DECISION TREE (POHON KEPUTUSAN)

Latar Belakang Pohon Keputusan

Di dalam kehidupan manusia sehari-hari, manusia selalu dihadapkan oleh berbagai macam masalah dari berbagai macam bidang. Masalah-masalah yang dihadapi oleh manusia memiliki tingkat kesulitan dan kompleksitas yang sangat bervariasi, mulai dari masalah yang teramat sederhana dengan sedikit faktor-faktor yang terkait, sampai dengan masalah yang sangat rumit dengan banyak sekali faktor-faktor yang terkait dan perlu untuk diperhitungkan. Untuk menghadapi masalah-masalah ini, manusia mulai mengembangkan sebuah sistem yang dapat membantu manusia agar dapat dengan mudah mampu untuk menyelesaikan masalah-masalah tersebut. Adapun pohon keputusan ini adalah sebuah jawaban akan sebuah sistem yang manusia kembangkan untuk membantu mencari dan membuat keputusan untuk masalah-masalah tersebut dan dengan memperhitungkan berbagai macam faktor yang ada di dalam lingkup masalah tersebut. Dengan pohon keputusan, manusia dapat dengan mudah mengidentifikasi dan melihat hubungan antara faktor-faktor yang mempengaruhi suatu masalah dan dapat mencari penyelesaian terbaik dengan memperhitungkan faktor-faktor tersebut. Pohon keputusan ini juga dapat menganalisa nilai resiko dan nilai suatu informasi yang terdapat dalam suatu alternatif pemecahan masalah. Peranan pohon keputusan sebagai alat bantu dalam mengambil keputusan (*decision support tool*) telah dikembangkan oleh manusia sejak perkembangan teori pohon yang dilandaskan pada teori graf. Kegunaan pohon keputusan yang sangat banyak ini membuatnya telah dimanfaatkan oleh manusia dalam berbagai macam sistem pengambilan keputusan.

Pengertian Pohon Keputusan

Pohon dalam analisis pemecahan masalah pengambilan keputusan adalah pemetaan mengenai alternatif-alternatif pemecahan masalah yang dapat diambil dari masalah tersebut. Pohon tersebut juga memperlihatkan faktor-faktor kemungkinan/probablitas yang akan mempengaruhi alternatif-alternatif keputusan tersebut, disertai dengan estimasi hasil akhir yang akan didapat bila kita mengambil alternatif keputusan tersebut.

Manfaat Pohon Keputusan

Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-*break down* proses pengambilan keputusan yang kompleks menjadi lebih simpel sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan. Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain. Sering terjadi tawar menawar antara keakuratan model dengan transparansi model. Dalam beberapa aplikasi, akurasi dari sebuah klasifikasi atau prediksi adalah satu-satunya hal yang ditonjolkan, misalnya sebuah perusahaan *direct mail* membuat sebuah model yang akurat untuk memprediksi anggota mana yang berpotensi untuk merespon permintaan, tanpa memperhatikan bagaimana atau mengapa model tersebut bekerja.

Kelebihan Pohon Keputusan

Kelebihan dari metode pohon keputusan adalah:

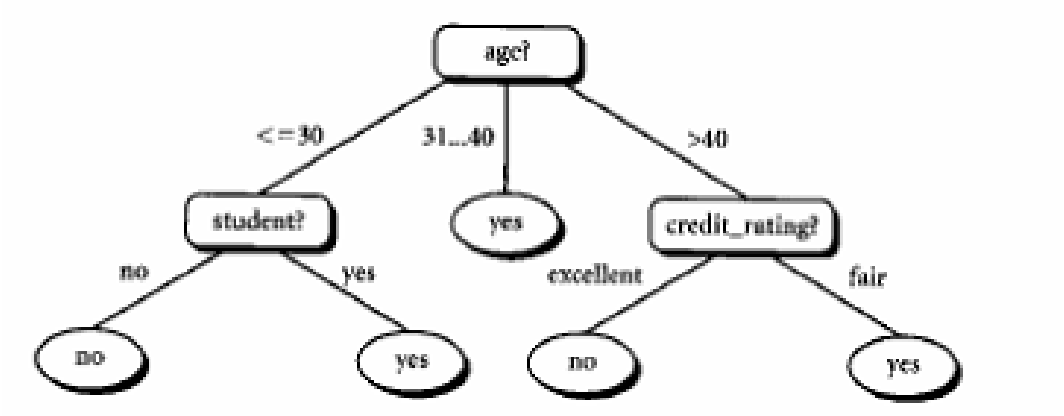
- Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.
- Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sample diuji hanya berdasarkan kriteria atau kelas tertentu.
- Fleksibel untuk memilih fitur dari internal node yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama. Kefleksibelan metode pohon keputusan ini meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional.
- Dalam analisis multivariat, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode pohon keputusan dapat menghindari munculnya permasalahan ini dengan menggunakan criteria yang jumlahnya lebih sedikit pada setiap node internal tanpa banyak mengurangi kualitas keputusan yang dihasilkan.

