BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Studi

Beberapa penilitian yang terkait dalam penelitian ini adalah sebagai berikut:

1. "Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit" [7]

   Penelitian membahas tentang bagaimana mereacang sebuah aplikasi dengan menerapan teknik data mining dalam memprediksi kriteria nasabah yang berpotensi melakukan peminjaman pada suatu bank di kabupaten Bandung. Tujuan dilakukan penelitian ini yaitu untuk membantu menyelesaikan permasalahan tingginya biaya operasional marketing dengan data mining, karena metode yang digunakan pada umumnya digunakan untuk menganalisis data nasabah dengan cara mengklasifikasikan semua nasabah yang telah melunasi angsuran kreditnya ke dalam target pemasaran. Teknik yang diterapkan pada aplikasi yang dibangun adalah klasifikasi. Sedangkan metode klasifikasi yang digunakan adalah pohon keputusan (*decision tree*) dan algoritma yang dipakai adalah algoritma C4.5.

2. "Klasifikasi Masa Studi Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta Menggunakan Algoritma C4.5" [8]

   Penelitian ini membahas tentang pengklasifikasian masa studi mahasiswa. Penelitian dilakukan dengan menggunakan metode pohon keputusan (*decision tree*) dan algoritma yang digunakan dengan menggunakan algoritma C4.5. data yang diambil sebanyak 341 data mahasiswa yang telah lulus dari total data 2358 data yang ada. Atribut yang digunakan terdiri dari

jurusan sekolah, jenis kelamin, asal sekolah, rerata jumlah SKS per semester, dan peran menjadi asisten.

3. "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa" [9]

   Penelitian ini membahas tentang penggunaan pendekatan teknik klasifikasi *data mining* dengan menggunakan algoritma C4.5 yang diterapkan dalam menentukan jurusan dalam bidang studi yang akan diambil oleh mahasiswa sesuai latar belakang mahasiswa supaya tidak salah saat memilih jurusan dalam perguruan tinggi. Parameter yang digunakan dala pemilihan jurusan adalah Indeks Prestasi Komutatif (IPK) pada semester 1 dan semester 2. Hasil eksperimen dan evaluasi menunjukan kesesuan jurusan mahsiswa dengan tingkat akurasi sebesar 93,31% dan akurasi rekomendasi jurusan sebesar 82,64%.

4. "Klasifikasi Data Nasabah Sebuah Asuransi menggunakan Algoritma C4.5" [4]

   Penelitian ini membahas tentang pengklasifikasian data nasabah pada sebuah asuransi sehingga dapat dicari pola status nasabah untuk dijadikan sebagai bahan analisis dalam menentukan calon nasabah di masa mendatang. Atribut yang digunakan dalam penelitian ini adalah penghasilan, premi dasar, cara pembayaran, mata uang, dan status. Hasil yang didapatkan dalam penelitian ini adalah aplikasi yang dapat menyimpulkan bahwa rata-rata nasabah memiliki status L (*Lapse*) dikarenakan pembayaran premi yang melebihi 10% dari penghasilan. Dengan presentase atribut premi dasar dan penghasilan, maka diketahui rata-rata status nasabah memiliki P (*Presistent*) atau (*Lapse*).

Dari beberapa penilitian terkait yang telah disebutkan diatas, maka didapatkan *state of the art* sebagai berikut :

Tabel 2.1 State Of The Art

No	Judul	Penulis	Tahun	Ringkasan
1.	Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit	Angga Ginanjar Mabrur Dan Rina Lubis	2012	Perancangan aplikasi dengan teknik data mining menggunakan pohon keputusan dan algoritma C4.5 untuk prediksi criteria nasabah
2	Klasifikasi Masa Studi Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta Menggunakan Algoritma C4.5	Yusuf Sulistyo Nugroho Dan Setyawan	2014	Metode pohon keputusan dan algoritma C4.5 untuk mengklasifikasi masa studi mahasiswa
3	Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa	Liliana Swastina	2012	Menerapkan algoritma C4.5 untuk menentukan jurusan yang akan diambil

				mahasiswa sesuai latar belakang mahasiswa agar tidak salah mengambil jurusan
4	Klasifikasi Data Nasabah Sebuahn Asuransi Menggunakan Algorima C4.5	Sunjana	2010	Mengklasifikasi data nasabah menggunakan algoritma C4.5 agar dapat mencari pola nasabah dalam menentukan calon nasabah di masa yang akan datang

2.2 Tinjauan Pustaka

2.2.1 Koperasi Simpan Pinjam

Koperasi simpan pinjam merupakan salah satu lembaga keuangan bukan bank yang bertugas memberikan pelayanan masyarakat, berupa pinjaman dan tempat penyimpanan uang bagi masyarakat [2]. Sumber dana koperasi simpan pinjam diperoleh dari simpanan sukarela, simpanan wajib, simpanan pokok dan bahkan dari berbagai lembaga pemerintah maupun lembaga swasta yang mengalami kelebihan dana.

