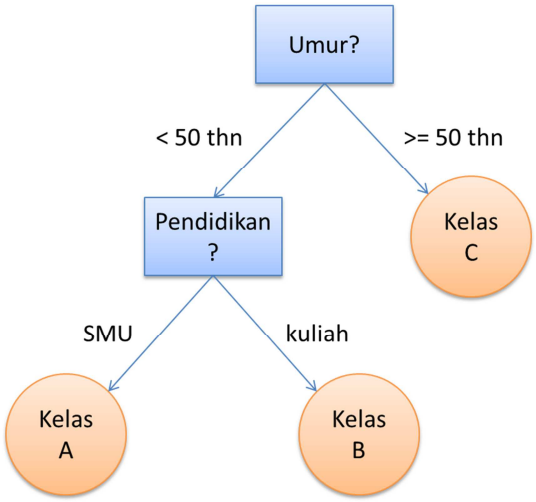
Belajar Mudah Algoritma Data Mining : C4.5

Algoritma *data mining* C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Dasar algoritma C4.5 adalah pembentukan pohon keputusan (*decision tree*). Cabang-cabang pohon keputusan merupakan pertanyaan klasifikasi dan daun-daunnya merupakan kelas-kelas atau segmen-segmennya.



Gambar 1. Contoh Pohon Keputusan

Algoritma C4.5 merupakan salah satu algoritma *machine learning*. Dengan algoritma ini, mesin (komputer) akan diberikan sekelompok data untuk dipelajari yang disebut *learning dataset*. Kemudian hasil dari pembelajaran selanjutnya akan digunakan untuk mengolah data-data yang baru yang disebut *test dataset*. Karena algoritma C4.5 digunakan untuk melakukan klasifikasi, jadi hasil dari pengolahan *test dataset* berupa pengelompokkan data ke dalam kelas-kelasnya.

Berikut ini adalah uraian langkah-langkah dalam algoritma C4.5 untuk menyelesaikan kasus suatu pertandingan tenis akan dilakukan atau tidak, berdasarkan keadaan cuaca, suhu, kelembaban, dan angin. Data yang telah ada pada Tabel 1, akan digunakan untuk membentuk pohon keputusan.

Pada Tabel 1, atribut-atributnya adalah Cuaca, Suhu, Kelembaban, dan Berangin. Setiap atribut memiliki nilai. Sedangkan kelasnya ada pada kolom Main yaitu kelas “Tidak” dan kelas “Ya”. Kemudian data tersebut dianalisis; dataset tersebut memiliki 14 kasus yang terdiri 10 “Ya” dan 4 “Tidak” pada kolom Main (lihat Tabel 2).

Tabel 1. Learning Dataset

No	Cuaca	Suhu	Kelembaban	Berangin	Main
1	Cerah	Panas	Tinggi	Salah	Tidak
2	Cerah	Panas	Tinggi	Benar	Tidak
3	Berawan	Panas	Tinggi	Salah	Ya
4	Hujan	Sejuk	Tinggi	Salah	Ya
5	Hujan	Dingin	Normal	Salah	Ya
6	Hujan	Dingin	Normal	Benar	Ya
7	Berawan	Dingin	Normal	Benar	Ya
8	Cerah	Sejuk	Tinggi	Salah	Tidak
9	Cerah	Dingin	Normal	Salah	Ya
10	Hujan	Sejuk	Normal	Salah	Ya
11	Cerah	Sejuk	Normal	Benar	Ya
12	Berawan	Sejuk	Tinggi	Benar	Ya
13	Berawan	Panas	Normal	Salah	Ya
14	Hujan	Sejuk	Tinggi	Benar	Tidak

Kemudian hitung entropi dengan rumus sebagai berikut :

$$Entropi\ (S) = \sum_{j=1}^{k} - p_j\ \log_2\ p_j$$

Keterangan :

- S adalah himpunan (dataset) kasus
- k adalah banyaknya partisi S
- p_j adalah probabilitas yang di dapat dari Sum(Ya) dibagi Total Kasus.

