir dari metode *decision tree*:



Gambar 4.4 Diagram Alir *Decision Tree C5.0*

Konsep dari metode *decision tree* C5.0 sebagai berikut:

Step 1. Menghitung nilai *entropy* total *dataset* menggunakan persamaan 4.2.

Step 2. Menghitung nilai *entropy* dan *information gain* pada setiap kriteria atribut menggunakan persamaan 4.2 dan 4.3.

Step 4. Menentukan node akar berdasarkan nilai *information gain* terbesar menggunakan persamaan 4.3.

Step 5. Menentukan *intemal node* hingga menghasilkan *leaf node* berdasarkan nilai *entropy* dan *information gain*.

Step 6. Proses berhenti jika atribut telah digunakan semua.

Contoh membuat pola *decision tree* C5.0:

Tabel 4.14 Contoh *Dataset Wine*

No.	Class	Alcohol	Total Phenols	Fllavanoids	Color Intensity	Diluted Wines	Proline
1	Y1	14,23	2,8	3,06	5,64	3,92	1065
2	Y1	13,2	2,65	2,76	4,38	3,4	1050
3	Y1	13,16	2,8	3,24	5,68	3,17	1185
4	Y1	13,24	2,8	2,69	4,32	2,93	735
5	Y1	14,39	2,5	2,52	5,25	3,58	1290
6	Y1	14,83	2,8	2,98	5,2	2,85	1045
7	Y1	13,86	2,98	3,15	7,22	3,55	1045
8	Y1	13,63	2,85	2,91	7,3	2,88	1310
9	Y1	14,3	2,8	3,14	6,2	2,65	1280
10	Y1	13,83	2,95	3,4	6,6	2,57	1130
11	Y1	14,19	3,3	3,93	8,7	2,82	1680
12	Y2	12,37	1,98	0,57	1,95	1,82	520
13	Y2	12,33	2,05	1,09	3,27	1,67	680
14	Y2	12,64	2,02	1,41	5,75	1,59	450
15	Y2	13,67	2,1	1,79	3,8	2,46	630
16	Y2	12,37	3,5	3,1	4,45	2,87	420
17	Y2	12,17	1,89	1,75	2,95	2,23	355
18	Y2	12,37	2,42	2,65	4,6	2,3	678
19	Y2	13,11	2,98	3,18	5,3	3,18	502
20	Y2	12,37	2,11	2	4,68	3,48	510
21	Y2	13,34	2,53	1,3	3,17	1,93	750
22	Y2	12,21	1,85	1,28	2,85	3,07	718
23	Y2	12,29	1,1	1,02	3,05	1,82	870
24	Y3	12,86	1,51	1,25	4,1	1,29	630
25	Y3	12,88	1,3	1,22	5,4	1,42	530
26	Y3	12,81	1,15	1,09	5,7	1,36	560
27	Y3	12,7	1,7	1,2	5	1,29	600
28	Y3	12,51	2	0,58	5,45	1,51	650
29	Y3	12,6	1,62	0,66	7,1	1,58	695
30	Y3	12,25	1,38	0,47	3,85	1,27	720

Tabel 4.14 merupakan contoh *dataset* klasifikasi dengan mengambil 30 sampel data yang memiliki dua variabel yaitu label *class* dan *atribut*. *Dataset wine* mengandung informasi tentang klasifikasi jenis minuman anggur yang memiliki 6 karakteristik atribut antara lain *alcohol*, *total phenols*, *flavonoids*, *color intensity*, *diluter wines* dan *proline* dan meiliki tiga

jenis *class* yaitu *class* 1, *class* 2 dan *class* 3. Berdasarkan tabel 4.14 untuk dapat menemukan pola klasifikasi maka terlebih dahulu mencari kategori pada masing-masing atribut sebagai nilai karakteristik dengan cara melakukan analisa pada *dataset*. Setelah dilakukan analisa maka data tersebut dikonversikan kedalam bentuk tabel klasifikasi seperti berikut:

Tabel 4.15 Perhitungan klasifikasi ke-1 node 1

Kriteria (*atribut*)	**Kategori**	**Total**	**Y1**	**Y2**	**Y3**	**Entropy**	**Information Gain**
		30	11	12	7	1,549	
alkohol	>12.6	19	11	4	4	1,403	-2,797
	≤12.6	11	0	8	2	0,781	
	>13.3	10	8	2	0	0,722	
	≤13.3	20	3	10	7	1,441	
total phenol	>2	19	11	8	0	0,982	-0,378
	≤2	11	0	4	7	0,946	
flavanolds	>0.8	26	11	11	4	1,466	-2,671
	≤0.8	4	0	1	3	0,811	
	>1.4	18	11	7	0	0,964	
	≤1.4	12	0	5	7	0,980	
color in	>3.5	24	11	6	7	1,534	0,015
	≤3.5	6	0	6	0	0,000	
diluted	>2	18	11	7	0	0,964	-0,395
	≤2	12	0	5	7	0,980	
proline	>755	11	10	1	0	0,439	-0,101
	≤755	19	1	11	7	1,211	

Langkah pertama mencari *entropy* total dengan menggunakan persamaan (2.4).

$$H(Total) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(Total) = (-\frac{11}{30}\times\log_2(\frac{11}{30})) + (-\frac{12}{30}\times\log_2(\frac{12}{30}) + (-\frac{7}{30}\times\log_2(\frac{7}{30}))$$

$$H(Total) = 1.549$$

Langkah kedua mencari nilai *entropy* pada setiap kriteria *atribut* dengan menggunakan persamaan (2.4) dan mencari nilai gain informasi menggunakan persamaan (2.5).

