BAB II

TINJAUAN PUSTAKA DAN LANDASAR TEORI

2.1 Penelitian Terkait

Penelitian mengenai penggunaan Metode Klasifikasi dengan algoritma C4.5 dalam pengelompokan data siswa berdasarkan prestasi dan kriteria tidak mampu bukanlah penelitian yang dilakukan pertama kalinya, sebelumnya sudah ada penelitian yang memiliki keterkaitan dalam topik penelitian ini, yaitu sebagai berikut.

- **Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif[8].**

  berdasarkan hasil penelitian ini adalah klasifikasi data mining untuk memprediksi mahasiswa nonaktif. Pada penelitian ini menyimpulkan bahwa decision tree merupakan algoritma yang paling akurat. Logistic regression merupakan algoritma yang paling dominan di antara algoritma yang lain meskipun akurasinya paling rendah. Berdasarkan nilai AUC, logistic regression, decision tree, naïve bayes, dan neural network masuk dalam kategori excellent classification.

- **Penerapan algoritma C4.5 untuk klasifikasi tingkat keganasan kanker payudara[9]**

  Penelitian ini mebahas tentang penklasifikasi tingkat kegaasan kanker payudara menggunkan algoritma c4.5 yang memiliki akurasi sebasaar 98.57%. akurasi diperoleh dari kesesuain antara prediksi klasifikasi dan hasil klasifikasi. Data yang digunakan dalam penelitian ini menggunakan sampel dari public dataset UCI, karena data dari pasien kanker payudara merupan data private yang sulit untuk didapatkan.Hasil penelitian ini dari permodelan dari algoritma c.45 dengan pembobotan atribut adalah pembentukan rules. Rules yang

terbentuk 23 rules. Pada nantinya rules tersebut akan diimplementasikan kedalam program. Solusi dari permasalahan penelitian ini dapat disimpulkan bahwa klasifikasi tingkat keganasan kanker payudara dapat diselesaikan menggukan teknik data mining yaitu algoritma c4.5

- **Penerapan Data Mining Untuk Rekomendasi Beasiswa Tepat Sasaran Menggunakan Algoritma C4.5 [7]**

  Dalam penelitian ini, penulis menerapan metode pohon keputusan yang dapat digunakan sebagai klasifikasi untuk menentukan rekomendasi beasiswa pada SMA Muhammadiyah Gubug Penelitian ini menggunakan dataset mahasiswa yang mengajukan beasiswa.Pada data tersebut dalam 4 tahun terakhir (2010-2013), Penerapan metode pohon keputusan terhadap data siswa SMA Muhammadiyah Gubug memiliki tingkat akurasi yang cukup baik dalam menyelesaikan klasifikasi rekomendasi beasiswa, dengan demikian metode pohon keputusan merupakan metode yang cukup sesuai untuk penyelesaian studi kasus dalam pemilihan siswa yang mendapatkan rekomendasi beasiswa. Tingkat akurasi yang dihasilan oleh metode tersebut adalah 77%.

- **Penerapan Algoritma C4.5 Pada Progam Klasifikasi mahasiswa Dropout [10]**

  berdasarkan hasil penelitian pada klasifikasi mahasiswa potensi dropout dapat diambil beberapa kesimpulan sebagai berikut, Klasifikasi mahasiswa yang menggunakan algoritma C4.5 untuk mengklasifikasikan mahasiswa aktif dan dropout. Hasil evaluasi dan validasi dengan confussion matrix menunjukkan tingkat akurasi pada algoritma C4.5 sebesar 97,75%. Hasil evaluasi dan validasi dengan ROC/AUC menunjukkan nilai lebih dari 0,9 sehingga dapat dimasukkan kedalam excellent classification Penerapan rule

dari algoritma C4.5 yang digunakan dalam klasifikasi mahasiswa potensi dropout terhadap data baru diperoleh hasil evaluasi dan validasi dengan confussion matrix menghasilkan tingkat akurasi sebesar 90,00%

- **Rancang Bangun Sistem Rekomendasi Beasiswa Menggunakan Algoritma Klasifikasi C4.5 Pada Universitas Dian Nuswantoro [6].**

  Dalam penelitian ini, penulis membangun sebuah sistem yang dapat digunakan sebagai klasifikasi untuk menentukan rekomendasi beasiswa pada universitas dianuswantoro. Penelitian ini menggunakan dataset mahasiswa yang mengajukan beasiswa. Pada penelitian ini menggunakan metode klasifikasi dengan algoritma C4.5. Penelitian tersebut menyimpulkan bahwa klasifikasi menggunakan algoritma c4.5 dipengaruhi oleh beberapa hal seperti jenis, jumlah, isi data set dan jumlah partisi dataset. Dari hasil uji coba partisi dataset ditemukan tingkat akurasi sebanyak 92,31%.