Jenis-jenis produk pinjaman yang ada di koperasi simpan pinjam [10]:

1. Pinjaman Musiman (*Revolving*)

Pinjaman musiman (*revolving*) merupakan pinjaman yang diberikan kepada anggota atau calon anggota sebagai modal kerja dengan jangka waktu 1,3,6, dan 12 bulan. Dimana setiap bulan debitur diwajibkan untuyk membayar bunganya saja dan pembayaran pokok pinjaman dilakukan pada akhir jangka waktu atau jatuh tempo.

2. Pinjaman Rekening Koiran

   Pinjaman rekening Koran merupakan pinjaman yang diberikan kepada anggota maupun calon anggota untuk pembiayaan modal kerja, khususnya untuk para pengembang atau developer dengan jangka waktu 12 bulan,. Dimana untuk pencairan pinjaman menggunakan sistem termin sesuai dengan perjanjian yang sudah ditentukan.

   Pada jenis pinjaman ini debitur hanya diwajibkan untuk membayar besarnya bunga setiap bulannya, akan tetapi jika berkeinginan untuk membayar pokok juga dapat dilakukan.

3. Pinjaman Angsuran

   Pinjaman angsuran merupakan pinjaman yang diberikan untuk anggota atau calon anggota sebagai modal kerja dengan jangka waktu sampai dengan 84 bulan atau 7 tahun dengan sistem pembayaran dilakukan secara dicicil atau diangsur.

2.2.2 Data

Himpunan data (data-set) merupakan kumpulan dari objek dan atributnya. Atribut meruapakan sifat atau karakteristik dari suatu objek. Atribut juga dikenal sebagai variable, field, karakteristik atau fitur. Sedangkan objek merupakan kumpulan dari atribut. Objek juga disebut dengan *record*, titik, kasus, *sample*, entitas atau *instance* [3].

Tipe-tipe dari himpunan data (dat-set) antara lain :

1. Data Matrix

   Jika objek data mempunyai himpunan atribut numerik yang sama, maka objek data tersebut dapat dianggap sebagai titik-titik dalam ruang multi dimensi, dimana masing-masing dimensi menyatakan satu atribut yang berbeda.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Gambar 2.1 Contoh Data Matrix

2. Data Dokumen

   Dimana dalam tipe data dokumen, tiap dokumen menjadi satu vector '*term*'. Tiap term merupakan satu komponen (atribut) dari vector tersebut. Nilai dari tiap komponen menyatakan beberapa kali kemunculan term tersebut dalam suatu dokumen.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Gambar 2.2 Contoh Data Dokumen

3. Data Transaksi

   Data transaksi merupakan sebuah tipe khusus dari *record* data, dimana tiap *record* (transaksi) meliputi satu set item.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Gambar 2.3 Contoh Data transaksi

4. Data *Graph*

   Data *graph* merupakan data dalam bentuk *graph* yang terdiri dari simpul (*node*) dan rusuk (*edge*).

5. Data Terurut (*Ordered* Data)

   Data-data yang memperhatikan urutan nilai-nilainya.

   Beberapa atribut data berdasarkan jumlah nilainya yaitu :

1. Atribut Diskrit

   Atribut diskrit yaitu atribut yang hanya menggunakan sebuah himpunan nilai berhingga dan himpunan nilai tak berhingga yang dapat dihitung.

2. Atribut Kontinyu

   Atrubut kontinyu yaitu atribut yang menggunakan bilangan riil sebagai nilai atrubut.

2.2.3 Data Mining

Data mining adalah proses yang memperkerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis.

Data mining merupakan proses iteratif dan interaktif untuk menemukan pola atau model baru yang sahih (sempurna), bermanfaat dan dapat dimengerti dala suatu *database* yang sangat besar (*massive databases*).

Data mining berisi pencarian trend atau pola yang diinginkan dalam database besar untuk membantu penggambilan keputusan di waktu yang akan datang. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa data yang berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mumgkin sajamenggunakan perangkat pendukung keputusan yang lainnya [3].