A. ***Atribut Alcohol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>12.6) = (-\frac{11}{19}\times\log_2(\frac{11}{19})) + (-\frac{4}{19}\times\log_2(\frac{4}{19}) + (-\frac{4}{19}\times\log_2(\frac{4}{19})) = 1.403$$

$$H(\leq 12.6) = (-\frac{11}{11}\times\log_2(\frac{11}{11})) + (-\frac{4}{11}\times\log_2(\frac{4}{11}) + (-\frac{4}{11}\times\log_2(\frac{4}{11})) = 0.781$$

$$H(>13.3) = (-\frac{8}{10}\times\log_2(\frac{8}{10})) + (-\frac{2}{10}\times\log_2(\frac{2}{10}) + (-\frac{0}{10}\times\log_2(\frac{0}{10})) = 0.722$$

$$H(\leq 13.3) = (-\frac{3}{20}\times\log_2(\frac{3}{20})) + (-\frac{10}{20}\times\log_2(\frac{10}{20}) + (-\frac{7}{20}\times\log_2(\frac{7}{20})) = 1.441$$

$$Gain(Alkohol) = H(Total) - \left(\frac{19}{30}\times H(>12.6) + \frac{11}{30}\times H(\leq 12.6) + \frac{10}{30}\times H(>13.3) + \frac{20}{30}\times H(\leq 13.3)\right)$$

$$Gain(Alkohol) = 1.549 - (1.403 + 0.781 + 0.722 + 1.441) = -2.797$$

B. ***Atribut Total Phenol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>2) = (-\frac{11}{19}\times\log_2(\frac{11}{19})) + (-\frac{8}{19}\times\log_2(\frac{8}{19}) + (-\frac{0}{19}\times\log_2(\frac{0}{19})) = 0.982$$

$$H(\leq 2) = (-\frac{0}{11}\times\log_2(\frac{0}{11})) + (-\frac{4}{11}\times\log_2(\frac{4}{11}) + (-\frac{7}{11}\times\log_2(\frac{7}{11})) = 0.946$$

$$Gain(Tot.Phenol) = H(Total) - \left(\frac{19}{30}\times H(>2) + \frac{11}{30}\times H(\leq 2)\right)$$

$$Gain(Tot.Phenol) = 1.549 - (0.982 + 0.946) = -0.378$$

C. ***Atribut Flavanoids***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>0.8) = (-\frac{11}{26}\times\log_2(\frac{11}{26})) + (-\frac{11}{26}\times\log_2(\frac{11}{26}) + (-\frac{4}{26}\times\log_2(\frac{4}{26})) = 1.466$$

$$H(\leq 0.8) = (-\frac{0}{4}\times\log_2(\frac{0}{4})) + (-\frac{4}{4}\times\log_2(\frac{4}{4}) + (-\frac{7}{4}\times\log_2(\frac{7}{4})) = 0.811$$

$$H(>1.4) = (-\frac{8}{18}\times\log_2(\frac{8}{18})) + (-\frac{2}{18}\times\log_2(\frac{2}{18}) + (-\frac{0}{18}\times\log_2(\frac{0}{18})) = 0.964$$

$$H(\leq 1.4) = (-\frac{0}{12} \times \log_2(\frac{0}{12})) + (-\frac{5}{12} \times \log_2(\frac{5}{12}) + (-\frac{7}{12} \times \log_2(\frac{7}{12})) = 0.980$$

$$Gain(Flavanolds) = H(Total) - \left(\frac{26}{30} \times H(> 0.8) + \frac{4}{30} \times H(\leq 0.8) + \frac{18}{30} \times H(> 1.4) + \frac{12}{30} \times H(\leq 1.4) \right)$$

$$Gain(Flavanolds) = 1.549 - (1.466 + 0.811 + 0.964 + 0.980) = -2.671$$

D. ***Atribut Color intensity***

$$H(S) = \sum_{i=1}^{n} -p(s_i) \log_2(p(s_i))$$

$$H(> 3.5) = (-\frac{11}{24} \times \log_2(\frac{11}{24})) + (-\frac{6}{24} \times \log_2(\frac{6}{24}) + (-\frac{7}{24} \times \log_2(\frac{7}{24})) = 1.534$$

$$H(\leq 3.5) = (-\frac{0}{6} \times \log_2(\frac{0}{6})) + (-\frac{6}{6} \times \log_2(\frac{6}{6}) + (-\frac{0}{6} \times \log_2(\frac{0}{6})) = 0$$

$$Gain(Color\_in) = H(Total) - \left(\frac{24}{30} \times H(> 3.5) + \frac{24}{30} \times H(\leq 3.5) \right)$$

$$Gain(Color\_in) = 1.549 - (1.534 + 0) = 0.015$$

E. ***Atribut Diluter Wines***

$$H(S) = \sum_{i=1}^{n} -p(s_i) \log_2(p(s_i))$$

$$H(> 2) = (-\frac{11}{18} \times \log_2(\frac{11}{18})) + (-\frac{7}{18} \times \log_2(\frac{7}{18}) + (-\frac{0}{18} \times \log_2(\frac{0}{18})) = 0.964$$

$$H(\leq 2) = (-\frac{0}{12} \times \log_2(\frac{0}{12})) + (-\frac{5}{12} \times \log_2(\frac{5}{12}) + (-\frac{7}{12} \times \log_2(\frac{7}{12})) = 0.980$$

$$Gain(Wines) = H(Total) - \left(\frac{18}{30} \times H(> 2) + \frac{18}{30} \times H(\leq 2) \right)$$

$$Gain(Wines) = 1.549 - (0.964 + 0.980) = -0.395$$

F. ***Atribut Proline***

$$H(S) = \sum_{i=1}^{n} -p(s_i) \log_2(p(s_i))$$

$$H(> 755) = (-\frac{10}{11} \times \log_2(\frac{10}{11})) + (-\frac{1}{11} \times \log_2(\frac{1}{11}) + (-\frac{0}{11} \times \log_2(\frac{0}{11})) = 0.439$$

$$H(\leq 755) = (-\frac{1}{19} \times \log_2(\frac{1}{19})) + (-\frac{11}{19} \times \log_2(\frac{11}{19}) + (-\frac{7}{19} \times \log_2(\frac{7}{19})) = 1.211$$

$$Gain(\text{Pr}\,oline) = H(Total) - \left(\frac{11}{30} \times H(> 755) + \frac{19}{30} \times H(\leq 755)\right)$$

$$Gain(\text{Pr}\,oline) = 1.549 - (0.439 + 1.211) = -0.101$$

Langkah ketiga setelah memperoleh nilai *entropy* dan *gain* maka langkah selanjutnya membuat struktur pohon untuk menghasilkan sebuah keputusan dan rule. Berdasarkan tabel 4.16 nilai *gain* terbesar adalah *color intensity* yang akan digunakan sebagai *root node*. *Color intensity* memiliki 2 kategori yaitu *>3.5 dan ≤3.5* yang selanjutnya digunakan sebagai cabang. Cabang *≤3.5* telah mengklasifikasikan kasus menjadi keputusan yaitu *class 2 (node 1.2)* sedangkan *>3.5 (node 1.1)* perlu dilakukan perhitungan lagi karena masih terdapat tiga keputusan yaitu class 1, class 2 dan class 3. .