Tabel 2.1 Penelitian Terkait

NO	Penulis	Judul	Masalah	Metode	Hasil
1	Khafizh Hastuti, 2012	Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif	Mahasiswa yang memiliki status non aktif mempunyai kecenderungan untuk drop out dan tingginya prosentase mahasiswa dengan status nonaktif	Klasifikasi	Logistic Regression merupakan algoritma yang paling dominan diantara algoritma yang lain meskipun akurasinya paling rendah. Berdasarkan

			mempengaruhi nilai akreditasi universitas		nilai AUC, *logistic regression, decision tree, naïve bayes, dan neural network* masuk dalam kategori excellent classification.
2	D. A. Nursela, 2014	Penerapan algoritma C4.5 untuk klasifikasi tingkat keganasan kanker payudara	Ada bnyak penderita kanker payudara, kanker payudara juga merupakan penyakit yang sangat ganas dan mengharuskan penderitanya untuk melakukan pemeriksaan yang intensif	Klasifikasi	Dari permasalahan tersebut dapat disimpulkan bahwa klasifikasi tingkat keganasan kanker payudra dapat diselesaikan menggunakan teknik data mining yaitu algoritma c4.5, karena rules yang terbentuk sederhana, akurasi yang dihasilkan yaitu sebesar 98,57%.

3	Pradega Shella	Sistem Pendukung Keputusan Menggunakan Decission Tree Dalam Pemberian Beasiswa Di Sekolah Menegah Pertama.	yaitu bagaimana menerapkan Algoritma C4.5 untuk prediksi pemberian beasiswa di Sekolah Menengah Pertama Negeri 2 Rembang sehingga mampu menjadi pendukung keputusan atas pihak SMP N 2 Rembang dalam proses pemberian beasiswa yang akan datang.	Klasifikasi	Akurasi yang dihasilkan dari pemodelan Algoritma C4.5 sebesar 86,91%. Dengan jumlah true positif (tp) sebanyak 68 record, false positif (fp) sebanyak 16 record, jumlah true negative (tn) sebanyak 171 record, dan jumlah false negative (fn) sebanyak 20 record.
4	Anik Andriani, 2012	Penerapan Algoritma C4.5 Pada Progam Klasifikasi mahasiswa	Tingginya jumlah mahasiswa Dropout pada perguruan tinggidapat diminimalisir	Klasifikasi	Penerapan rule dari algoritma C4.5 yang digunakan dalam klasifikasi mahasiswa potensi dropout

			dengan kebijakan dari perguruan tinggi untuk mengarahkan dan mencegah mahasiswa dari dropout		terhadap data baru diperoleh hasil evaluasi dan validasi dengan confussion matrix menghasilkan tingkat akurasi sebesar 90,00.
5	Yosoa Putra Raharja 2014	Rancang Bangun Sistem Rekomendasi Beasiswa Mengguna kan Algoritma Klasifikasi C4.5 Pada Universitas Dianuswator o	Banyaknya mahasiswa memiliki status mangkir. Hal Tersebut disebabkan karena kondisi ekonomi yang tidak mampu	klasifikasi	Dari hasil uji coba partisi data set ditemukan tingkat akurasi tertinggi pada jumlah partisi data sebanyak 90% dan menghasilkan akurasi sebanyak 92,31%. Dan semakin besar jumlah partisi data maka akan menghasilkan jumlah akurasi yang semakin tinggi pula.

2.2 Data Mining

Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran computer (*machine learning*) untuk menganalisis dan mengekstrasi pengetahuan (*knowledge*) secara otomatis. Definisi lain diantaranya adalah pembelajaran berbasis induksi (*induction-basedlearning*) adalah proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsepkonsep yang akan dipelajari. *Knowledge Discovery in Databases (KDD)* adalah penerapan metode *saintifik* pada data mining. Dalam konteks ini data mining merupakan satu langkah dari proses KDD **[**2**]**.