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang yang dapat dilakukan, yaitu [6]:

1. Deskripsi

   Terkadang peneliti dan analisis secara sederhana ingin mencoba nencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.

   Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

   Estimasi hamper sama dengan klasifikasi, kecuali variable target estimasi lebih kea rah numeric daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variable target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variable target dibuat berdasarkan nilai variable predisi.

3. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil aka nada di masa mendatang.

   Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

   Dalam klasifikasi, terdapat target variable kategori.

5. Pengklusteran

   Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *recor-record* dalam kluster lain

6. Asosiasi

   Tugas asosiasi dala data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisinis lebih umum disebut analisis keranjang belanja.

Operasi-operasi data mining menurut sifatnya dibedakan menjadi [3]:

1. Prediksi (*prediction driven*)

   Prediksi untuk menjawab pertanyaan apa dan sesuatu yang bersifat remang-remang atau transparan.

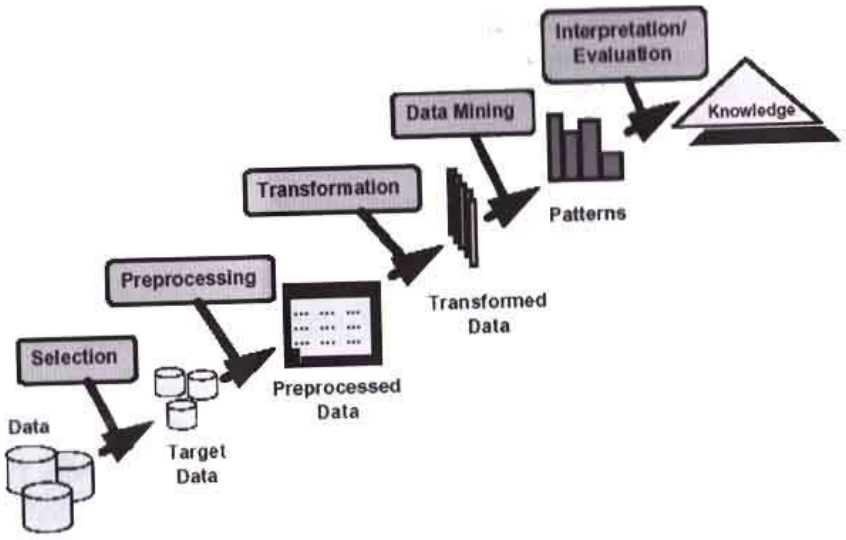
2. Penemuan (*discovery driven*)

   Penemuan bersifat transparan dan untuk menjawab pertanyaan.

Tahaapan proses dalam penggunaan *data mining* yang merupaka proses *Knowledge Discovery in Databases* (KDD) sebagai berikut [3]:

1. Memahami *domain* aplikasi untuk mengetahui dan memanggil pengetahuan awal serta apa sasaran pengguna.
2. Membuat target data-set yang meliputi pemilaihan data dan fokus pada sub-set data.

3. Pembersihan transformasi data meliputi pemilihan eliminasi derau, outliers, missing value serta pemilihan fitur dan reduksi dimensi.
4. Penggunaan algoritma *data mining* yang terdiri dari asosiasi, sekuensial, klasifikasi, klasterisasi, dll.
5. Interpretasi, evaluasi dan visualisasi pola untuk melihat apakah ada sesuatu yang baru dan menarik dan dilakukan iterasi jika diperlukan.

