BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Penelitian yang berkaitan dengan *data mining* sudah banyak dilakukan di perusahaan, instansi pemerintahan dan pendidikan. Tinjauan pustaka bertujuan sebagai referensi dan rujukan terhadap hasil penelitian sebelumnya yang berkaitan dengan penelitian yang akan dilakukan.

Menurut penelitian yang dilakukan oleh (Saefulloh & Moedjiono, 2013) yang berjudul "Penerapan Metode Klasifikasi *Data Mining* Untuk Prediksi Kelulusan Tepat Waktu". Metode yang digunakan dalam penelitian ini adalah algoritma C4.5, *Naive Bayes*, dan *Neural Network* untuk memperkirakan kelulusan tepat waktu mahasiswa dengan melihat pengaruh dari IMK dan IPK. Dari penelitian ini menunjukan bahwa algoritma terbaik adalah algoritma yang paling tinggi tingkat *accuracy* pada model klasifikasi yaitu C4.5 dan *Neural Network* dengan tingkat *accuracy* 100% sedangkan *Naive Bayes* 99.8878%. Hasil data mining dari algoritma terpilih dalam penelitian ini menggunakan C4.5, *interface* dirancang menggunakan *java engine* yang dapat menampilkan prediksi kelulusan tepat waktu beserta jumlah kelulusan tepat waktu setiap Program Studi.

Menurut penelitian yang dilakukan oleh (Nugroho, 2014) yang berjudul "Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta". Metode yang digunakan dalam penelitian ini adalah algoritma C4.5 untuk klasifikasi mahasiswa berdasarkan predikat kelulusannya. Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan bahwa telah diperoleh klasfikasi predikat kelulusan mahasiswa Fakultas Komunikasi dan Informatika UMS. Variabel yang paling tinggi pengaruhnya terhadap predikat kelulusan adalah partisipasi mahasiswa menjadi asisten. Interprestasi hasil penelitian ini adalah mengindikasikan bahwa variabel yang perlu digunakan sebagai pertimbangan bagi

Fakultas Komunikasi dan Informatika UMS untuk memperoleh tingkat predikat kelulusan yang maksimal adalah peran serta mahasiswa untuk menjadi asisten. Secara umum probabilitas predikat "*Cumlaude*" pada kelompok mahasiswa yang pernah menjadi asisten lebih tinggi dibandingkan dengan yang tidak pernah menjadi asisten jika berasal dari jurusan IPA semasa sekolah menengah atas memiliki probabilitas predikat kelulusan "*Cumlaude*" yang lebih tinggi dibandingkan dengan mahasiswa dari jurusan lainnya.

Menurut penelitian yang dilakukan oleh (Luvia, Windarto, Solikhun, & Hartama, 2017) yang berjudul "Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Keberhasilan Mahasiswa Di Amik Tunas Bangsa". Metode yang digunakan dalam penelitian ini adalah C4.5 untuk mengetahui berdasarkan kriteria apa saja mahasiswa layak mendapatkan predikat keberhasilannya dengan beberapa atribut seperti jenis kelamin, kehadiran, sesi perkuliahan, rerata NEM, dan Asal sekolah. Hasil penelitian yang diperoleh disimpulkan bahwa telah didapat klasifikasi predikat keberhasilan mahasiswa di Amik Tunas Bangsa Pematangsiantar. Variabel yang memiliki prioritas utama terhadap predikat keberhasilan mahasiswa adalah mahasiswa yang memilih sesi perkuliahan pada pagi hari dan didukung dengan nilai kehadiran > 50 maka mahasiswa tersebut mendapatkan predikat *cumlaude* dibanding dengan mahasiswa yang berada di sesi perkuliahan siang dan malam. Pengaruh ini dapat dilihat dari besarnya semangat belajar mahasiswa pagi yang memiliki banyak waktu untuk diskusi diluar jam belajar sehingga kepedulian dan kedisiplinan mahasiwa tersebut berhak mendapatkan predikat keberhasilan *cumlaude*.

