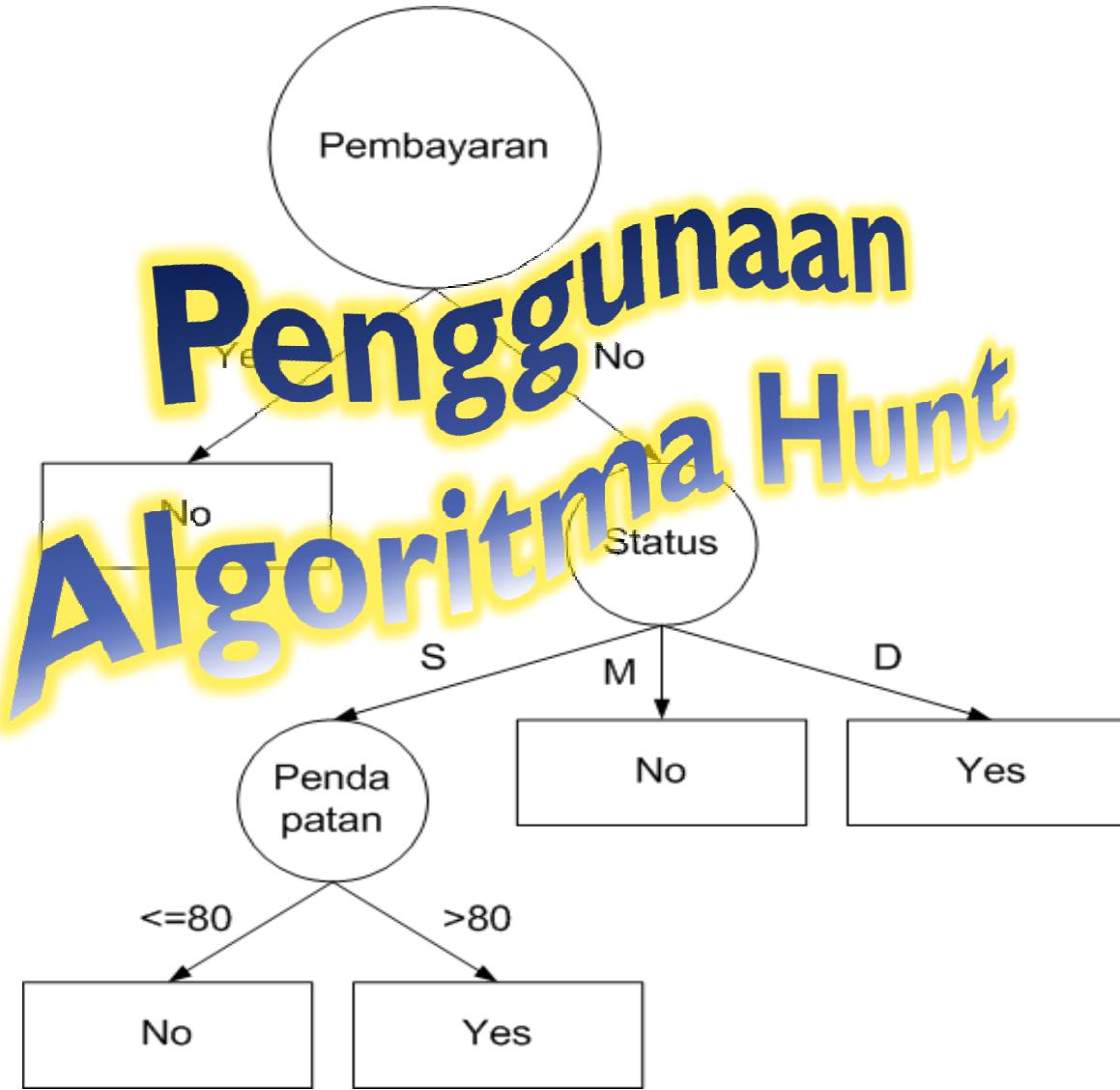


DATA MINING METODE KLASIFIKASI

DECISION TREE: PEMILIHAN ATRIBUT BERDASARKAN INFORMATION GAIN



Decision Tree



Sesederhana itukah ?