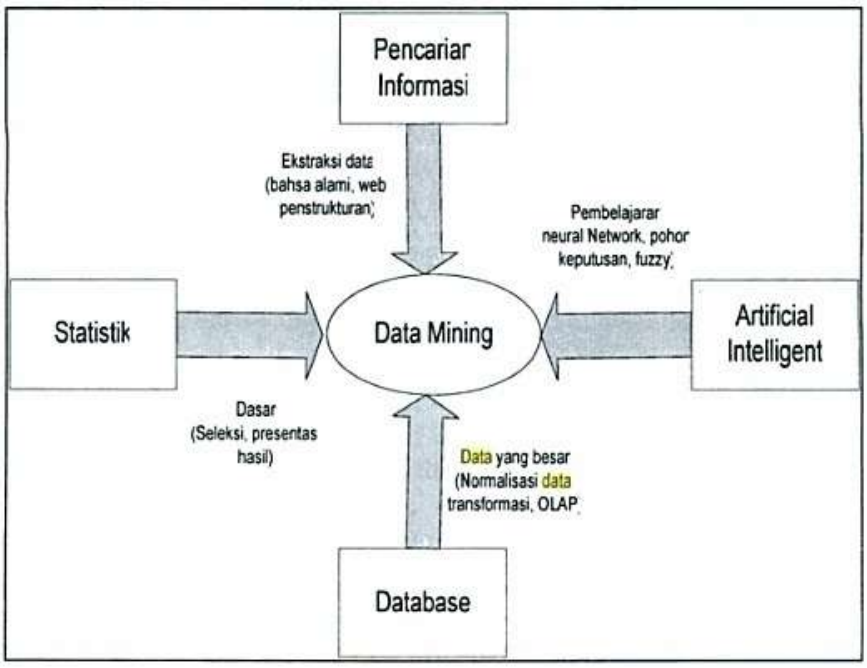


Gambar 2.4 Proses *Data mining*

Beberapa faktor yang mendorong kemajuan luar biasa dalam bidang *data mining*, antara lain [6]:

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam database yang handal.
3. Adanya peningkatan akses data melalui navigasi web dan internet.
4. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).

Konsep data mining yaitu data mining memiliki akar yang panjang dari bidang yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistic, database, dan juga *information retrieval* (Pramudiono,2006).



Gambar 2.5 Bidang Ilmu Data Mining

2.2.4 Klasifikasi

Klasifikasi merupakan salah satu proses data mining, pada klasifikasi diberikan sejumlah record yang dinamakan *training set*, yang terdiri dari kelas untuk *record*. Tujuan klasifikasi untuk menemukan model dari *training set* yang membedakan *record* kedalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasi *record* yang kelasnya belum diketahui sebelumnya.

Komponen-komponen utama dari proses klasifikasi antara lain [5]:

1. Kelas, merupakan variabel tidak bebas yang merupakan label dari hasil klasifikasi. Sebagai contoh adalah kelas loyalitas pelanggan, kelas badai atau gempa bumi, dan lain-lain.
2. Prediktor, merupakan variabel bebas suatu model berdasarkan berdasarkan dari karakteristik atribut data yang diklasifikasi, misalnya merokok, minum-minuman beralkohol, tekanan darah, status perkawinan, dan sebagainya.
3. Set data pelatihan, merupakan sekumpulan data lengkap yang berisi kelas-kelas predictor untuk dilatih agar model dapat mengelompokan ke dalam kelas yang tepat. Contohnya adalah grup pasien yang telah di-test terhadap serangan jantung, grup pelanggan di suatu supermarket, dan sebagainya.
4. Set data uji, berisi data-data baru yang akan dikelompokan oleh model guna mengetahuai akurasi dari model yang telah dibuat.

2.2.5 Pohon Keputusan (*Decision Tree*)

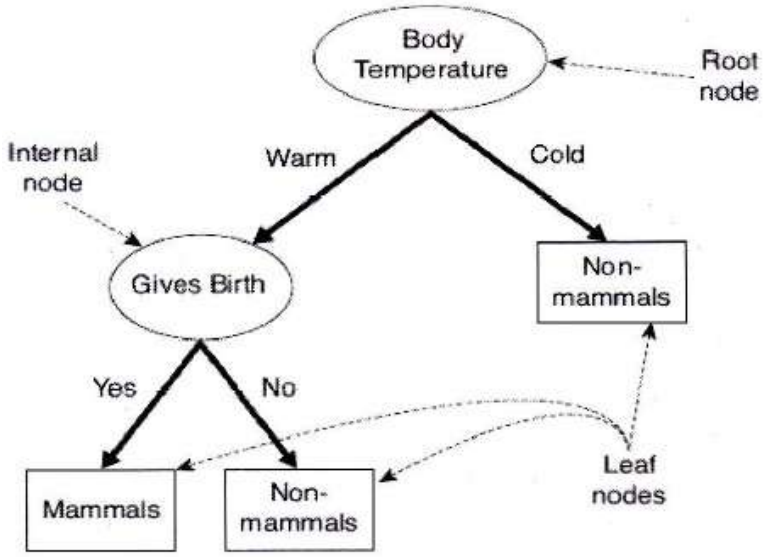
Pohon keputusan (*Decision Tree*) merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input menjadi variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, sangat bagus sekali sebgai langkah awal dalam

proses pemodelan bahkan ketika diadakan sebagai model akhir dari beberapa teknik lain.

Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain.

Algoritm yang dapat dipakai dalam pembetukan pohon keputusan, yaitu algoritma ID3, CART, dan C4.5 (merupakan pengembangan dari algoritma ID3).

Proses pada pohon keputusan dengan mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule* [6].



Gambar 2.6 Contoh Pohon Keputusan

Pada pohon keputusan terdapat 3 jenis *nose*, yaitu [11]:

1. *Rote Node*, merupakan *node* yang paling atas. Pada *node* ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu.

2. *Internal Node*, merupakan *node* pencabangan. Pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua.
3. *Leaf Node* atau *Terminal Node*, merupakan *node* akhir. Pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

2.2.6 Algoritma C4.5

Algoritma C4.5 merupakan kelompok algoritma pohon Keputusan (*decision tree).* Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data [4].

Algoritma C4.5 merupakan pengembangan dari algoritma ID3, dimana pengembangan dilakukan dalam hal bisa mengatasi missing data, bisa mengatasi data kontinyu, pruning [11]. Algoritma C4.5 memiliki kelebihan yaitu mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar (pohon keputusan) [5].

Secara umum algoritma C4.5 Untuk membangun pohon keputusan adalah sebagai berikut [6]:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \text{.....} (2.1)$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si|: jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Sedangkan untuk menghitung entropy digunakan rumus:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \text{.....} (2.2)$$

Keterangan:

S: himpunan kasus

A: fitur

n: jumlah partisi S

pi: proporsi dari S_i terhadap S

2.2.7 Tahapan Data Mining Pada Penelitian

Beberapa tahapan data mining pada penelitian diantaranya :

1. Mempersiapkan data-data yang akan digunakan dalam penelitian. Misal, dat pengajuan pinjaman.
2. Melakukan pengolahan data awal atau pembersihan data, yaitu dengan menentukan atribu-atribut apa saja yang akan digunakan dan diperlukan pada penelitian, menhilangkan atribut-atribut yang tidak perlu dan membersihkan data yang kosong atau missing. Misal, atribut yang digunakan yaitu atribut jenis kelamin, data blacklist bank dan sebagainya.

3. Membagi data menjadi dua bagian yaitu 80% data training dan 20% data testing dari keseluruhan jumlah data.
4. Penerapan Algoritma C4.5 pada data dengan menghitung nilai entropy dan gain. Nilai gain tertinggi akan digunakan sebagai node akar.
5. Pembuatan pohon keputusan dari hasil nilai entropy dan gain.
6. Penentuan rules dari pohon keputusan yang telah dibentuk untuk mengklasifikasi data testing.
7. Proses pengujian yaitu dengan menguji data testing berdasarkan dari pembentukan rules pada data training.
8. Perhitungan nilai *accuracy, precision,*dan *recall.*\

2.2.8 Contoh Kasus Algoritma C4.5

Untuk memudahkan penjelasan algoritma C4.5, berikut ini contoh kasus algoritma C4.5 :

Tabel 2.2 Keputusan Bermain Tenis

NO	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Cloudy	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	Yes
7	Cloudy	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Cloudy	Mild	High	True	Yes
13	Cloudy	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Berikut ini adalah penjelasan lebih rinci mengenai masing-masing langkah:

1. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Enteropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, HUMIDITY dan WINDY. Setelah itu lakukan perhitungan Gain untuk masing-masing atribut.

Tabel 2.3 Perhitungan Node 1

NODE			JUMLAH KASUS (S)	NO (S_1)	YES (S_2)	ENTROPY	GAIN
1	TOTAL		14	4	10	0.863120569	
	OUTLOOK						0.258521037
		CLOUDY	4	0	4	0	
		RAINY	5	1	4	0.721928095	
		SUNNY	5	3	2	0.970950594	
	TEMPERATURE						0.183850925
		COOL	4	0	4	0	
		HOT	4	2	2	1	
		MILD	6	2	4	0.918295834	
	HUMIDITY						0.370506501
		HIGH	7	4	3	0.985228136	
		NORMAL	7	0	7	0	
	WINDY						0.005977711
		FALSE	8	2	6	0.811278124	
		TRUE	6	4	2	0.918295834	

Baris TOTAL kolom Entropy dihitung, sebagai berikut :

$$Entropy(Total) = \left(-\frac{4}{14} * log_2\left(\frac{4}{14}\right)\right) + \left(-\frac{10}{4} * log_2\left(\frac{10}{4}\right)\right)$$