Beberapa teknik dan sifat data mining adalah sebagai berikut:

- *Classification [Predictive]*
- *Clustering [Descriptive]*
- *AssociationRule Discovery [Descriptive]*
- *Regression [Predictive]*
- *DeviationDetection [Predictive]*

2.2.1 Pengelompokan Data Mining

Ada beberapa pengelempokan data mining, data mining akan dikelompokan berdaasarkan tugas-tugas yang dikerjakan , antaralain [13] :

1. **Deskripsi**

   mancari cara untuk menentukan sebuah pola dan kecenderungan yang ada dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan

2. **Estimasi**

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik daripada kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilau variabel prediksi.

3. **Prediksi**

   Prediksi hampir sama dengan klasifikasi dan estimasi. Kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

4. **Klasifikasi**

   Dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan suhu dalam tiga kelas, yaitu suhu panas, suhu sejuk, suhu dingin.

5. **Pengklusteran**

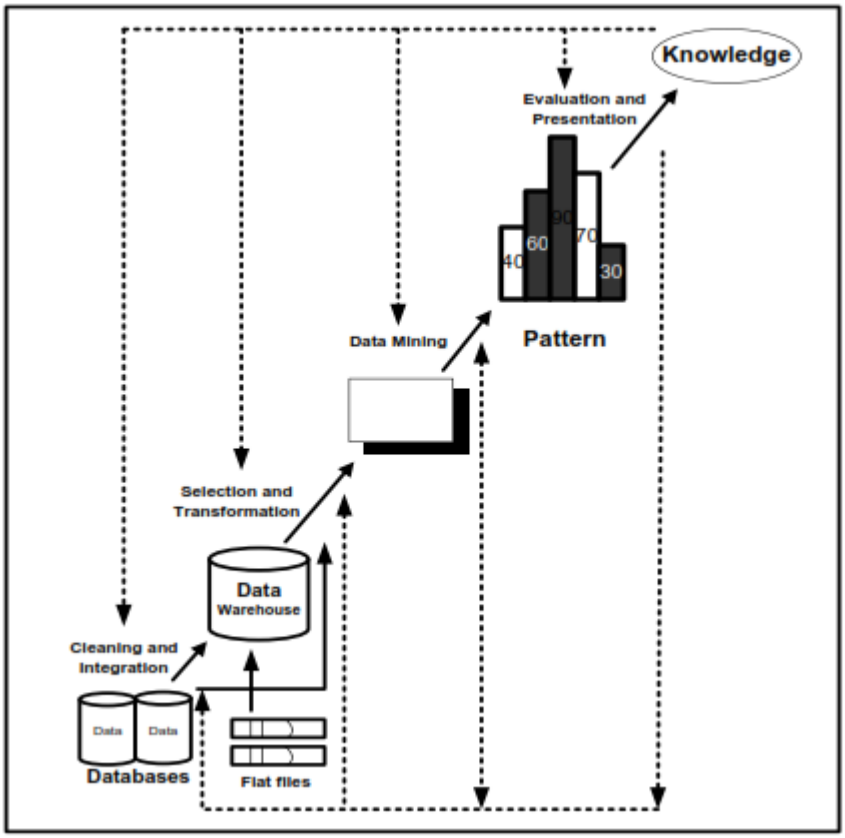
   Pengklusteran adalah pengelompokkan sebuah record, pengamatan dan membentuk kelas kedalam sebuah objek yang mempunyai kemiripan. algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. **Asosiasi**

   Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu.

2.2.2 Tahap-Tahap Data Mining

Karena data mining adalah sebuah untain proses, maka pecah menjadi beberapa tahap. Tahapan tersebut akan bersifat interaktif, pengguna akan terlibat langsung atau dengan perantara *KDD*[17].



Gambar 2.1 Tahapan *Data Mining*.