TIDAK !!!!

Bagaimana Memilih Atribut ?

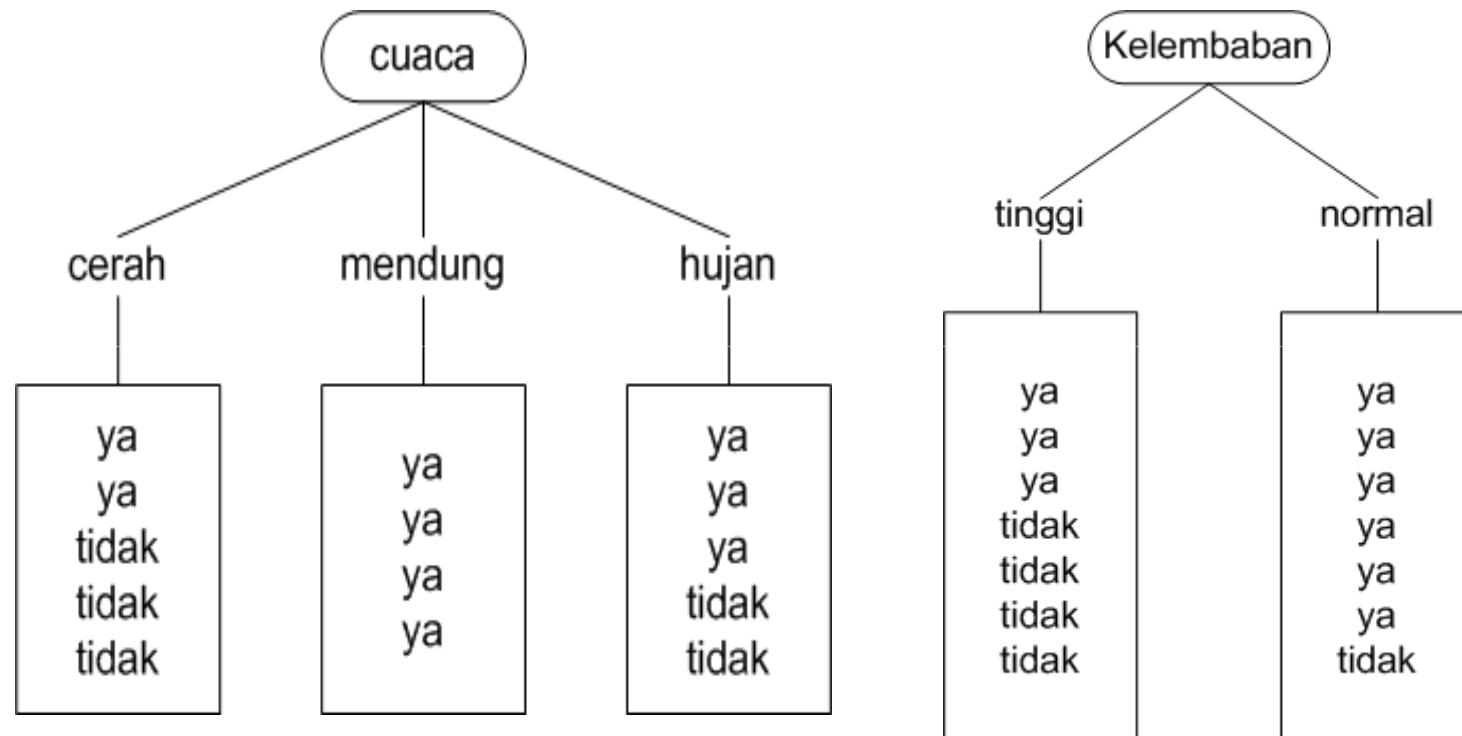


DATA TRAINING

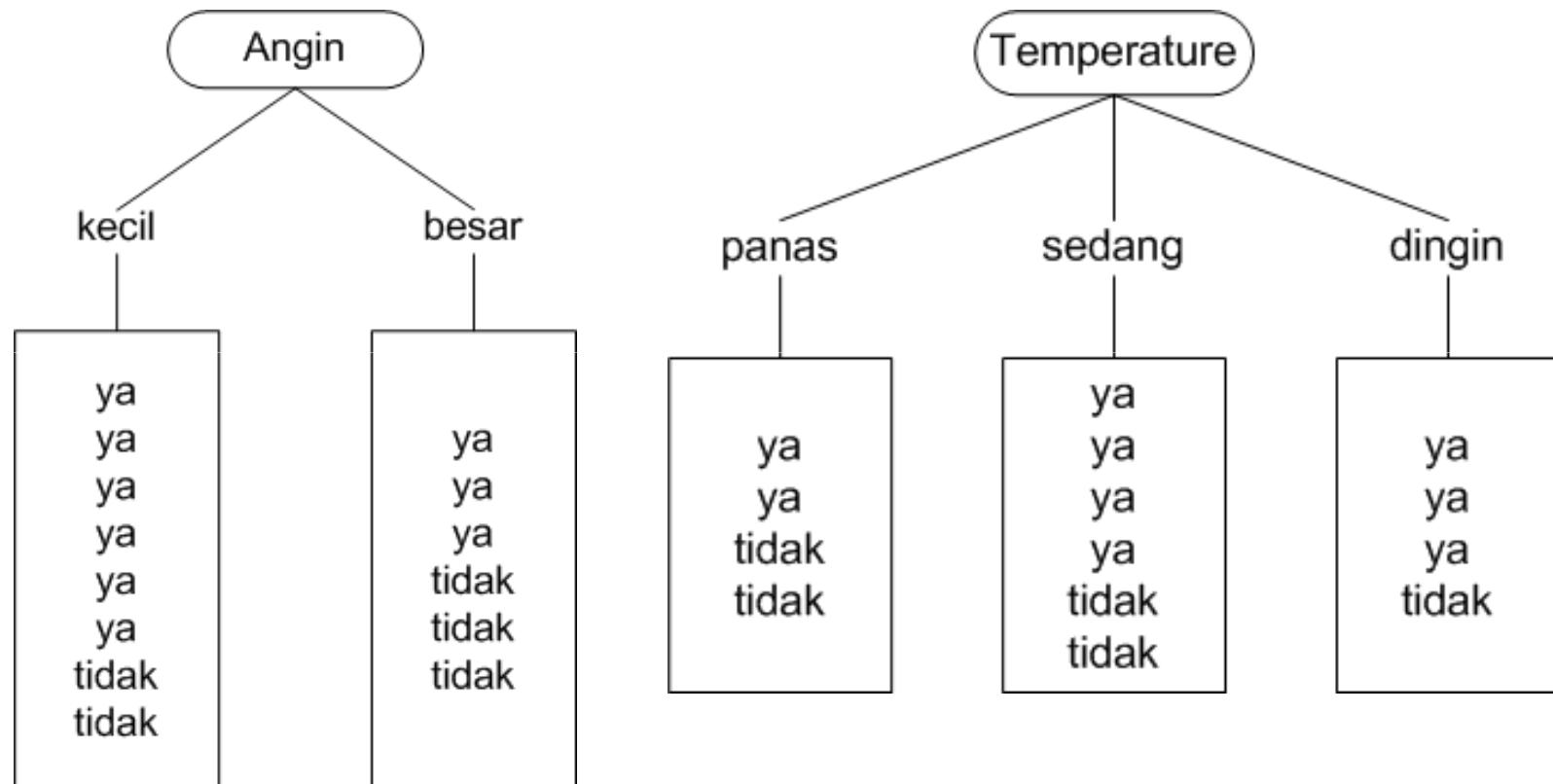
Tabel I. Data Cuaca dan Keputusan

No	Cuaca X1	Temperature X2	Kelembaban X3	Angin X4	Main / Tidak Y
1	Cerah	Panas	Tinggi	Kecil	Tidak
2	Cerah	Panas	Tinggi	Besar	Tidak
3	Mendung	Panas	Tinggi	Kecil	Ya
4	Hujan	Sedang	Tinggi	Kecil	Ya
5	Hujan	Dingin	Normal	Kecil	Ya
6	Hujan	Dingin	Normal	Besar	Tidak
7	Mendung	Dingin	Normal	Besar	Ya
8	Cerah	Sedang	Tinggi	Kecil	Tidak
9	Cerah	Dingin	Normal	Kecil	Ya
10	Hujan	Sedang	Normal	Kecil	Ya
11	Cerah	Sedang	Normal	Besar	Ya
12	Mendung	Sedang	Tinggi	Besar	Ya
13	Mendung	Panas	Normal	Kecil	Ya
14	Hujan	Sedang	Tinggi	Besar	Tidak

SELEKSI ATRIBUT INDEPENDENT



SELEKSI ATRIBUT INDEPENDENT





ENTROPI

- Sebelum penghitungan information gain, kita perlu menghitung dulu nilai informasi dalam satuan bits dari suatu kumpulan objek.