Tabel 4.16 Perhitungan klasifikasi ke-2 node 2.1.

Kriteria (*atribut*)	**Kategori**	**Total**	**Y1**	**Y2**	**Y3**	**Entropy**	**Information Gain**
Color In	>3.5	24	11	6	7	1,534	
Alkohol	>12.6	18	11	3	4	1,347	-2,822
	≤12.6	6	0	3	3	1,000	
	>13.3	9	8	1	0	0,503	
	≤13.3	15	3	5	7	1,506	
Total Phenol	>2	17	11	6	0	0,937	0,598
	≤2	7	0	0	7	0,000	
Flavanolds	>0.8	21	11	6	4	1,461	-0,863
	≤0.8	3	0	0	3	0,000	
	>1.4	17	11	6	0	0,937	
	≤1.4	7	0	0	7	0,000	
Diluter	>2	16	11	5	0	0,896	0,095
	≤2	8	0	1	7	0,544	
Proline	>755	10	10	0	0	0,000	0,239
	≤755	14	1	6	7	1,296	

$$H(Total, color) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(Total, color) = (-\frac{11}{24} \times \log_2(\frac{11}{24})) + (-\frac{6}{24} \times \log_2(\frac{6}{24}) + (-\frac{7}{24} \times \log_2(\frac{7}{24}))$$

$$H(Total, color) = 1.534$$

A. ***Atribut Alkohol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>12.6) = (-\frac{11}{18}\times\log_2(\frac{11}{18})) + (-\frac{3}{18}\times\log_2(\frac{3}{18}) + (-\frac{4}{18}\times\log_2(\frac{4}{18})) = 1.347$$

$$H(\leq 12.6) = (-\frac{0}{6}\times\log_2(\frac{0}{6})) + (-\frac{3}{6}\times\log_2(\frac{3}{6}) + (-\frac{3}{6}\times\log_2(\frac{3}{6})) = 1$$

$$H(>13.3) = (-\frac{8}{9}\times\log_2(\frac{8}{9})) + (-\frac{1}{9}\times\log_2(\frac{1}{9}) + (-\frac{0}{9}\times\log_2(\frac{0}{9})) = 0.503$$

$$H(\leq 13.3) = (-\frac{3}{15}\times\log_2(\frac{3}{15})) + (-\frac{5}{15}\times\log_2(\frac{5}{15}) + (-\frac{7}{15}\times\log_2(\frac{7}{15})) = 1.506$$

$$Gain(Alkohol) = H(Total) - \left(\frac{18}{24}\times H(>12.6) + \frac{6}{24}\times H(\leq 12.6) + \frac{9}{24}\times H(>13.3) + \frac{15}{24}\times H(\leq 13.3)\right)$$

$$Gain(Alkohol) = 1.534 - (1.347 + 1 + 0.503 + 1.506) = -2.822$$

B. ***Atribut Total Phenol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>2) = (-\frac{11}{17}\times\log_2(\frac{11}{17})) + (-\frac{6}{17}\times\log_2(\frac{6}{17}) + (-\frac{0}{17}\times\log_2(\frac{0}{17})) = 0.937$$

$$H(\leq 2) = (-\frac{0}{7}\times\log_2(\frac{0}{7})) + (-\frac{0}{7}\times\log_2(\frac{0}{7}) + (-\frac{7}{7}\times\log_2(\frac{7}{7})) = 0$$

$$Gain(Tot.Phenol) = H(Total) - \left(\frac{17}{24}\times H(>2) + \frac{7}{24}\times H(\leq 2)\right)$$

$$Gain(Tot.Phenol) = 1.534 - (0.937 + 0) = 0.598$$

C. ***Atribut Flavanoids***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>0.8) = (-\frac{11}{21}\times\log_2(\frac{11}{21})) + (-\frac{6}{21}\times\log_2(\frac{6}{21}) + (-\frac{4}{21}\times\log_2(\frac{4}{21})) = 1.463$$

$$H(\leq 0.8) = (-\frac{0}{3}\times\log_2(\frac{0}{3})) + (-\frac{0}{3}\times\log_2(\frac{0}{3}) + (-\frac{3}{3}\times\log_2(\frac{3}{3})) = 0$$

$$H(>1.4) = (-\frac{11}{17}\times\log_2(\frac{11}{17})) + (-\frac{6}{17}\times\log_2(\frac{6}{17}) + (-\frac{0}{17}\times\log_2(\frac{0}{17})) = 0.937$$

$$H(\leq 1.4) = (-\frac{7}{7} \times \log_2(\frac{7}{7})) + (-\frac{0}{7} \times \log_2(\frac{0}{7}) + (-\frac{0}{7} \times \log_2(\frac{0}{7})) = 0$$

$$Gain(Flavanolds) = H(Total) - \left(\frac{21}{24} \times H(> 0.8) + \frac{3}{24} \times H(\leq 0.8) + \frac{17}{24} \times H(> 1.4) + \frac{7}{24} \times H(\leq 1.4) \right)$$

$$Gain(Flavanolds) = 1.534 - (1.461 + 0 + 0.937 + 0) = -0.863$$

D. ***Atribut Diluter Wines***

$$H(S) = \sum_{i=1}^{n} -p(s_i) \log_2(p(s_i))$$

$$H(> 2) = (-\frac{11}{16} \times \log_2(\frac{11}{16})) + (-\frac{5}{16} \times \log_2(\frac{5}{16}) + (-\frac{0}{16} \times \log_2(\frac{0}{16})) = 0.896$$

$$H(\leq 2) = (-\frac{0}{8} \times \log_2(\frac{0}{8})) + (-\frac{1}{8} \times \log_2(\frac{1}{8}) + (-\frac{7}{8} \times \log_2(\frac{7}{8})) = 0.544$$

$$Gain(Diluter) = H(Total) - \left(\frac{16}{24} \times H(> 2) + \frac{8}{24} \times H(\leq 2) \right)$$