Menurut penelitian yang dilakukan oleh (Agarwal, Babu, & Reddy, 2016) yang berjudul "*Classification Techniques in Data Mining-Case Study*". Metode yang digunakan dalam penelitian ini adalah ID3, *decision tree* (DT), C4.5, *Bayesian classification*, SVM, ANM, KNN untuk memberikan pemeriksaan lengkap berbagai mekanisme klasifikasi dalam *data mining*. Algoritma klasifikasi ini dapat diimplementasikan pada jenis dataset yang berbeda seperti data pasar saham, data pasien, data keuangan, dll. Oleh karena itu teknik klasifikasi ini menunjukkan

bagaimana data ditentukan dan dikelompokkan ketika sekelompok data baru tersedia. Setiap teknik memiliki fitur dan keterbatasan tersendiri.

Menurut penelitian yang dilakukan oleh (Ayu Rizqi Oktaviana, 2016) dengan judul penelitian "Penerapan *Data Mining* Klasifikasi Pola Nasabah Menggunakan Algoritma C4.5 Pada Bank BRI Batang". Metode yang digunakan dalam penelitian ini adalah *Decison Tree* untuk melakukan klasifikasi data nasabah dari Bank BRI Batang yang menjadi bahan acuan untuk menganalisa pola nasabah pemohon kredit. Pemohon kredit termasuk dalam kategori lancar atau macet. Dalam penelitian ini, *decision tree* menggunakan bahasa pemrograman *java*. Kemudian hasil akurasi dari aplikasi yang telah diimplementasikan akan dibandingkan dengan hasil yang menggunakan *software rapidminer*. Sehingga diperoleh akurasi dengan *decision tree* sebesar 89,5%.

Menurut penelitian yang dilakukan oleh (Nikam, 2015) yang berjudul "*A Comparative Study of Classification Techniques in Data Mining Algorithms*". Metode yang digunakan dalam penelitian ini adalah C4.5, ID3, *K-nearest neighbor classifer, Naive Bayes*, SVM, dan ANN untuk mengklasifikasikan data ke dalam kelas yang berbeda sesuai dengan beberapa kendala. Teknik klasifikasi ini menunjukkan bagaimana data dapat ditentukan dan dikelompokkan ketika satu set data baru tersedia. Setiap teknik memliki fitur dan keterbatasan tersendiri. Berdasarkan ketentuan, kinerja, dan fitur yang sesuai dengan kebutuhan masing-masing dapat dipilih.

Menurut penelitian yang dilakukan oleh (Arifin & Fitrianah, 2018) dengan judul penelitian "Penerapan Algoritma Klasifikasi C4.5 dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada". Metode yang digunakan dalam penelitan ini adalah C4.5 untuk melakukan proses melakukan penerimaan mitra yang sesuai dengan prosedur untuk menentukan calon mitra penjualan tersebut diterima atau ditolak. Hasil penelitian yang diperoleh disimpulkan yaitu pengklasifikasiannya divalidasi menggunakan *ten-fold cross validation* dengan tingkat akurasi 96,26%, presisi 100% , dan *recall* 71,43% dari

hasil tersebut dapat dinyatakan bahwa perhitungan yang dilakukan akan mampu memprediksi dan merekomendasikan penerimaan mitra penjual dengan baik.

Menurut penelitian yang dilakukan (Pambudi & Setiawan, 2018) yang berjudul "Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal". Metode yang digunakan adalah C4.5 untuk mengetahui tingkat akurasi prediksi kemampuan siswa sekolah menengah. Parameter pemilihan adalah faktor-faktor dalam bidang studi matematika. Berdasarkan hasil penelitian menunjukkan bahwa Algoritma *Decision Tree* C4.5 akurat diterapkan untuk prediksi nilai akhir siswa sekolah menengah dengan tingkat akurasi 60%.

Menurut penelitian yang dilakukan oleh (Novandya & Oktria, 2017) berjudul "Penerapan Algoritma Klasifikasi *Data Mining* C4.5 Pada Dataset Cuaca Wilayah Bekasi". Metode yang digunakan adalah C4.5 yang bertujuan untuk mendapatkan pola klasifikasi cuaca. Berdasarkan hasil pengujian algoritma C4.5 yang menggunakan 10-fold validation dan dibuktikan dengan pembuatan aplikasi web untuk pengujian sehingga menghasilkan nilai yang akurasinya sebesar 88,89%.