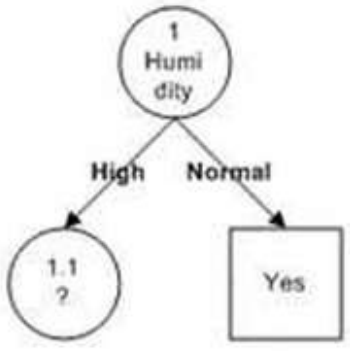
$Entropy(Total) = 0.863120569$

Sementara itu, nilai Gain pada baris OUTLOOK dihitung sebagai berikut :

$$Gain(Total, Outlook)$$
$$= Entropy(Total) - \sum_{i=1}^{n} \frac{|Outlook_i|}{|Total|} * Entropy(Outlook_i)$$
$$Gain(Total, Outlook)$$
$$= 0.863120569 - \left(\left(\frac{4}{10} * 0\right) + \left(\frac{5}{14} * 0.723\right) + \left(\frac{5}{14} * 0.97\right)\right)$$
$$Gain(Total, Outlook) = 0.23$$

Dari hasil di dapat diketahui bahwa atribut dengan Gain tertinggi adalah HUMIDITY yaitu sebesar 0.37. dengan demikian HUMIDITY dapat menjadi node akar. Ada 2 nilai atribut dari HUMIDITY yaitu HIGH dab NORMAL. Dari kedua nilai atribut tersebut, nilai atribut NORMAL sudah mengklasifikasikan kasus menjadi1 yaitu keputusan-nya Yes, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut HIGH masih perlu dilakukan perhitungan lagi.

Dari hasil tersebut dapat digambarkan pohon keputusan sementara seperti :



Gambar 2.7 Pohon Keputusan Hasil Perhitungan Node 1

2. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari

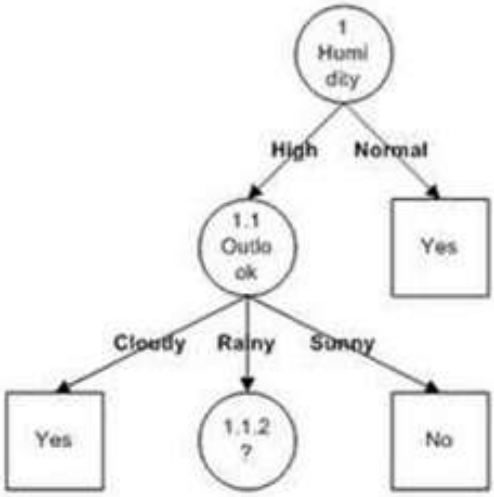
semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH. Setelah itu lakukan perhitungan Gain untuk masing-masing atribut.

Tabel 2.4 Perhitungan Node 1.1

Node			Jml Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1.1	HUMIDITY-HIGH		7	4	3	0.985228136	
	OUTLOOK						0.69951385
		CLOUDY	2	0	2	0	
		RAINY	2	1	1	1	
		SUNNY	3	3	0	0	
	TEMPERATURE						0.020244207
		COOL	0	0	0	0	
		HOT	3	2	1	0.918295834	
		MILD	4	2	2	1	
	WINDY						0.020244207
		FALSE	4	2	2	1	
		TRUE	3	2	1	0.918295834	

Dari hasil pada tabel diatas dapat diketahui bahwa atribut dengan Gain tertinggi adalah OUTLOOK yaitu sebesar 0.67. dengan demikian OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH. Ada 3 nilai atribut dari OUTLOOK yaitu CLOUDY, RAINY, dan SUNNY. Dari ketiga nilai atribut tersebut, nilai atribut CLOUDY sudah mengklasifikasikan kasus menjadi 1 yaitu keputusan-nya Yes dan nilai atribut SUNNY sudah mengklasifikasikan kasus menjadi satu dengan keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut, tetapi untuk nilai atribut RAINY masih perlu dilakukan perhitungan lagi.