$$Gain(Diluter) = 1.534 - (0.896 + 0.544) = 0.095$$

E. ***Atribut Proline***

$$H(S) = \sum_{i=1}^{n} -p(s_i) \log_2(p(s_i))$$

$$H(> 755) = (-\frac{10}{10} \times \log_2(\frac{10}{10})) + (-\frac{0}{10} \times \log_2(\frac{0}{10}) + (-\frac{0}{10} \times \log_2(\frac{0}{10})) = 0$$

$$H(\leq 755) = (-\frac{1}{14} \times \log_2(\frac{1}{14})) + (-\frac{6}{14} \times \log_2(\frac{6}{14}) + (-\frac{7}{14} \times \log_2(\frac{7}{14})) = 1.296$$

$$Gain(\text{Pr}oline) = H(Total) - \left(\frac{10}{24} \times H(> 755) + \frac{14}{24} \times H(\leq 755) \right)$$

$$Gain(\text{Pr}oline) = 1.534 - (0 + 1.296) = 0.239$$

Tabel 4.17 Perhitungan klasifikasi ke-2 node 3.1.

Kriteria (*atribut*)	Kategori	Total	Y1	Y2	Y3	Entropy	Information Gain
total phenol	>2	17	11	6	0	0,937	
alkohol	>12.6	14	11	3	0	0,750	-1,271
	≤12.6	3	0	3	0	0,000	
	>13.3	9	8	1	0	0,503	
	≤13.3	8	3	5	0	0,954	
flavanolds	>0.8	17	11	6	0	0,937	-0,937
	≤0.8	0				0,000	
	>1.4	17	11	6	0	0,937	
	≤1.4	0				0,000	
diluted	>2	16	11	5	0	0,896	0,041
	≤2	1	0	1	0	0,000	
proline	>755	10	10	0	0	0,000	0,345
	≤755	7	1	6	0	0,592	

$$H(Total, phenol) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(Total, phenol) = (-\frac{11}{24}\times\log_2(\frac{11}{24})) + (-\frac{6}{24}\times\log_2(\frac{6}{24}) + (-\frac{7}{24}\times\log_2(\frac{7}{24}))$$

$$H(Total, phenol) = 1.534$$

A. ***Atribut Alkohol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>12.6) = (-\frac{11}{14}\times\log_2(\frac{11}{14})) + (-\frac{3}{14}\times\log_2(\frac{3}{14}) + (-\frac{0}{14}\times\log_2(\frac{0}{14})) = 0.750$$

$$H(\leq 12.6) = (-\frac{0}{3}\times\log_2(\frac{0}{3})) + (-\frac{3}{3}\times\log_2(\frac{3}{3}) + (-\frac{0}{3}\times\log_2(\frac{0}{3})) = 0$$

$$H(>13.3) = (-\frac{8}{9}\times\log_2(\frac{8}{9})) + (-\frac{1}{9}\times\log_2(\frac{1}{9}) + (-\frac{0}{9}\times\log_2(\frac{0}{9})) = 0.503$$

$$H(\leq 13.3) = (-\frac{3}{8}\times\log_2(\frac{3}{8})) + (-\frac{5}{8}\times\log_2(\frac{5}{8}) + (-\frac{0}{8}\times\log_2(\frac{0}{8})) = 0.954$$

$$Gain(Alkohol) = H(Total) - \left(\frac{14}{17}\times H(>12.6) + \frac{3}{17}\times H(\leq 12.6) + \frac{9}{17}\times H(>13.3) + \frac{8}{17}\times H(\leq 13.3)\right)$$

$$Gain(Alkohol) = 0.937 - (0.750 + 0 + 0.503 + 0.954) = -1.271$$

B. ***Atribut Flavanoids***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>0.8) = (-\frac{11}{17}\times\log_2(\frac{11}{17})) + (-\frac{6}{17}\times\log_2(\frac{6}{17}) + (-\frac{0}{17}\times\log_2(\frac{0}{17})) = 0.937$$

$$H(>1.4) = (-\frac{11}{17}\times\log_2(\frac{11}{17})) + (-\frac{6}{17}\times\log_2(\frac{6}{17}) + (-\frac{0}{17}\times\log_2(\frac{0}{17})) = 0.937$$

$$Gain(Flavanolds) = H(Total) - \left(\frac{21}{24}\times H(>0.8) + \frac{17}{24}\times H(>1.4)\right)$$

$$Gain(Flavanolds) = 1.534 - (0.937 + 0.937) = -0.937$$

C. ***Atribut Diluter Wines***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>2) = (-\frac{11}{16}\times\log_2(\frac{11}{16})) + (-\frac{5}{16}\times\log_2(\frac{5}{16}) + (-\frac{0}{16}\times\log_2(\frac{0}{16})) = 0.896$$

$$H(\leq 2) = (-\frac{0}{1}\times\log_2(\frac{0}{1})) + (-\frac{1}{1}\times\log_2(\frac{1}{1}) + (-\frac{0}{1}\times\log_2(\frac{0}{1})) = 0$$

$$Gain(Diluter) = H(Total) - \left(\frac{16}{17}\times H(>2) + \frac{1}{17}\times H(\leq 2)\right)$$

$$Gain(Diluter) = 0.937 - (0.896 + 0) = 0.041$$

D. ***Atribut Proline***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>755) = (-\frac{10}{10}\times\log_2(\frac{10}{10})) + (-\frac{0}{10}\times\log_2(\frac{0}{10}) + (-\frac{0}{10}\times\log_2(\frac{0}{10})) = 0$$

$$H(\leq 755) = (-\frac{1}{7}\times\log_2(\frac{1}{7})) + (-\frac{6}{7}\times\log_2(\frac{6}{7}) + (-\frac{0}{7}\times\log_2(\frac{0}{7})) = 0.592$$

$$Gain(\mathrm{Pr}oline) = H(Total) - \left(\frac{10}{17}\times H(>755) + \frac{7}{17}\times H(\leq 755)\right)$$

$$Gain(\mathrm{Pr}oline) = 1.534 - (0 + 0.592) = 0.345$$

Tabel 4.18 Perhitungan klasifikasi ke-2 node 4.1.