- Cara menghitung dilakukan dengan menggunakan konsep entropi.
- Entropi menyatakan *impurity* suatu kumpulan objek.



ENTROPI

- Jika diberikan sekumpulan objek dengan label / output y yang terdiri dari objek berlabel 1,2 sampai n , entropi dari obyek dengan n kelas dihitung dengan rumus:

$$Entropi(y) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \cdots - p_n \log_2 p_n$$

- Dimana p_1, p_2, \dots, p_n masing-masing menyatakan proporsi kelas 1, kelas 2, ..., kelas n dalam output.



ENTROPI

- Berdasarkan data pada Tabel 1, pada variabel y ada 9 keputusan main dan 5 keputusan tidak main, sehingga diperoleh:

$$Entropi[9,5] = -\frac{9}{14} \log_2 \left(\frac{9}{14} \right) - \frac{5}{14} \log_2 \left(\frac{5}{14} \right) = 0.940$$

- Jika perbandingan dua kelas, rasionalya sama maka nilai entropinya 1.
- Jika satu set hanya terdiri dari satu kelas maka entropinya 0.



ENTROPI

- Hasil Perhitungan
 - Cuaca = cerah
 $\text{entropi}[2,3] = 0.971 \text{ bits}$
 - Cuaca = mendung
 $\text{entropi}[4,0] = 0 \text{ bits}$
 - Cuaca = hujan
 $\text{entropi}[3,2] = 0.971 \text{ bits}$

$$\text{entropi}([2,3], [4,0], [3,2]) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694 \text{ bits}$$

Entropi Total Cuaca = 0.694 bits



A. INFORMATION GAIN

- Information gain diperoleh dari output data atau variabel dependent y yang dikelompokkan berdasarkan atribut A , dinotasikan dengan $gain(y, A)$.
- Information gain, $gain(y, A)$, dari atribut A relatif terhadap output data y adalah:

$$gain(y, A) = entropi(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entropi(y_c)$$



A. INFORMATION GAIN

- Berdasarkan Tabel I, kita pilih Angin sebagai atribut A yang terdiri dari 9 keputusan ya (main) dan 5 keputusan tidak main.