Jadi $Entropi\ (S) = (-\left(\frac{10}{14}\right) \times\ \log_2\ \left(\frac{10}{14}\right)) + (-\left(\frac{4}{14}\right) \times\ \log_2\ \left(\frac{4}{14}\right)) = 0.863120569$

Tabel 2. Hasil Perhitungan pada Dataset

Total Kasus	Sum(Ya)	Sum(Tidak)	Entropi Total
14	10	4	0.863120569

Setelah mendapatkan entropi dari keseluruhan kasus, lakukan analisis pada setiap atribut dan nilai-nilainya dan hitung entropinya seperti yang ditampilkan pada Tabel 3.

Tabel 3. Analisis Atribut, Nilai, Banyaknya Kejadian Nilai, Entropi dan Gain

Node	Atribut	Nilai	Sum(Nilai)	Sum(Ya)	Sum(Tidak)	Entropi	Gain
1	Cuaca	Berawan	4	4	0	0	
		Hujan	5	4	1	0.721928095	
		Cerah	5	2	3	0.970950594	
							0.258521037
	Suhu	Dingin	4	4	0	0	
		Panas	4	2	2	1	
		Sejuk	6	4	2	0.918295834	
							0.183850925
	Kelembaban	Tinggi	7	3	4	0.985228136	
		Normal	7	7	0	0	
							0.370506501
	Berangin	Salah	8	6	2	0.811278124	
		Benar	6	2	4	0.918295834	
							0.005977711

Untuk menghitung gain setiap atribut rumusnya adalah :

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropi(S_i)$$

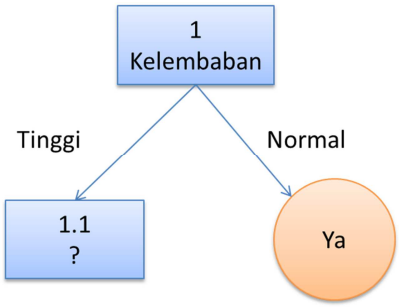
Jadi :

$$Gain\ (Cuaca) = 0.863120569 - \left(\left(\frac{4}{14}\right) \times 0 + \left(\frac{5}{14}\right) \times 0.721928095 + \left(\frac{5}{14}\right) \times 0.970950594\right)$$

$$Gain\ (Cuaca) = 0.258521037$$

Hitung pula Gain (Suhu), Gain (Kelembaban), dan Gain (Berangin). Hasilnya dapat dilihat pada Tabel 3. Karena nilai gain terbesar adalah Gain (Kelembaban). Maka Kelembaban menjadi node akar (*root node*).

Kemudian pada kelembaban normal, memiliki 7 kasus dan semuanya memiliki jawaban Ya (Sum(Total) / Sum(Ya) = 7/7 = 1). Dengan demikian kelembaban normal menjadi daun atau *leaf*. Lihat Tabel 3 yang selnya berwarna hijau.



Gambar 2. Pohon Keputusan Node 1 (*root node*)

Berdasarkan pembentukan pohon keputusan node 1 (*root node*), Node 1.1 akan dianalisis lebih lanjut. Untuk mempermudah, Tabel 1 difilter, dengan mengambil data yang memiliki Kelembaban = Tinggi sehingga jadilah Tabel 4.

Tabel 4. Data yang Memiliki Kelembaban = Tinggi

No	Cuaca	Suhu	Kelembaban	Berangin	Main
1	Cerah	Panas	Tinggi	Salah	Tidak
2	Cerah	Panas	Tinggi	Benar	Tidak
3	Berawan	Panas	Tinggi	Salah	Ya
4	Hujan	Sejuk	Tinggi	Salah	Ya
5	Cerah	Sejuk	Tinggi	Salah	Tidak
6	Berawan	Sejuk	Tinggi	Benar	Ya
7	Hujan	Sejuk	Tinggi	Benar	Tidak

Kemudian data di Tabel 4 dianalisis dan dihitung lagi entropi atribut Kelebaban Tinggi dan entropi setiap atribut serta gainnya sehingga hasilnya seperti data pada Tabel 5. Setelah itu tentukan pilih atribut yang memiliki gain tertinggi untuk dibuatkan node berikutnya.