Kriteria (*atribut*)	Kategori	Total	Y1	Y2	Y3	Entropy	Information Gain
proline	≤755	7	1	6	0	0,592	
alkohol	>12.6	4	1	3	0	0,811	
	≤12.6	3	0	3	0	0,000	-0,870
	>13.3	1	0	1	0	0,000	
	≤13.3	6	1	5	0	0,650	
flavanolds	>0.8	7	1	6	0	0,592	
	≤0.8	0				0,000	-0,592
	>1.4	7	1	6	0	0,592	
	≤1.4	0				0,000	
diluted	>2	6	1	5	0	0,650	-0,058

$$H(Total, proline) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(Total, proline) = (-\frac{1}{7}\times\log_2(\frac{1}{7})) + (-\frac{6}{7}\times\log_2(\frac{6}{7}) + (-\frac{0}{7}\times\log_2(\frac{0}{7}))$$

$$H(Total, proline) = 0.592$$

A. ***Atribut Alkohol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>12.6) = (-\frac{1}{4}\times\log_2(\frac{1}{4})) + (-\frac{3}{4}\times\log_2(\frac{3}{4}) + (-\frac{0}{4}\times\log_2(\frac{0}{4})) = 0.811$$

$$H(\leq 12.6) = (-\frac{0}{3}\times\log_2(\frac{0}{3})) + (-\frac{3}{3}\times\log_2(\frac{3}{3}) + (-\frac{0}{3}\times\log_2(\frac{0}{3})) = 0$$

$$H(>13.3) = (-\frac{0}{1}\times\log_2(\frac{0}{1})) + (-\frac{1}{1}\times\log_2(\frac{1}{1}) + (-\frac{0}{1}\times\log_2(\frac{0}{1})) = 0$$

$$H(\leq 13.3) = (-\frac{1}{6}\times\log_2(\frac{1}{6})) + (-\frac{5}{6}\times\log_2(\frac{5}{6}) + (-\frac{0}{6}\times\log_2(\frac{0}{6})) = 0.650$$

$$Gain(Alkohol) = H(Total) - \left(\frac{4}{7}\times H(>12.6) + \frac{3}{7}\times H(\leq 12.6) + \frac{1}{7}\times H(>13.3) + \frac{6}{7}\times H(\leq 13.3)\right)$$

$$Gain(Alkohol) = 0.592 - (0.811 + 0 + 0 + 0.650) = -0.870$$

B. ***Atribut Flavanoids***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>0.8) = (-\frac{1}{7}\times\log_2(\frac{1}{7})) + (-\frac{6}{7}\times\log_2(\frac{6}{7}) + (-\frac{0}{7}\times\log_2(\frac{0}{7})) = 0.592$$

$$H(>1.4) = (-\frac{1}{7}\times\log_2(\frac{1}{7})) + (-\frac{6}{7}\times\log_2(\frac{6}{7}) + (-\frac{0}{7}\times\log_2(\frac{0}{7})) = 0.592$$

$$Gain(Flavanolds) = H(Total) - \left(\frac{7}{7}\times H(>0.8) + \frac{7}{7}\times H(>1.4)\right)$$

$$Gain(Flavanolds) = 0.592 - (0.592 + 0.592) = -0.058$$

C. ***Atribut Diluter Wines***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>2) = (-\frac{1}{7}\times\log_2(\frac{1}{7})) + (-\frac{6}{7}\times\log_2(\frac{6}{7}) + (-\frac{0}{7}\times\log_2(\frac{0}{7})) = 0.650$$

$$Gain(Diluter) = H(Total) - \left(\frac{6}{7}\times H(>2)\right)$$

$$Gain(Diluter) = 0.937 - (0.650) = -0.058$$

Tabel 4.19 Perhitungan klasifikasi ke-2 node 5.2.

Kriteria (*atribut*)	**Kategori**	**Total**	**Y1**	**Y2**	**Y3**	**Entropy**	**Information Gain**
diluted	>2	6	1	5	0	0,650	
alkohol	>12.6	3	1	2	0	0,918	-0,990
	≤12.6	3	0	3	0	0,000	
	>13.3	1	0	1	0	0,000	
	≤13.3	5	1	4	0	0,722	
flavanolds	>0.8	6	1	5	0	0,650	-0,650
	≤0.8	0				0,000	
	>1.4	6	1	5	0	0,650	
	≤1.4	0				0,000	

$$H(Total, Diluter) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(Total, Diluter) = (-\frac{1}{6}\times\log_2(\frac{1}{6})) + (-\frac{5}{6}\times\log_2(\frac{5}{6}) + (-\frac{0}{6}\times\log_2(\frac{0}{6}))$$

$$H(Total, Diluter) = 0.650$$

A. ***Atribut Alkohol***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>12.6) = (-\frac{1}{3}\times\log_2(\frac{1}{3})) + (-\frac{2}{3}\times\log_2(\frac{2}{3}) + (-\frac{0}{3}\times\log_2(\frac{0}{3})) = 0.918$$

$$H(\leq 12.6) = (-\frac{0}{3}\times\log_2(\frac{0}{3})) + (-\frac{3}{3}\times\log_2(\frac{3}{3}) + (-\frac{0}{3}\times\log_2(\frac{0}{3})) = 0$$

$$H(>13.3) = (-\frac{0}{1}\times\log_2(\frac{0}{1})) + (-\frac{1}{1}\times\log_2(\frac{1}{1}) + (-\frac{0}{1}\times\log_2(\frac{0}{1})) = 0$$

$$H(\leq 13.3) = (-\frac{1}{5}\times\log_2(\frac{1}{5})) + (-\frac{4}{5}\times\log_2(\frac{4}{5}) + (-\frac{0}{5}\times\log_2(\frac{0}{5})) = 0.722$$

$$Gain(Alkohol) = H(Total) - \left(\frac{3}{6}\times H(>12.6) + \frac{3}{6}\times H(\leq 12.6) + \frac{1}{6}\times H(>13.3) + \frac{5}{6}\times H(\leq 13.3)\right)$$

$$Gain(Alkohol) = 0.650 - (0.918 + 0 + 0 + 0.722) = -0.990$$

B. ***Atribut Flavanoids***

$$H(S) = \sum_{i=1}^{n} -p(s_i)\log_2(p(s_i))$$

$$H(>0.8) = (-\frac{1}{6}\times\log_2(\frac{1}{6})) + (-\frac{5}{6}\times\log_2(\frac{5}{6}) + (-\frac{0}{6}\times\log_2(\frac{0}{6})) = 650$$

$$H(>1.4) = (-\frac{1}{6}\times\log_2(\frac{1}{6})) + (-\frac{5}{6}\times\log_2(\frac{5}{6}) + (-\frac{0}{6}\times\log_2(\frac{0}{6})) = 650$$