$$nilai(A) = [kecil, besar]$$

$$y = [9, 5]$$

$$y_{kecil} \leftarrow [6, 2]$$

$$y_{besar} \leftarrow [3, 3]$$

A. INFORMATION GAIN

$$gain(y, A) = entropi(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entropi(y_c)$$

$$gain(y, A) = entropi(y) - \frac{8}{14} entropi(y_{kecil}) - \frac{6}{14} entropi(y_{besar})$$

$$gain(y, A) = 0.940 - \left(\frac{8}{14}\right) 0.811 - \left(\frac{6}{14}\right) 1.0 = 0.048$$



A. INFORMATION GAIN

Dari semua atribut, jika dihitung *information gain*-nya adalah:

- Cuaca = 0.247 bits
- Temperatur = 0.029 bits
- Kelembaban = 0.152 bits
- Angin = 0.048 bits



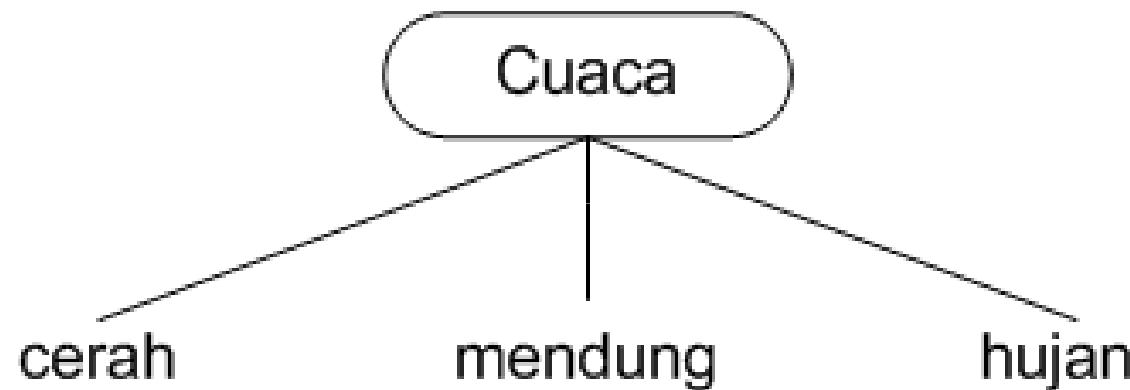
A. INFORMATION GAIN

Dari *information gain* semua atribut dapat dilihat bahwa nilai $gain(y, A)$ yang terbesar adalah :

- Cuaca = 0.247 bits
- Sehingga atribut cuaca dapat dipilih sebagai atribut pemecahan pertama dalam *Decision Tree*

A. INFORMATION GAIN

- Pemecahan pertama pada *Decision Tree*





A. INFORMATION GAIN

- Iterasi dilakukan kembali pada setiap cabang cuaca.
- Pada cabang cuaca cerah, kita hitung $gain(y,A)$ setiap atribut x_2 , x_3 , dan x_4 .
- Diperoleh information gain, $gain(y,A)$:
 - Gain(temperatur) = 0.571 bits
 - Gain(kelembaban) = 0.971 bits
 - Gain(angin) = 0.02 bits