Tahapan *data mining* dibagi menjadi enam bagian yaitu :

1. Pembersihan data (*data cleaning*)

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

2. Integrasi data (*data integration*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalamsatu *database* baru. Tidak jarang data yang diperlukan untuk *data mining* tidakhanya berasal dari satu *database* tetapi juga berasal dari beberapa *database* ataufile teks. Integrasi data dilakukan pada atribut-

aribut yang mengidentifikasikanentitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggandan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan padaintegrasi data bisa menghasilkan hasil yang menyimpang dan bahkanmenyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi databerdasarkan jenis produk ternyata menggabungkan produk dari kategori yangberbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi Data (*Data Selection*)

   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena ituhanya data yang sesuai untuk dianalisis yang akan diambil dari *database*. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orangmembeli dalam kasus *market basket analysis,* tidak perlu mengambil namapelanggan, cukup dengan id pelanggan saja.

4. Transformasi data (*Data Transformation*)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam*data mining*. Beberapa metode *data mining* membutuhkan format data yangkhusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standarseperti analisis asosiasi dan *clustering* hanya bisa menerima input datakategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagibagimenjadi beberapa interval. Proses Ini seringdisebut transformasi data.

5. Proses *mining*.

   adalah sebuah proses yang paling utama pada saat metode diterapkan untuk mencari pengetahuan tersembunyi dan berharga dari data.

6. Evaluasi pola (*pattern evaluation*),

   Untuk mengidentifikasi pola-pola menarik kedalam *knowledge based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.

7. Presentasi pengetahuan (*knowledge presentation*),

   Merupakan penyajian dan visualisasi pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami *data mining*. Karenanya presentasi hasil *data mining* dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses *data mining*. Dalam presentasi ini,visualisasi juga bisa membantu mengkomunikasikan hasil *data mining* [19].

2.3 Klasifikasi

Klasifikasi adalah tugas pembelajaran sebuah fungsi target f yang memetakan setiap himpunan atribut x ke salah satu label kelas y yang telah didefinisikan sebelumnya. Klasifikasi adalah proses yang terdiri dari dua tahap, Langkah pertama, membangun model untuk mendeskripsikan predetermined set kelas data atau konsep. Model dibangun dengan menganalisis tuple database yang mendeskripsikan atribut. Tiap tuple diasumsikan memiliki predefined class, yang ditentukan oleh satu atribut, yang dinamakan atribut label kelas. Data tuple dianalisis untuk membangun model secara kolektif berdasarkan training data set. Beberapa tuple yang membentuk training set disebut training sample dan secara random dipilih dari populasi sample. Karena label kelas dari tiap training sample disediakan, langkah ini dikenal sebagai supervised learning. Model belajar direpresentasikan dalam classification rules, decision tree, atau formula matematika. Langkah kedua, pemakaian model untuk klasifikasi. Sebelum dipakai, dibuat estimasi keakuratan model dengan teknik holdout. Jika keakuratan model dapat diterima, model dapat digunakan untuk mengklasifikasikan data tuple baru, yang label kelasnya belum diketahui Atribut utama dari proses klasifikasi antara lain [10]:

1) Kelas, adalah atribut yang akan dijadikan target, sering juga disebut dengan label dari hasil klasifikasi. Sebagai contoh adalah kelas loyalitas pelanggan, kelas badai atau gempa bumi, dan lain-lain.
2) Prediktor, adalah variable yang bebas berdasarkan karakter atribut data yang akan diklasifikasikan. Misalnya merokok, minumminuman beralkohol, tekanan darah, status perkawinan, dansebagainya.
3) Set data pelatihan, adalah kumpulan data yang isinya predictor dan kelas untuk diolah agar model dapat kelompokan kedalam kelas yang tepat. Contohnya adalah grup mahasiswa fakultas ilmu komputer disebuah universitas, grup jurusan siswa di suatu sekolahan SMA.
4) Set data uji, merupakan data baru yang dikelompokan oleh modelUntuk mencari tau akurasi dari model yang sudah bentuk.

2.4 Algoritma C4.5

Algoritma C4.5 merupakan algortima yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metodeklasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusa mengubah fakta yang sangat besar menjadi pohon keputusanyang merepresentasikan aturan. Aturan dapat dengan mudah dipahamidengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentukbahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu [13].

Cara algoritma C4.5 untuk membangun pohon keputusan yaitu:

a. Pilih atribut yang akan digunakan sebagai akar
b. Buatlah sebuah cabang untuk setiap nilai
c. Bagilah kasus dalam sebuah cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 [14].