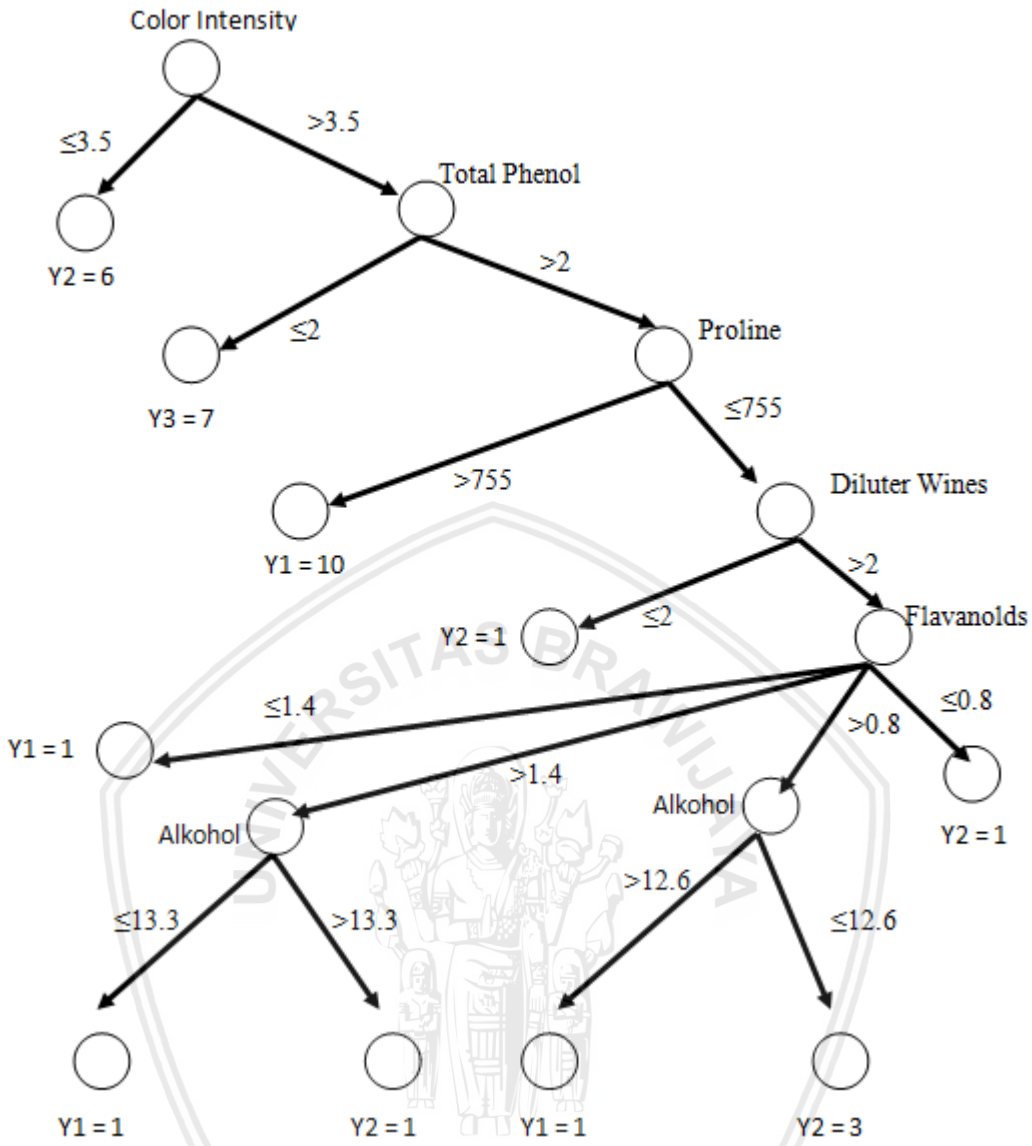
$$Gain(Flavanolds) = H(Total) - \left(\frac{6}{6}\times H(>0.8) + \frac{6}{6}\times H(>1.4)\right)$$

$$Gain(Flavanolds) = 0.650 - (0.650 + 0.650) = -0.650$$

Tabel 4.20 Hasil Perhitungan Node

Node			Total	Y1	Y2	Y3			Total	Y1	Y2	Y3
									30	11	12	7
root	node 1		24	11	6	7	Color In	>3.5	24	11	6	7
								≤3.5	6	0	6	0
leaf	node 1.1	>3.5	24	11	6	7	total phenol	>2	17	11	6	0
								≤2	7	0	0	7
leaf	node 2.1	>2	17	11	6	0	proline	>755	10	10	0	0
								≤755	7	1	6	0
leaf	node 3.1	≤755	7	1	6	0	diluted	>2	6	1	5	0
								≤2	1	0	1	0
leaf	node 4.1	>2	6	1	5	0	flavanolds	>0.8	6	1	5	0
								≤0.8	0	0	0	0
								>1.4	6	1	5	0
								≤1.4	0	0	0	0
leaf	node 5.1	>0.8	6	1	5	0	>12.6	3	1	0	0	0
							≤12.6	3	0	3	0	0
leaf	node 5.2	>1.4	6	1	5	0	>13.3	1	0	1	0	0
							≤13.3	5	1	0	0	0

Langkah ke empat setelah seluruh kriteria atribut dihitung sehingga diperoleh nilai *entropy* dan *information gain.* Tahap selanjutnya dapat dibuat pola model pohon keputusan berdasarkan nilai pada Tabel 4.20 sehingga diperoleh pola model pohon keputusan seperti gambar 4.4.