Menurut penelitian yang dilakukan oleh (Eka Sabna dan Muhardi, 2016) berjudul "Penerapan *Data Mining* Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar". Metode yang digunakan dalam penelitian ini adalah C4.5. Terdapat 5 variabel yaitu data yang terkait dengan peran dosen, motivasi, kedisplinan, sosial ekonomi, dan hasil belajar masa lalu. Berdasarkan hasil penelitian yang diperoleh bahwa variabel nilai rapor (hasil belajar masa lalu) menjadi *node* awal artinya dari 5 variabel yang menentukan prestasi akademik mahasiswa maka nilai rapor menjadi *node* yang terpilih sebagai penentu terhadap presentasi akademik mahasiswa.

Menurut penelitian yang dilakukan oleh (Adhatrao, Gaykar, Dhawan, Jha, & Honrao, 2013) berjudul "*Predicting Student's Performance Using ID3 And C4.5 Classification Algorithms*". Metode yang digunakan adalah ID3 dan C4.5 untuk memprediksi kinerja siswa. Data siswa yang digunakan adalah jenis kelamin, nilai yang dicetak dalam ujian kelas X dan XII, nilai peringkat dalam ujian masuk, dan hasil pada tahun pertamadari kelompok siswa sebelumnya. Berdasarkan penelitian tersebut mendaptakan hasil presentase akurasi rata-rata yang dicapai dalam evaluasi massal dan pribadi adalah sekitar 75,275% dari total 182 siswa.

Untuk itu yang menjadi fokus pada penelitian ini adalah bagaimana mengimplementasikan algoritma C4.5 dalam mengklasifikasi data jenis pekerjaan alumni di Universitas Muhammadiyah Yogyakarta karena belum pernah dilakukan klasifikasi terhadap jenis pekerjaan alumni pada penelitian sebelumnya. Penelitian ini menggunakan algoritma C4.5 karena algoritma tersebut memiliki kelebihan dapat membentuk pohon keputusan (*decision tree*) yang lebih efisien dan mudah dipahami sehingga dapat digunakan untuk mengidentifikasi dan melihat hubungan antara faktor-faktor yang mempengaruhi suatu kasus, selain itu algoritma C4.5 juga dapat mengolah data diskrit dan kontinu.

2.2. Landasan Teori

2.2.1. *Data Mining*

Data mining merupakan proses penggalian dan pertambangan pengetahuan dari sejumlah data yang besar, *database* atau *repository database* lainnya. Tujuan utama dari penambangan data ini untuk menemukan pengetahuan baru yang tersembunyi dari *database* tersebut (Elisa, 2017).

Data mining adalah suatu rangkaian dari proses kemudian dapat dipisah-pisah menjadi beberapa tahapan. Tahapan-tahapan yang ada dalam *data mining* bersifat interaktif terhadap pengguna yang terlibat langsung dengan perantara *knowledge base*. Tahap-tahap dalam *data mining* antara lain :

1. Pembersihan Data

   Tahap pembersihan data dilakukan untuk membuang data yang tidak konsisten dan *noise*. Selain itu, terdapat atribut data yang tidak sesuai dengan hipotesis *data mining* yang ada. Pembersihan data dapat mempengaruhi kinerja dari sistem *data mining* karena data yang diolah akan berkurang jumlah dan kompleksitasnya.

2. Integrasi Data

   Integrasi data digunakan untuk menggabungkan data dari beberapa sumber karena dapat terjadi data yang dibutuhkan dapat berasal dari beberapa *database* atau *file task*. Tahap ini dilakukan pada atribut-atribut yang unik seperti nama, jenis produk, dan nomor pelanggan. Untuk menghasilkan data yang tepat dan tidak menyimpang maka harus dilakukan dengan cermat pada tahap ini.

3. Transformasi Data

   Transformasi data dilakukan dengan mengubah data menjadi bentuk atau format yang sesuai. Sebagai contoh beberapa teknik dasar seperti analisis asosiasi dan klastering hanya dapat menerima input data kategorikal. Karena data yang berupa angka numerik perlu dipecah menjadi beberapa interval. Proses tersebut yang dinamakan *binning*. Transformasi dan pemilihan data ini menentukan ketepatan hasil dari *data mining* karena ada beberapa karakteristik dari teknik-teknik yang ada pada *data mining* tertentu bergantung dengan tahap ini.