Kekurangan Pohon Keputusan

- Terjadi overlap terutama ketika kelas-kelas dan criteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah memori yang diperlukan.
- Pengakumulasian jumlah eror dari setiap tingkat dalam sebuah pohon keputusan yang besar.
- Kesulitan dalam mendesain pohon keputusan yang optimal.
- Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana pohon tersebut didesain.

Model Pohon Keputusan

Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Contoh dari pohon keputusan dapat dilihat di Gambar 1 berikut ini.



Gambar 1. Model Pohon Keputusan (Pramudiono,2008)

Disini setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Contoh di Gambar 1 adalah identifikasi pembeli komputer,dari pohon keputusan tersebut diketahui bahwa salah satu kelompok yang potensial membeli komputer adalah

orang yang berusia di bawah 30 tahun dan juga pelajar. Setelah sebuah pohon keputusan dibangun maka dapat digunakan untuk mengklasifikasikan *record* yang belum ada kelasnya. Dimulai dari *node root*, menggunakan tes terhadap atribut dari *record* yang belum ada kelasnya tersebut lalu mengikuti cabang yang sesuai dengan hasil dari tes tersebut, yang akan membawa kepada *internal node* (*node* yang memiliki satu cabang masuk dan dua atau lebih cabang yang keluar), dengan cara harus melakukan tes lagi terhadap atribut atau *node* daun. *Record* yang kelasnya tidak diketahui kemudian diberikan kelas yang sesuai dengan kelas yang ada pada *node* daun. Pada pohon keputusan setiap simpul daun menandai label kelas. Proses dalam pohon keputusan yaitu mengubah bentuk data (tabel) menjadi model pohon (*tree*) kemudian mengubah model pohon tersebut menjadi aturan (*rule*).

Algoritma C4.5

Untuk memudahkan penjelasan mengenai algoritma C4.5 berikut ini disertakan contoh kasus yang dituangkan dalam Tabel 1:

Tabel 1. 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

Dalam kasus yang tertera pada Tabel 1, akan dibuat pohon keputusan untuk menentukan main tenis atau tidak dengan melihat keadaan cuaca (outlook), temperatur, kelembaban (humidity) dan keadaan angin (windy).

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk masing-masing nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk masing-masing 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 seperti tertera dalam Rumus 1.

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

Dengan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

Sedangkan perhitungan nilai entropy dapat dilihat pada rumus 2 berikut:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \qquad (2)$$

dengan :
S : Himpunan Kasus
A : Fitur
n : Jumlah partisi S
p_i : Proporsi dari S_i terhadap S

Berikut ini adalah penjelasan lebih rinci mengenai masing-masing langkah dalam pembentukan pohon keputusan dengan menggunakan algoritma C4.5 untuk menyelesaikan permasalahan.

a. 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, HUMIDITY dan WINDY. Setelah itu lakukan penghitungan Gain untuk masing-masing atribut. Hasil perhitungan ditunjukkan oleh Tabel 2.

Tabel 2. 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 pada Tabel 2 dihitung dengan rumus 2, sebagai berikut:

$$Entropy(Total) = (-\frac{4}{14} * \log_2(\frac{4}{14})) + (-\frac{10}{14} * \log_2(\frac{10}{14}))$$
$$Entropy(Total) = 0.863120569$$

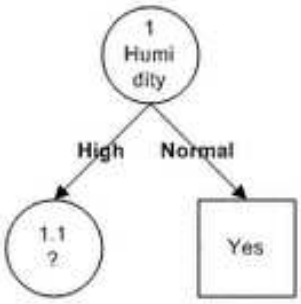
Sedangkan nilai Gain pada baris OUTLOOK dihitung dengan menggunakan rumus 1, sebagai berikut:

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

Sehingga didapat Gain(Total, Outlook) = 0.258521037

Dari hasil pada Tabel 2 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 dan NORMAL. Dari kedua nilai atribut tersebut, nilai atribut NORMAL sudah mengklasifikasikan kasus menjadi 1 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.