1. Menyiapkan data training. Data ini diambil dari data yang sudah pernah ada sebelumnya dan sudah dikelompokan kedalam kelas tertentu.
2. Setalah itu tentukan akar dari pohon. Pilih akar dari atribut, cara adalah dengan menghitung nilai gain dari semua atribut, yang menjadi akar pertama adalah nilai gain yang paling. Sebelum menentukan nilai gain, terlebih dahulu hitung nilai entropy. Untuk menentukan nilai entropy gunakan rumus

   $$Entropy(S) = \sum_{i=1}^{n} - \ pi.\log 2pi$$

   Keterangan :

   S = himpunan kasus

   n = jumlah partisi S

   pi = proporsi S_i terhadap S
3. Setelah itu tentukan nilai *gain* menggunakan rumus :

   $$gain(S.A) = Entropy(s) - \sum_{i}^{n} \frac{|si|}{|s|} * Entropy(Si)$$

   Keterangan :

   S = Himpunan kasus

   A = fitur

   N = jumlah partisi atribut A

   |Si| = proporsi Si terhadap S

   |S| = jumlah kasus dalam S
4. Setelah itu ulangilah langkah ke-2 sampai semua record terpartisi secara sempurna.
5. Proses partisi pohon keputusan akan berhenti saat :
   a) Semua *record* dalam simpul N mendapat kelas yang sama.
   b) Tidak ada atribut di dalam record yang dipartisi lagi.
   c) Tidak ada *record* di dalam cabang yang kosong.

2.5 Contoh Kasus Yang Diselesaikan Dengan Algoritma C4.5

Dalam kasus main tenis ini akan diperoleh sebuah pohon keputusan menentukan apakah akan bermain tenis atau tidak bermain tenis yang

dipengaruhi oleh keadaan cuaca, temperatur, kelembaban dan keadaan angin.

Tabel 2.2 Kondisi Lapangan

N0	OUTLOOK	TEMPERATURE	HUMADITY	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	Cloudy	Mild	High	TRUE	No

Untuk masalah di tabel 2.2 deselaiakan dengan menggunakan algoritma C4.5 dan akn di jelaskan semua langkah-langkahnya.

a. Cara menghitung jumlah kasus, jumlah kasus dengan keputusan Yes dan keputusan No, dan Entropy dari semua kasu yang ada dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE, HUMIDITY dan WINDY. Lalu dilakukan penghitungan Gain untuk semua atribut. Hasil perhitungan ditunjukkan oleh Tabel 2.3.

Tabel 2.3 Perhitungan Node 1

Node			**Jml kasus (S)**	**Tidak (S_1)**	**Ya (S_2)**	**Entropy**	**Gain**
1	TOTAL		14	4	10	0.863120569	
	OUTLOOK						0.258521037
		CLOUDY	4	0	4		
		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	

Pada baris *TOTAL* dalam kolom Entropy pada diatas ditentukan dengan rumus berikut:

$$Entropy(Total) = \left(-\frac{4}{14} * \log 2\ (\frac{4}{14})\right) + (\frac{10}{14}\ *\ \log 2\ (\frac{10}{14}))$$

$$Entropy(Total) = 0.863120569$$

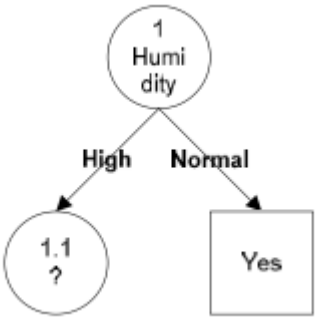
untuk menentukan nilai gain pada baris OUTLOOK dengan gunakan rumus berikut:

$$Gain\ (Total, outlook) = Entropy(Total) - \sum_{i-1}^{n} \frac{Outlook_i}{Total} *\ Entropy(Outlook)$$

$$Gain\ (Total, outlook) = 0.863120569 - ((\frac{4}{14} * 0) + (\frac{5}{14} *\ 0{,}723) +\ \ (\frac{5}{14} * 0.97))$$

$$Gain\ (Total, outlook) = 0.23$$

Dari hasil semua perhitungan didapat hasil Gain yang tertinggi dalam baris HUMIDITY yaitu sebesar 0.370506501. Dengan demikian HUMIDITY dapat menjadi node akar. Ada 2 nilai atribut dari HUMIDITY yaitu HIGH dan NORMAL. Karena atribut NORMAL mengklasifikasikan 1 yaitu keputusan-nya Yes, maka tidak perlu dilakukan perhitungan lagi, sedangkan untuk HIGH dilakukan perhitungan lagi. Untuk hasilnya akan tampak seperti gambar 2.2.