4. Aplikasi Teknik *Data Mining*

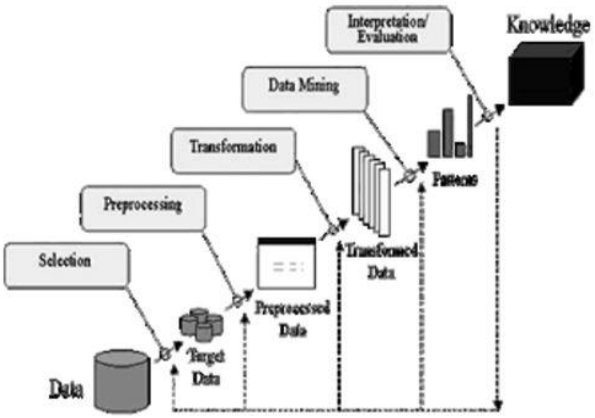
   Tahapan aplikasi teknik *data mining* adalah bagian dari salah satu proses *data mining*. Sehingga harus diperhatikan bahwa teknik-teknik yang ada tidak selamanya dapat mencukupi untuk melaksanakan *data mining* tertentu.

5. Evaluasi Pola yang Ditemukan

Tahap evaluasi pola yang ditemukan digunakan untuk menemukan pola-pola yang dengan ciri khas maupun prediksi yang bernilai. Apabila hasil yang ada tidak cocok dengan hipotesis yang ada maka terdapat cara lain yang dapat dilakukan.

6. Presentasi Pola yang Ditemukan

Selanjutnya tahap presentasi pola yang ditemukan digunakan untuk menghasilkan tindakan atau langkah yang harus dilakukan dari analisis yang diperoleh dengan bentuk pengetahuan yang dapat dipahami semua orang. Dalam presentasi ini visualisasi membantu menampilkan hasil *data mining*. Gambar 2.1 merupakan proses *data mining*.



Gambar 2. 1 Proses *Data Mining*

Dalam *data mining* terdapat beberapa metode pengolahan. Berikut adalah pengelompokan metode pengolahan *data mining* antara lain :

a. Classification

Classification adalah suatu teknik dengan melihat atribut dari kelompok data yang telah didefinisikan. Teknik ini dilakukan pada data dengan memanipulasi data yang ada, kemudian diklasifikasi sehingga dapat memperoleh hasil berupa sejumlah aturan. Salah satu contoh yang mudah

dan populer adalah *decision tree*. *Decision tree* merupakan model prediksi menggunakan struktur pohon atau struktur berhirarki. Perbedaan antara metode *clustering* dan *classification* terletak pada data karena metode *clustering* tidak ada variabel target dalam pengklusteran, sedangkan *classification* harus ada target variabel kategori.

b. Association

Association sebuah metode yang digunakan untuk mengetahui beberapa kejadian-kejadian khusus atau proses yang muncul pada setiap kejadian yang berhubungan dengan asosiasi. Salah satu contoh adalah *Market Basket Analysis*, yaitu salah satu metode asosiasi yang digunakan untuk menganalisis kemungkinan para pelanggan untuk membeli sejumlah barang secara bersamaan.

c. Clustering

Clustering sebuah metode yang digunakan untuk menganalisis pengelompokkan pada data yang berbeda, hampir sama dengan klasifikasi tetapi dalam proses pengelompokkannya belum diketahui saat dijalankan pada *tool data mining*. Metode yang sering digunakan adalah metode statistik atau *neural network.*

d. Predictive Modelling

Predictive modelling sebuah metode berupa metode pengolahan data mining dengan melakukan dengan cara prediksi atau peramalan. Dan tujuan dari metode ini yaitu untuk membentuk sebuah model prediksi suatu nilai yang mempunyai ciri-ciri khusus.