Pohon keputusan yang terbentuk sampai tahap ini ditunjukan sebagai berikut :



Gambar 2.8 Pohon Keputusan Hasil Perhitungan Node 1.1

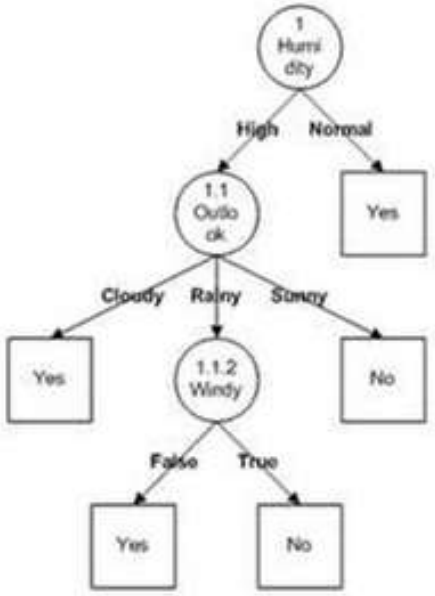
3. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keoutusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY. Setelah itu lakukan perhitungan Gain untuk masing-masing atribut.

Tabel 2.5 Perhitungan Node 1.1.2

Node			Jml Kasus (S)	Tidak (S1)	Ya (S2)	Entropy	Gain
1.1.2	HUMIDITY-HIGH dan OUTLOOK-RAINY		2	1	1	1	
	TEMPERATURE						0
		COOL	0	0	0	0	
		HOT	0	0	0	0	
		MILD	2	1	1	1	
	WINDY						1
		FALSE	1	0	1	0	
		TRUE	1	1	0	0	

Dari hasil pada tabel diatas dapat diketahui bahwa atribut dengan Gain tertinggi adalah WINDY yaitu sebesar 1. Dengan demikian WINDY dapat menjadi node cabang dari nilai atribut RAINY. Ada 2 nilai atribut dari WINDY yaitu

FALSE dan TRUE. Dari kedua nilai atribut tersebut, nilai atribut FALSE sudah mengklasifikasikan kasus menjadi 1 yaitu keputusan-nya Yes dan nilai atribut TRUE sudah mengklasifikasikan kasus menjadi satu dengan keputusan No, sehingga tidak perlu dilakukan perhitungan lebih lanjut untuk nilai atribut ini.



Gambar 2.9 Pohon Keputusan Hasil Perhitungan Node 1.1.2

Dengan memperhatikan pohon keputusan pada gambar diatas, diketahui bahwa semua kasus sudah masuk dalam kelas. Dengan demikian, pohon keputusanpada gambar diatas merupakan pohon keputusan terakhir yang dibentuk [6].

2.2.9 Confusion Matrix

Confusion Matrix adalah *tools* yang digunakan untuk evaluasi model klasifikasi untuk memerkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelasyang asli dari inputan atau dengan kata lain berisi informasi nilai *actual* dan prediksis pada klasifikasi [12].

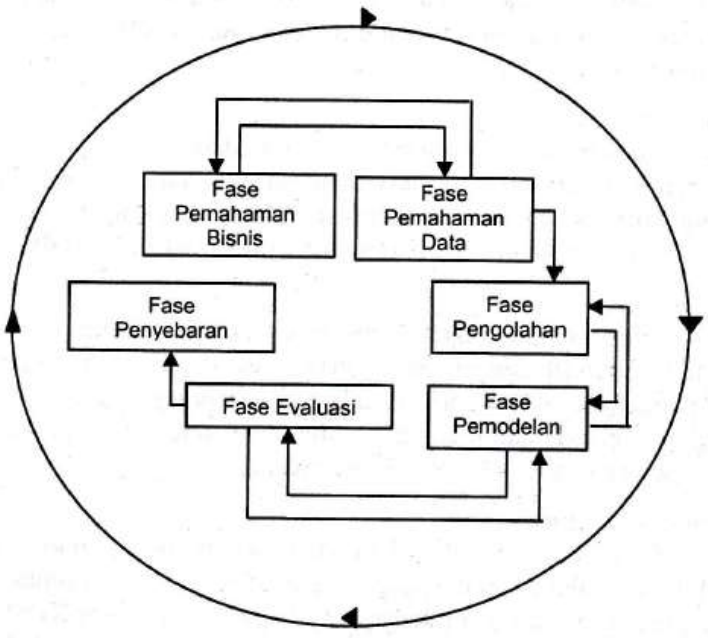
Tabel 2.6 Cofusion Matrix

Classification	*Predicted class*	
	Class = Yes	Class = No
Class = Yes	a (t*rue positives*_TP)	b (*false negatives*_FN)
Class = No	c (*false positives*_FP)	d (*true negatives*_TN)