Gambar 2.2 Hasil Perhitungan Node 1

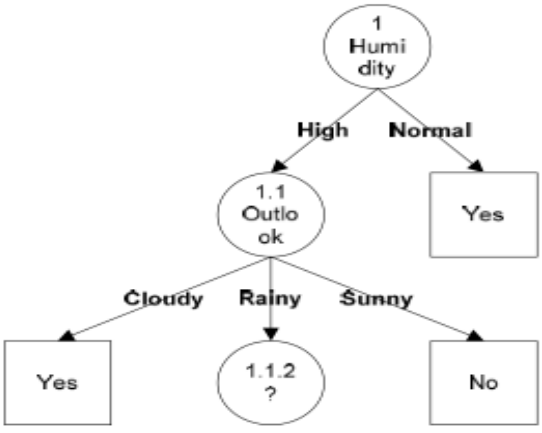
b. Untuk menentukan jumlah kasus, jumlah kasus yang ditentukan adalah jumlah keputusan *Yes* dan *No*, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut *OUTLOOK*, *TEMPERATURE* dan *WINDY* yang akan menjadi node akar atribut HIGH. Lalu lakukan penghitungan Gain untuk masing-masing atribut. Hasil perhitungan akan dipaparkan dalam Tabel 2.4.

Tabel 2.4 Perhitungan Node 1.1

Node			**Jumlah kasus (S)**	**Tidak (S_1)**	**Ya (S_2)**	**Entropy**	**Gain**
1.1	HUMIDITYHIGH		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 semua perhitungan didapat hasil Gain yang tertinggi dalam baris OUTLOOK yaitu sebesar 0.69951385. Dari hasil tersebut maka OUTLOOK menjadi node cabang dari atribut HIGH. Ada 3 nilai atribut dari OUTLOOK yaitu CLOUDY, RAINY dan SUNNY. karena atribut CLOUDY mengklasifikasikan kasus menjadi 1 yaitu keputusan Yes dan nilai atribut SUNNY sudah mengklasifikasikan kasus menjadi satu keputusan No, oleh karena itu tidak perlu lagi dilakukan perhitungan, seddangkan atribut RAINY dilakukan perhitungan lagi. Untuk hasilnya akan tampak seperti gambar 2.3.

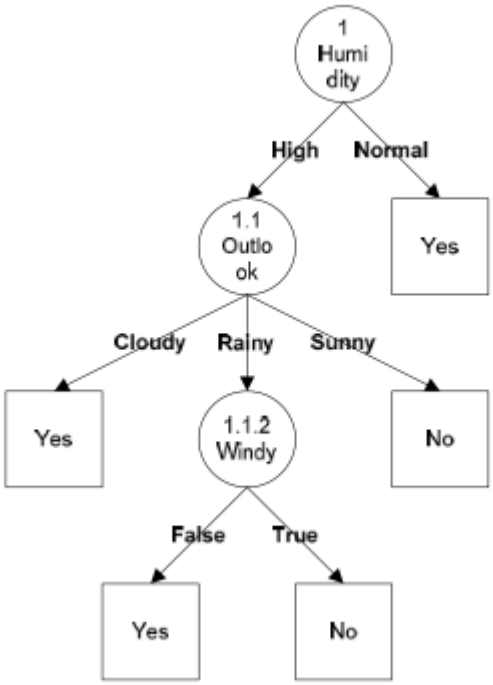


Gambar 2.3 Hasil Perhitungan Node 1.1

c. Untuk menentukan jumlah kasus, jumlah kasus yang ditentukan adalah jumlah keputusan *Yes* dan *No,* dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut *TEMPERATURE* dan *WINDY* yang dapat menjadi node cabang dari atribut *RAINY*. Lalu lakukan penghitungan Gain untuk setiap atribut. Untuk hasilnya akan tampak seperti tabel 2.5. Dari hasil pada tabel 2.4 dapat peroleh atribut dengan Gain tertinggi adalah *WINDY* yaitu sebesar 1. Dari hasil tersebut maka *WINDY* menjadi node cabang dari atribut *RAINY*. Dari atribut *WINDY* terdapat dua nilai yaitu *TRUE* dan *FALSE*. karena atribut *FALSE* sudah mengklasifikasikan kasus menjadi 1 yaitu keputusan-nya *Yes* dan nilai atribut *TRUE* sudah mengklasifikasikan kasus menjadi 1 dengan keputusan *No,* oleh karena itu tidak perlu lagi dilakukan perhitungan atribut ini.