2.2.2. Klasifikasi

Klasifikasi merupakan suatu teknik yang diterapkan untuk memprediksi properti atau *class* pada setiap *instace* data. Teknik ini dilakukan dengan memanipulasi data yang tersedia dan telah diklasifikasikan sehingga dapat

menghasilkan sejumlah aturan (Ayu Rizqi Oktaviana, 2016). Berikut contoh yang mudah dan sering digunakan adalah *decision tree* atau pohon keputusan. *Decision tree* merupakan model prediksi yang menggunakan struktur pohon atau struktur berhirarki.

a. Definisi Data

Data adalah sesuatu objek yang diwakilkan dan suatu peristiwa yang mempunyai arti yang sangat penting bagi pengguna (*user*). Untuk mengetahui definisi dari data dalam klasifikasi maka dapat dilihat pada tabel 2.1 di bawah ini :

Tabel 2.1 Definisi Data

	2	2	3
1	Muda	Minum Alkohol	Sakit *Liver*
1	Muda	Tidak Minum Alkohol	Tidak Sakit *Liver*

Berdasarkan tabel 2.1 dapat diketahui ada 3 elemen yaitu :

1. *Instance* adalah sebuah data itu sendiri. Dan setiap *instance* dapat memiliki atribut dan *class*.
2. Atribut merupakan keterangan yang ada pada data itu sendiri. Maka setiap data dapat mempunyai atribut lebih dari satu dan atribut tersebut memiliki variabel diskret atau tidak saling berhubungan.
3. *Class* merupakan status pada setiap *instance*. *Class* merupakan kesimpulan pada setiap data, karena satu class biasanya ada pada satu data dan *class* tersebut memiliki variabel diskret atau variabel yang tidak saling berhubungan.

b. Tahapan Klasifikasi

Klasifikasi dalam *data mining* memiliki beberapa tahapan, berikut adalah tahapannya antara lain :

- Tahapan Pembangunan Model

  Tahapan pembangunan model merupakan sebuah tahapan pemodelan dalam data untuk menyelesaikan masalah pada klasifikasi *class* atau atribut. Tahapan ini diproses sesuai dengan *training set. Training set* telah memiliki informasi yang lengkap, dalam atribut maupun *class*nya.

- Tahapan Penerapan Model

  Penerapan pemodelan ini telah diproses terlebih dahulu dan untuk membuat atribut atau *class* dari sebuah atribut dari data baru atau *class* yang sebelumnya belum diketahui.

- Tahapan Evaluasi

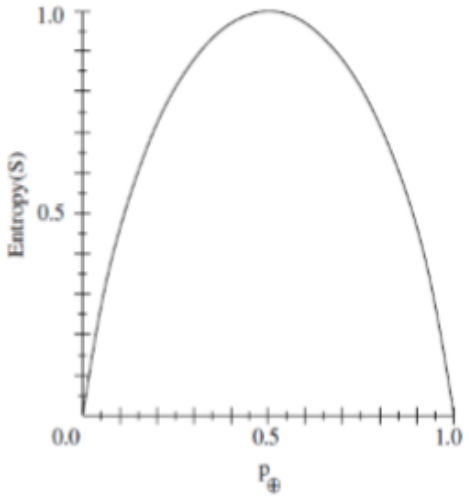
  Evaluasi merupakan hasil yang berasal dari tahapan sebelumnya. Dan evaluasi diolah dengan parameter yang terukur untuk menentukan dapat diterimanya sebuah model tersebut.

2.2.3. Algoritma C4.5

Algoritma C4.5 merupakan sebuah algoritma data yang sering digunakan dan diterapkan untuk sebuah proses klasifikasi data dengan atribut numerik dan kategorial. Hasil dari proses klasifikasi dapat berupa aturan-aturan yang digunakan untuk memprediksi nilai atribut bertipe diskret atau tidak saling berhubungan dari *record* yang baru. Algoritma C4.5 berasal dari pengembangan algoritma ID3, antara lain dapat mengatasi *missing data*, dapat mengatasi data kontinu, dan *pruning* (Saefulloh & Moedjiono, 2013).

A. *Entropy*

Entropy (S) adalah jumlah bit yang dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari beberapa data acak pada ruang sampel S. Sehingga *entropy* dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai *entropy* maka akan semakin baik *entropy* yang digunakan dalam mengekstrak suatu kelas. *Entropy* digunakan untuk mengukur ketidakaslian S. Gambar 2.2 merupakan tampilan dari grafik *entropy*.