Gambar 4.5 Model *Decision Tree Dataset Wine*

Langkah kelima berdasarkan pola model *decision tree* seperti Gambar 4.4 dapat diperoleh aturan-aturan sebagai berikut:

1. Jika color intensity ≤ 3.5 maka class Y2
2. Jika color intensity > 3.5 dan total phenol ≤ 2 maka class Y3
3. Jika color intensity > 3.5 dan total phenol > 2 dan proline >755 maka class Y1
4. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines ≤ 2 maka class Y2

5. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds ≤0.8 maka class Y2
6. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds >0.8 dan alkohol ≤12.6 maka class Y2
7. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds ≤1.4 maka class Y2
8. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds >1.4 dan alkohol >13.3 maka class Y2
9. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds >1.4 dan alkohol >12.6 maka class Y1
10. Jika color intensity > 3.5 dan total phenol > 2 dan proline ≤755 dan diluter wines > 2 dan flavanolds >1.4 dan alkohol ≤13.3 maka class Y1

4.7 Konsep Algoritma *Metacost*

Metacost merupakan algoritma dari metode *cost sensitive leaming* dengan menggunakan teknik *thresholding meta leaming* untuk meminimalkan *cost.* Prisip kerja dari *metacost* adalah menghitung nilai probabilitas setiap class *j* pada label *class* (S_j) dari model *decision tree* (M_i). Persamaan untuk menghitung probabilitas sebagai berikut.

$$P(j \mid x) = \frac{1}{\sum_i 1} \sum_i P(j \mid x, M_i) \quad \text{.....(4.4)}$$

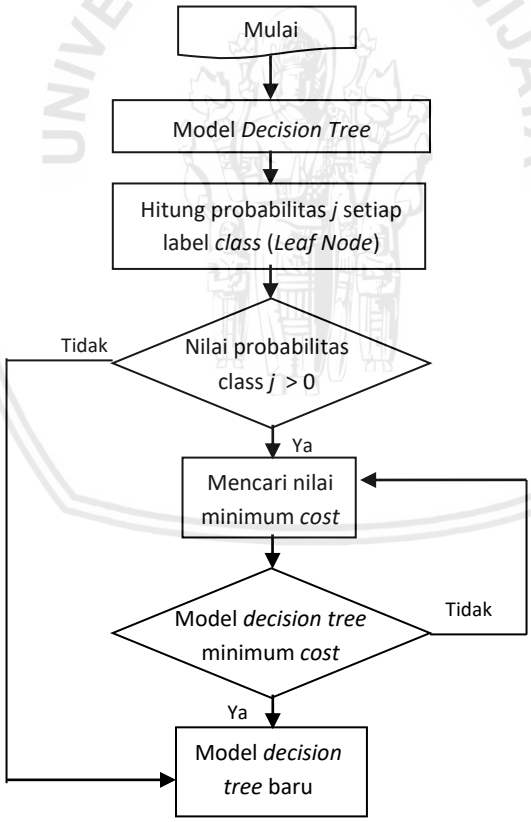
Cost dinotasikan sebagai $C(_{i,j})$, dimana *i* adalah aktual *class* tetapi diprediksi menjadi *class j* sehingga menyebabkan terjadinya *misclassification*. Ketika probabilitas pada label *class* $P(j \mid x, M_i) > 0$, maka akan dilakukan evaluasi dengan cara melakukan teknik *pruning* dan *relabeling* sampai mendapatkan minimum *cost.* Persamaan untuk mencari nilai minimum *cost* sebagai berikut .

$$S_i = \arg\min_i \sum_j P(j \mid x) C(i, j) \quad \text{.....(4.5)}$$

Berikut *Pseudo code* dari algoritma *Metacost*:

For Model pohon keputusan
Class i merupakan aktual class, nilai probabilitas = 1
Class j merupakan prediksi class, nilai probabilitas = 0
Evaluasi model pohon keputusan
End
Repeat
For setiap label *class* pada model pohon keputusan
Hitung nilai probabilitas label *class*
If nilai probabilitas label *class j* = 0
Pola model pohon keputusan tetap
End
If nilai probabilitas label class *j* > 0
Mencari *cost* minimum dengan teknik *pruning* dan *relabel*
End
Pilih model dengan nilai *cost* minimum
End

Berikut ini adalah diagram alir dari algoritma *Metacost*:



Gambar 4.6 Diagram Alir Algoritma *Metacost*

Tahapan dari algoritma *metacost* untuk meminimalkan *cost* dari kesalahan klasifikasi adalah sebagai berikut:

Step 1. Evaluasi model pohon keputusan

Step 2. Hitung nilai probabilitas setiap label *class j*

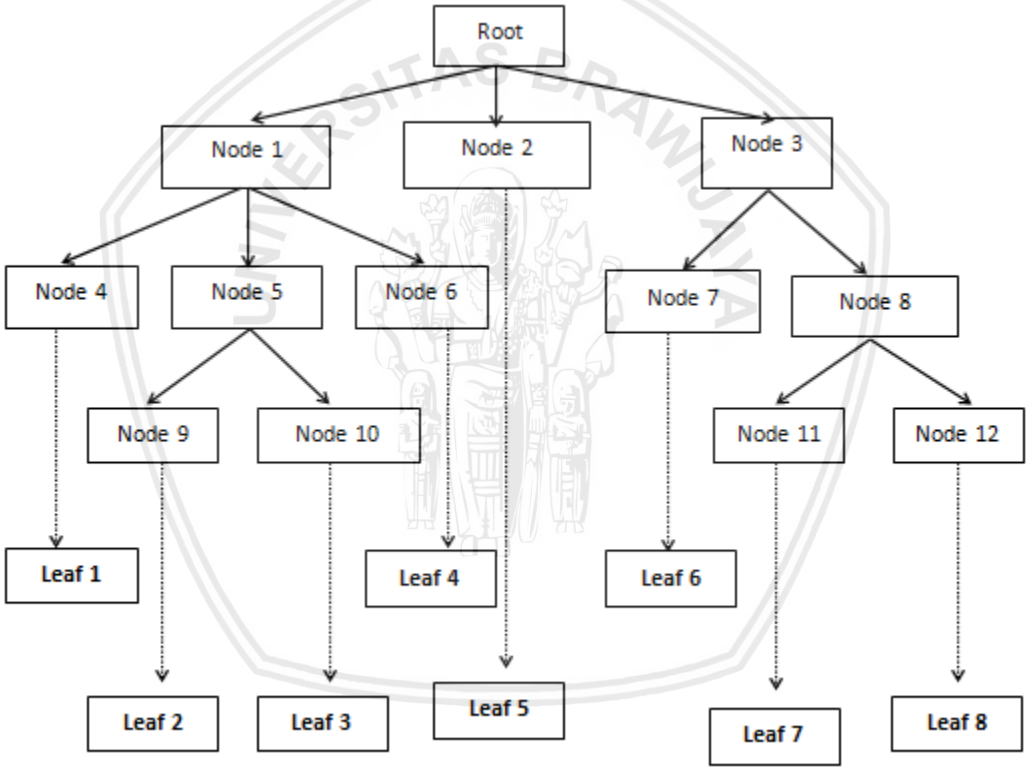
Step 3. Jika probabilitas class *j* = 0 maka model tetap