Gambar 2. Pohon Keputusan Hasil Perhitungan Node 1

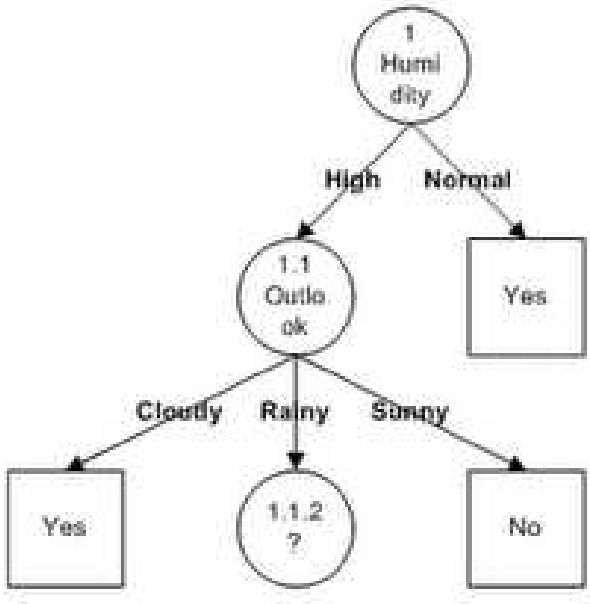
b. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut OUTLOOK, TEMPERATURE dan WINDY yang dapat menjadi node akar dari nilai atribut HIGH. Setelah itu lakukan penghitungan Gain untuk masing-masing atribut. Hasil perhitungan ditunjukkan oleh Tabel 3.

Tabel 3. 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 3 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 ditunjukkan pada gambar 3.



Gambar 3. Pohon Keputusan Hasil Perhitungan Node 1.1

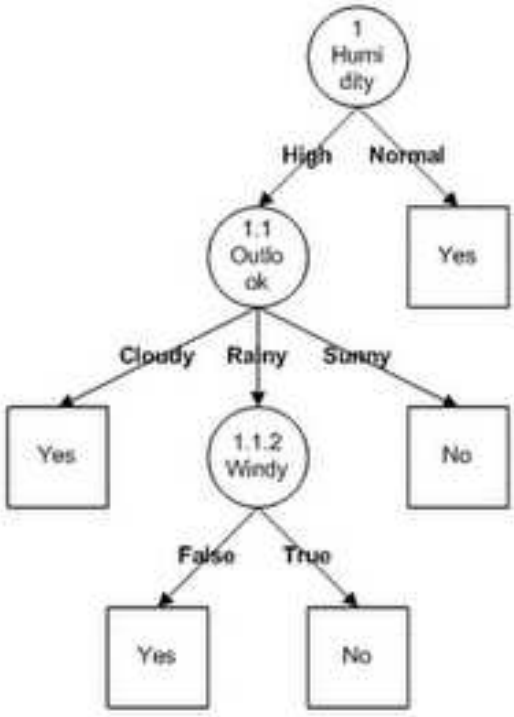
c. Menghitung jumlah kasus, jumlah kasus untuk keputusan Yes, jumlah kasus untuk keputusan No, dan Entropy dari semua kasus dan kasus yang dibagi berdasarkan atribut TEMPERATURE dan WINDY yang dapat menjadi node cabang dari nilai atribut RAINY. Setelah itu lakukan penghitungan Gain untuk masing-masing atribut. Hasil perhitungan ditunjukkan oleh Tabel 4.

Tabel 4. 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 4 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.

Pohon keputusan yang terbentuk sampai tahap ini ditunjukkan pada Gambar 4.



Gambar 4. Pohon Keputusan Hasil Perhitungan Node 1.1.2

Dengan memperhatikan pohon keputusan pada Gambar 4, diketahui bahwa semua kasus sudah masuk dalam kelas. Dengan demikian, pohon keputusan pada Gambar 4 merupakan pohon keputusan terakhir yang terbentuk.

Referensi:

- Kusrini dan Emha Taufiq Luthfi. 2009. *Algoritma Data Mining*. Penerbit Andi Offset, Yogyakarta.
- Larose, Daniel T. 2005. Discovering Knowledge ini Data: An Introduction to Data Mining. Wiley.
- Pramudiono, Iko. *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data*. http://www.ilmukomputer.com
- Santosa, Budi. 2007. Data Mining : *Teknik Pemanfaatan Data untuk keperluan Bisnis*. Graha Ilmu. Yogyakarta.
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2004. *Introduction to Data Mining*.