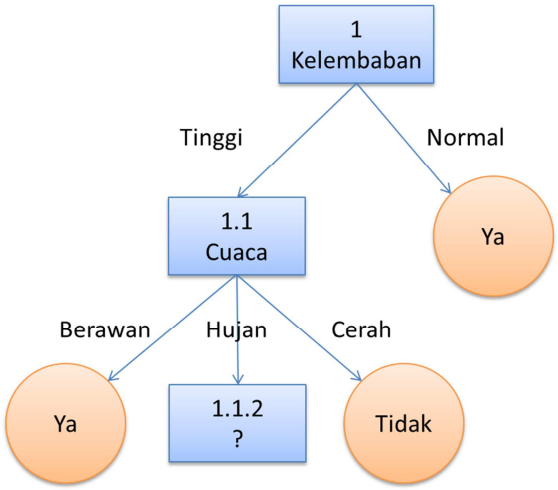
Tabel 5. Hasil Analisis Node 1.1

Kelembaban Tinggi	Sum(Ya)	Sum(Tidak)	Entropi
7	3	4	0.985228136

Node	Atribut	Nilai	Sum(Variabel)	Sum(Ya)	Sum(Tidak)	Entropi	Gain
1.1	Cuaca	Berawan	2	2	0	0	
		Hujan	2	1	1	1	
		Cerah	3	0	3	0	
							0.69951385
	Suhu	Dingin	0	0	0	0	
		Panas	3	1	2	0.918295834	
		Sejuk	4	2	2	1	
							0.020244207
	Berangin	Salah	4	2	2	1	
		Benar	3	2	1	0.918295834	
							0.020244207

Dari Tabel 5, gain tertinggi ada pada atribut Cuaca, dan Nilai yang dijadikan daun atau leaf adalah Berawan dan Cerah. Jika divualisasi maka pohon keputusan tampak seperti Gambar 3.

Untuk menganalisis node 1.1.2, lakukan lagi langkah-langkah yang sama seperti sebelumnya. Hasilnya ditampilkan pada Tabel 6 dan Gambar 4.



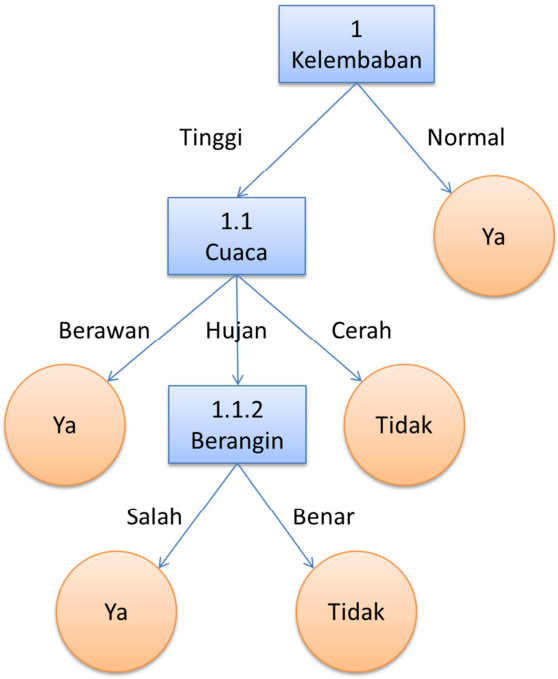
Gambar 3. Pohon Keputusan Analisis Node 1.1

Tabel 6. Hasil Analisis Node 1.1.2.

No	Cuaca	Suhu	Kelembaban	Berangin	Main
1	Hujan	Sejuk	Tinggi	Salah	Ya
2	Hujan	Sejuk	Tinggi	Benar	Tidak

Kelembaban Tinggi & Hujan	Sum(Ya)	Sum(Tidak)	Entropi
2	1	1	1

Node	Atribut	Nilai	Sum(Nilai)	Sum(Ya)	Sum(Tidak)	Entropi	Gain
1.1.2	Suhu	Dingin	0	0	0	0	
		Panas	0	0	0	0	
		Sejuk	2	1	1	1	
							0
	Berangin	Salah	1	1	0	0	
		Benar	1	0	1	0	
							1



Gambar 4. Pohon Keputusan Akhir