Step 4. Jika probabilitas class *j* > 0 maka dilakukan teknik *pruning* dan *relabel*

Step 5. Mencari nilai minimum *cost*

Step 6. Membuat model pohon keputusan baru berdasarkan minimum *cost*

Contoh meminimalkan *cost* :



Gambar 4.7 Contoh Model *Decision Tree*

Gambar 4.5 merupakan contoh model *decision tree* yang terdiri dari 1 *root node*, 12 *intemal node* dan 8 *leaf node*. Dimisalkan pola model pohon keputusan diujikan dengan 30 data pengujian akan diperoleh hasil pengujian sebagai berikut:

***Rule* 1:** Root > Node 1 > Node 4 > **Leaf 1** hasil (i=1, j=0)

***Rule* 2:** Root > Node 1 > Node 5 > Node 9 > **Leaf 2** hasil (i=4, j=0)

***Rule* 3:** Root > Node 1 > Node 5 > Node 10 > **Leaf 3** hasil (i=8, j=2)

***Rule* 4:** Root > Node 1 > Node 6 > **Leaf 4** hasil (i=3, j=0)

***Rule* 5:** Root > Node 2 > **Leaf 5** hasil (i=2, j=0)

***Rule* 6:** Root > Node 3 > Node 7 > **Leaf 6** hasil (i=1, j=3)

***Rule* 7:** Root > Node 3 > Node 8 > Node 11 > **Leaf 7** hasil (i=3, j=0)

***Rule* 8:** Root > Node 3 > Node 8 > Node 12 > **Leaf 8** hasil (i=3, j=0)

Langkah pertama hitung nilai probabilitas class *j* setiap *label class* atau *leaf node* menggunakan persamaan 4.4. Hasil perhitungan nilai probabilitas *j* sebagai berikut:

***Rule* 1:** (i=1, j=0), P(j|x) = 1/1 x 0 = 0

***Rule* 2:** (i=4, j=0), P(j|x) = 1/4 x 0 = 0

***Rule* 3:** (i=8, j=2), P(j|x) = 1/10 x 2 = 0.2

***Rule* 4:** (i=3, j=0), P(j|x) = 1/3 x 0 = 0

***Rule* 5:** (i=2, j=0), P(j|x) = 1/2 x 0 = 0

***Rule* 6:** (i=1, j=3), P(j|x) = 1/4 x 3 = 0.75

***Rule* 7:** (i=3, j=0), P(j|x) = 1/3 x 0 = 0

***Rule* 8:** (i=3, j=0), P(j|x) = 1/3 x 0 = 0

Langkah kedua berdasarkan perhitungan nilai probabilitas class *j* diperoleh rule yang memiliki nilai probabilitas class j > 0 yaitu pada rule 3 dan rule 6 yang selanjutnya akan dilakukan evaluasi untuk mencari kesalahan terkecil dengan cara menghitung minimum *cost* menggunakan persamaan 4.5. Hasil perhitungan sebagai berikut:

Evaluasi rule 3 dengan teknik *pruning*

***Rule* 3:** Root > Node 1 > Node 5 > Node 10 > **Leaf 3** hasil (i=8, j=2)

Dimisalkan,

Evaluasi 1 menghilangkan node 1

Rule 3: Root > Node 5 > Node 10 > Leaf 3 hasil (i=3, j=7)

Probabilitas: (i=3, j=7), S_i= 7 x 7 = 49

Evaluasi 2 menghilangkan node 5

Rule 3: Root > Node 1 > Node 10 > Leaf 3 hasil (i=9, j=1)

Probabilitas: (i=9, j=1), S_i = 1 x 1 = 1

Evaluasi 3 menghilangkan node 10

Rule 3: Root > Node 1 > Node 5 > Leaf 3 hasil (i=5, j=5)

Probabilitas: (i=5, j=5), S_i = 5 x 5 = 25

Tabel 4.21 Hasil Perhitungan *Cost* Teknik *Pruning*

Evaluasi	Cost
1	49
2	1
3	25

Berdasarkan hasil evaluasi diperoleh nilai minimum cost pada evaluasi ke 2 dengan cost sebesar 1. Sehingga evaluasi ke-2 dipilih sebagai hasil terbaik untuk model *decision tree*.

Evaluasi rule 6 dengan teknik *relabel*

***Rule* 6:** Root > Node 3 > Node 7 > **Leaf 6** hasil (i=1, j=3)

Evaluasi menganti label class kemudian dilakukan perhitungan *cost* menggunakan persamaan 4.5, dimisalkan:

1. **Leaf 6** diganti menjadi **Leaf 1** hasil (i=0, j=4), S_i = 4 x 4 =16
2. **Leaf 6** diganti menjadi **Leaf 2** hasil (i=0, j=4), S_i = 4 x 4 =16
3. **Leaf 6** diganti menjadi **Leaf 3** hasil (i=0, j=4), S_i = 4 x 4 =16
4. **Leaf 6** diganti menjadi **Leaf 4** hasil (i=0, j=4), S_i = 4 x 4 =16
5. **Leaf 6** diganti menjadi **Leaf 5** hasil (i=0, j=4), S_i = 4 x 4 =16
6. **Leaf 6** diganti menjadi **Leaf 7** hasil (i=4, j=0), S_i = 0 x 0 =0
7. **Leaf 6** diganti menjadi **Leaf 8** hasil (i=2, j=2), S_i = 2 x 2 = 4

Tabel 4.22 Hasil Perhitungan *Cost* Teknik *Relabel*

Evaluasi	Cost
1	16
2	16
3	16
4	16
5	16
6	0
7	4

Berdasarkan hasil evaluasi diperoleh nilai minimum cost pada evaluasi ke 6 dengan cost sebesar 0. Sehingga evaluasi ke-6 dipilih sebagai hasil terbaik untuk model *decision tree*. Gambar 4.5 merupakan model *decision tree* baru dengan minimum *cost* setelah proses *pruning* dan *relabel.*