Gambar 2. 2 Tampilan Grafik *Entropy*

Berikut definisi nilai *Entropy* :

$$\textit{Entropy}\ (S) = \sum_{i=0}^{n} -pi * log_2(pi) \quad \text{.....(1)}$$

Rumus (1) merupakan rumus yang digunakan dalam menghitung *entropy* untuk menentukan seberapa informatif atribut tersebut. Berikut keterangannya :

S : Himpunan kasus

n : Jumlah partisi

pi : Jumlah kasus pada partisi ke-*i*

B. *Information Gain*

Information Gain adalah informasi yang didapatkan dari perubahan *entropy* di suatu kumpulan data, baik melalui observasi maupun disimpulkan dengan cara melakukan partisipasi terhadap suatu set data.

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} -\frac{|Si|}{|S|} * Entropy\ (S) \quad \text{.....(2)}$$

Rumus (2) merupakan rumus dalam perhitungan *information gain* setelah menemukan nilai *entropy*. Berikut keterangannya :

S : Himpunan kasus

n : Jumlah partisi atribut A

|S*i*| : Jumlah kasus pada partisi ke-*i*

|S| : Jumlah kasus dalam S

C. *Split Info*

Split Info merupakan rumus yang menyatakan informasi potensial atau *entropy*. dapat dilihat dalam rumus (3). Dan keterangannya :

$$Split\ Info(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} log_2 \frac{S_i}{S} \text{.....(3)}$$

S : Himpunan kasus

A : Atribut

Si : Jumlah kasus pada partisi ke- *i*

D. *Gain Ratio*

Gain Ratio adalah modifikasi dari *information gain* yang digunakan untuk mengurangi bias atribut yang memiliki banyak cabang. *Gain ratio* memiliki sifat :

- Bernilai besar jika data menyebar rata
- Bernilai kecil jika semua data masuk ke dalam satu cabang

GainRatio (S, A) = $\frac{Gain(S,A)}{SpiltInfo(S,A)}$ .....(4) dan Keterangannya sebagai berikut:

S : Himpunan kasus

A : Atribut

Gain (S, A) : Information gain pada atribut A

SpiltInfo (S, A) : SplitInfo pada atribut A

2.2.4. *Software RapidMiner*

Software RapidMiner adalah sebuah *software* yang digunakan untuk mengolah *data mining*. Biasanya berkaitan tentang analisis teks, mengekstrak pola-pola dari *dataset* yang besar dan digabungkan dengan metode statistika, *database*, dan kecerdasan buatan. Analisis teks ini bertujuan untuk mendapatkan informasi bermutu tinggi dari teks yang diolah (Nugroho, 2014).

Di dalam *software RapidMiner* menyediakan prosedur *data mining* dan *machine lerning* termasuk ETL (*extraction, transformation, loading*), data *prepocessing*, visualisasi, *modelling* dan evaluasi. Prosesnya tersusun dari operator-operator, dideskripsikan dengan XML, dan dibuat dengan GUI. Disajikan dengan tulisan bahasa pemrograman *Java*.

2.2.5. Pohon Keputusan (*Decision Tree*)

Pohon keputusan adalah sebuah struktur yang digunakan untuk mengubah data menjadi pohon keputusan sehingga akan menghasilkan aturan-aturan keputusan besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Pohon keputusan yang dihasilkan oleh algoritma C4.5 dapat digunakan untuk klasifikasi (Nugroho, 2014). Berikut langkah-langkah dalam membuat pohon keputusan pada algoritma C4.5, yaitu :

- Pertama adalah memilih atribut sebagai akar, dan yang akan dipilih sebagai akar adalah atribut yang memiliki nilai *gain ratio* tertinggi dari semua atribut yang ada.
- Membuat cabang pada masing-masing nilai, artinya membuat cabang sesuai dengan jumlah nilai variabel *gain ratio* tertinggi.
- Membagi setiap kasus dalam cabang, berdasarkan perhitungan nilai *gain ratio* tertinggi dan perhitungan dilakukan setelah perhitungan nilai *gain ratio* tertinggi awal dan kemudian dilakukan proses perhitungan *gain ratio* tertinggi kembali tanpa menyertakan nilai variabel *gain ratio* awal.
- Terakhir adalah mengulangi proses pada setiap cabang sehingga semua kasus pada cabang memiliki kelas yang sama, mengulangi semua proses perhitungan *gain ratio* tertinggi untuk masing-masing cabang kasus sampai tidak bisa dilakukan proses perhitungan.