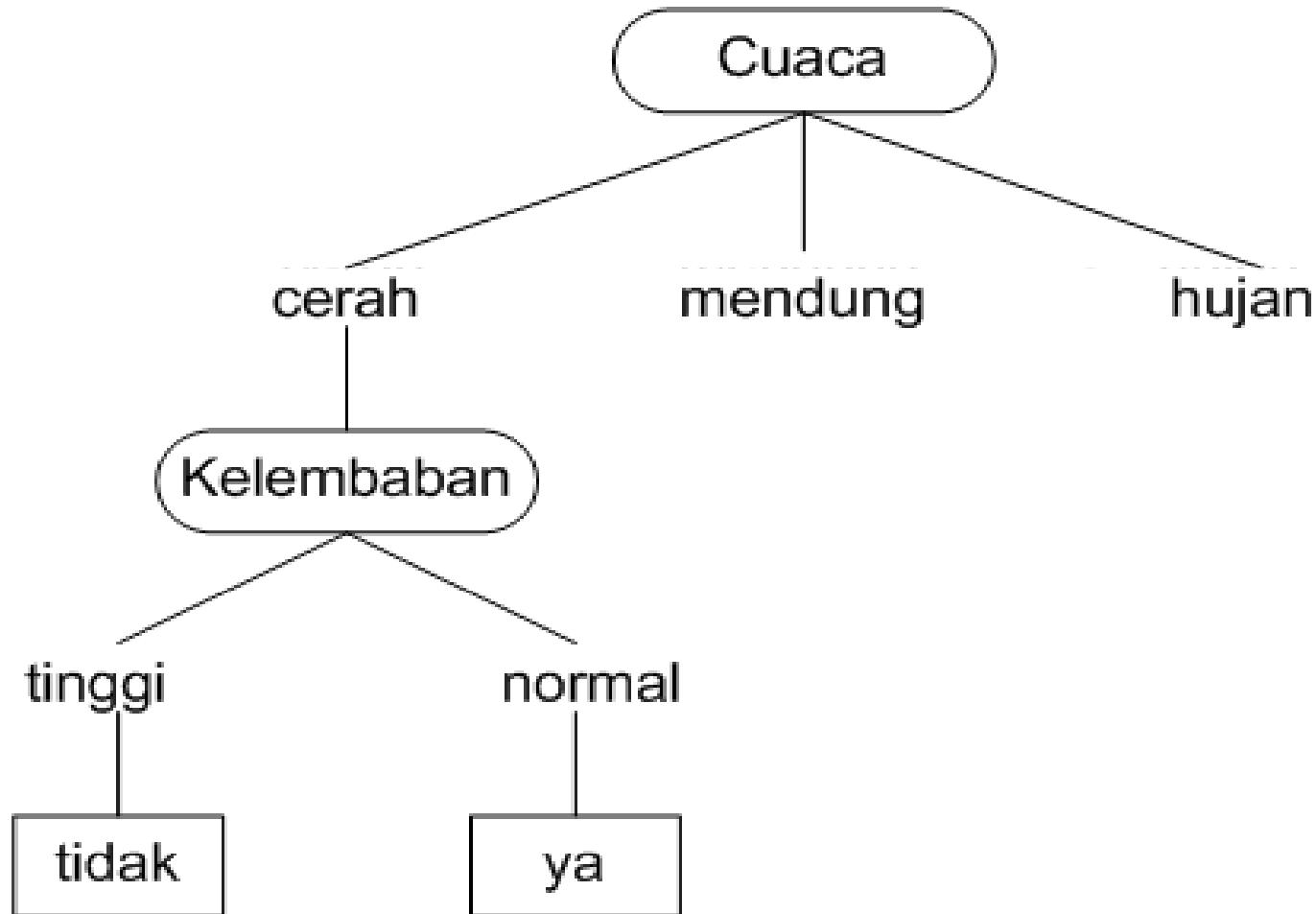
A. INFORMATION GAIN

Dari *information gain* tersebut dapat dilihat bahwa nilai $gain(y, A)$ yang terbesar adalah :

- Gain(kelembaban) = 0.971 bits
- Sehingga atribut **kelembaban** dapat dipilih sebagai atribut pemecahan kedua pada cabang Cuaca = cerah dalam *Decision Tree*