2.2.10 Corrs-Industry Standard Process for *Data Mining* (CRISP-DM)

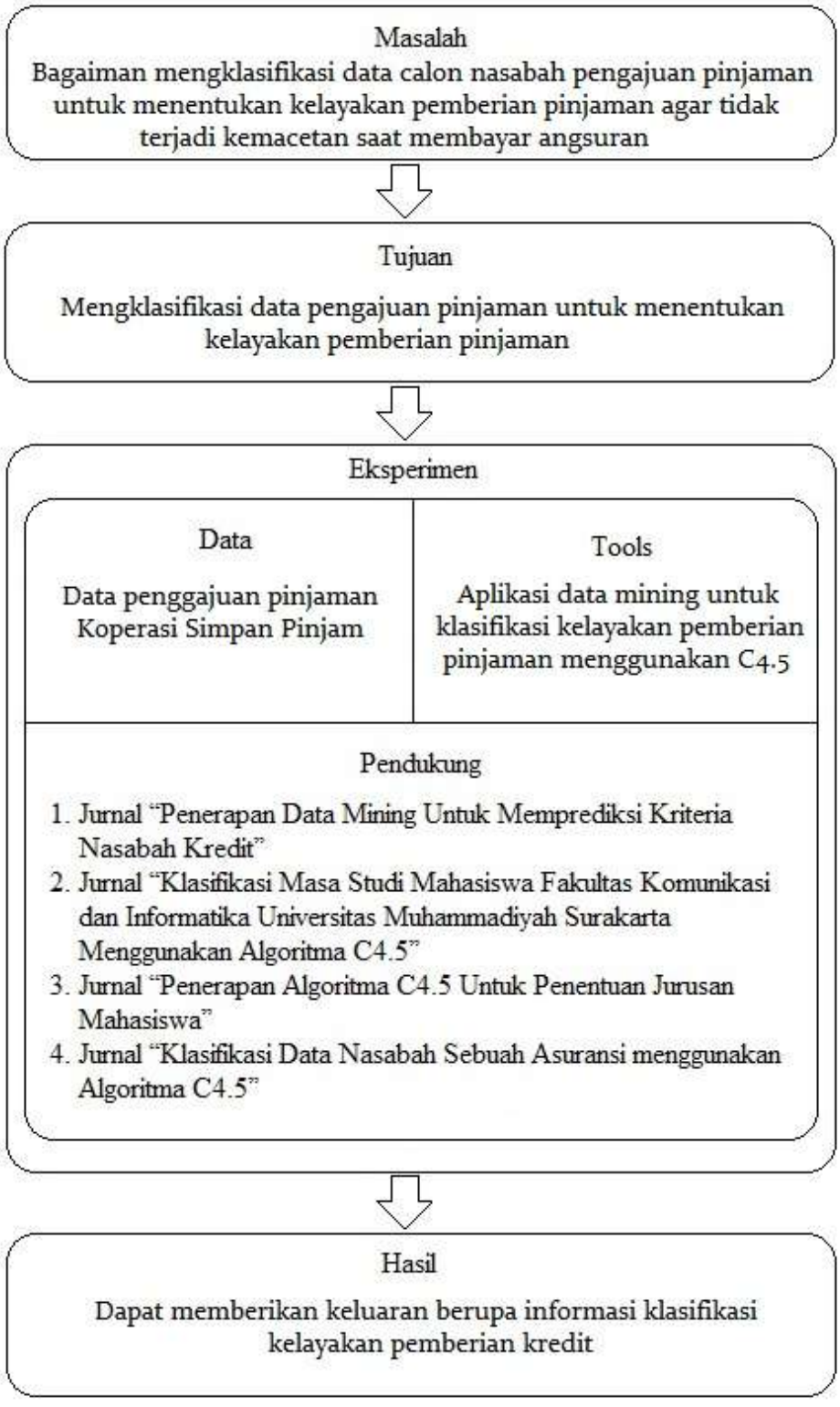
CRISP-DM dikembangkan pada tahun 1996 oleh analisis dari beberapa industry seperti Daimler Chrysler, SPPS, dan NCR. CRISP-DM menyediakan standar proses *data mining* sebagai strategi pemecahan masalah secara umun dari bisnis atau unit penelitian.

Dalam CRISP-DM, sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase. Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan bergantung pada keseluruhan dari fase sebelumnya. Hubungan penting antar fase digambarkan dengan panah [6].



Gambar 2.10 Proses *Data Mining* menurut CRISP-DM

2.3 Kerangka Pemikiran



Gambar 2.11 Kerangka Pemikiran