Tabel 2.5 Perhitungan Node 1.1.2

Node			**Jml kasus (S)**	**Tidak (S_1)**	**Ya (S_2)**	**Entropy**	**Gain**
1.1.2	HUMIDITYHIGH 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	



Gambar 2.4 Hasil Perhitungan Node 1.1.2

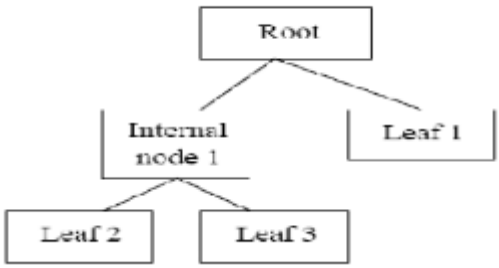
Pohon keputusan yang terbentuk sampai tahap ini ditunjukkan pada Gambar 2.4. Dengan memperhatikan pohon keputusan pada Gambar 2.4, diketahui bahwa semua kasus sudah masuk dalam kelas. Dengan demikian, pohon keputusan pada Gambar 4 merupakan pohon keputusan terakhir yang terbentuk

2.6 Decision Tree

Pohon *(tree)* adalah sebuah struktur data yang terdiri dari simpul *(node)* dan rusuk *(edge)*. Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar *(root node)*, simpul percabangan/ internal *(branch/ internal node)* dan simpul daun *(leaf node)*, [15].Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi label nilai atribut yang mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda [15].

Gambar 2.5 Konsep Pohon Keputusan.

Pohon keputusan adalah sebuah stuktur flowchart yang setiap node nya merepresentasikan test dalam atribut (contoh, koin bila kita bolak balikan akan menghasilkan kepala, atau ekor), Setiap cabang (branch) mewakili hasil test dan setiap daun node (leaf) mewakili kelas label (hasil keputusan setelah menghitung semua atribut). Bagian dari akar (root) hingga ke daun merepresentasikan dari rules (aturan) yang terbentuk. Pohon keputusan menurut saya bukanlah sebuah algoritma melainkan metode yang nantinya menghasilkan beberapa algoritma yang dapat digunakan dalam pengembangannya. Metode ini akan membantu kita untuk mengeksplorasi data, dan menemukan relasi tersembunyi antar sejumlah variabel input dan target.Pohon keputusan adalah himpunan aturan *IF...THEN*. Setiap *path* dalam *tree* dihubungkan dengan sebuah aturan, di mana premis terdiri atas sekumpulan *node-node* yang ditemui, dan kesimpulan dari aturam terdiri atas kelas yang terhubung dengan *leaf* dari *path* [16].



Gambar 2.6 Konsep Dasar Pohon Keputusan

Paling atas dari pohon keputusan diesbut dengan titik akar *(root)*, sedangkan setiap cabang dari pohon keputusan adalah pembagian berdasarkan hasil uji, dan titik akhir *(leaf)* merupakan pembagian kelas yang dihasilkan.

2.7 Confusion Matrix

Confusion matrix adalah suatu metode yang biasanya digunakan sebagai perhitungan akurasi pada konsep data mining. Informasi dalam confusion matrix diperlukan untuk menentukan kinerja model klasifikasi. Ringkasan informasi ini ke dalam sebuah nilai digunakan untuk membandingkan kinerja dari model-model yang berbeda. Hal ini dapat dilakukan dengan menggunakan performace metric [17]

Tabel 2.6 Confusion Matrik

Classsification	*Predicted class*	
	Class = Yes	Class = No
Class = Yes	a (*true positive* – TP)	b (*false negative* – TN)
Class = No	c (*false positive* – FP)	d (*true negative*– TN)

2.8 Kerangka Pemikiran

Tabel 2.7 Kerangka Pemikiran