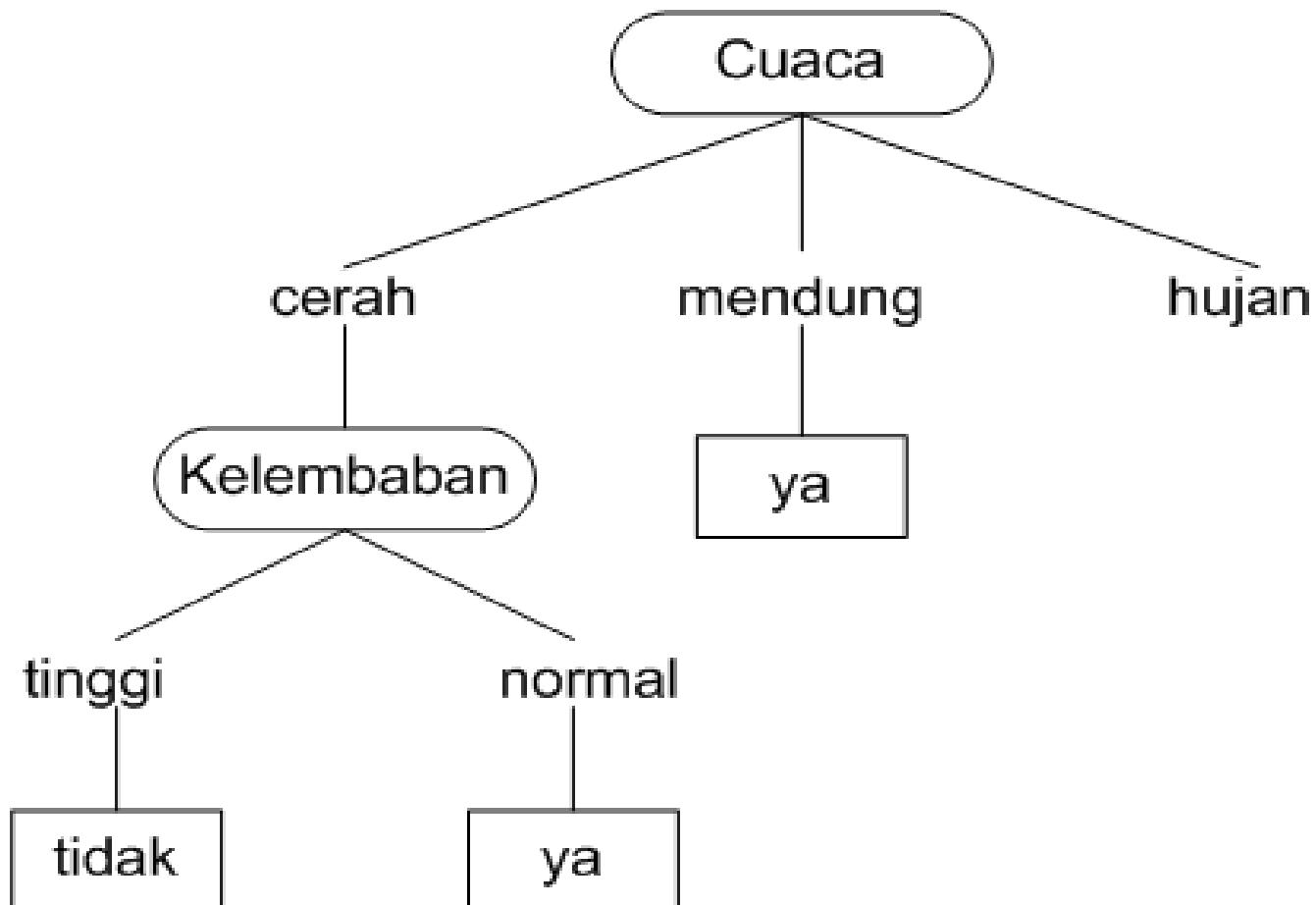
A. INFORMATION GAIN

- Pemecahan kedua pada *Decision Tree*



A. INFORMATION GAIN

- Pemecahan kedua pada *Decision Tree*





A. INFORMATION GAIN

- Iterasi dilakukan kembali pada cabang cuaca hujan.
- Pada cabang cuaca hujan, kita hitung $gain(y,A)$ atribut x_2 , dan x_4 .
- Diperoleh information gain, $gain(y,A)$:
 - Gain(temperatur) = 0.02 bits
 - Gain(angin) = 0.971 bits



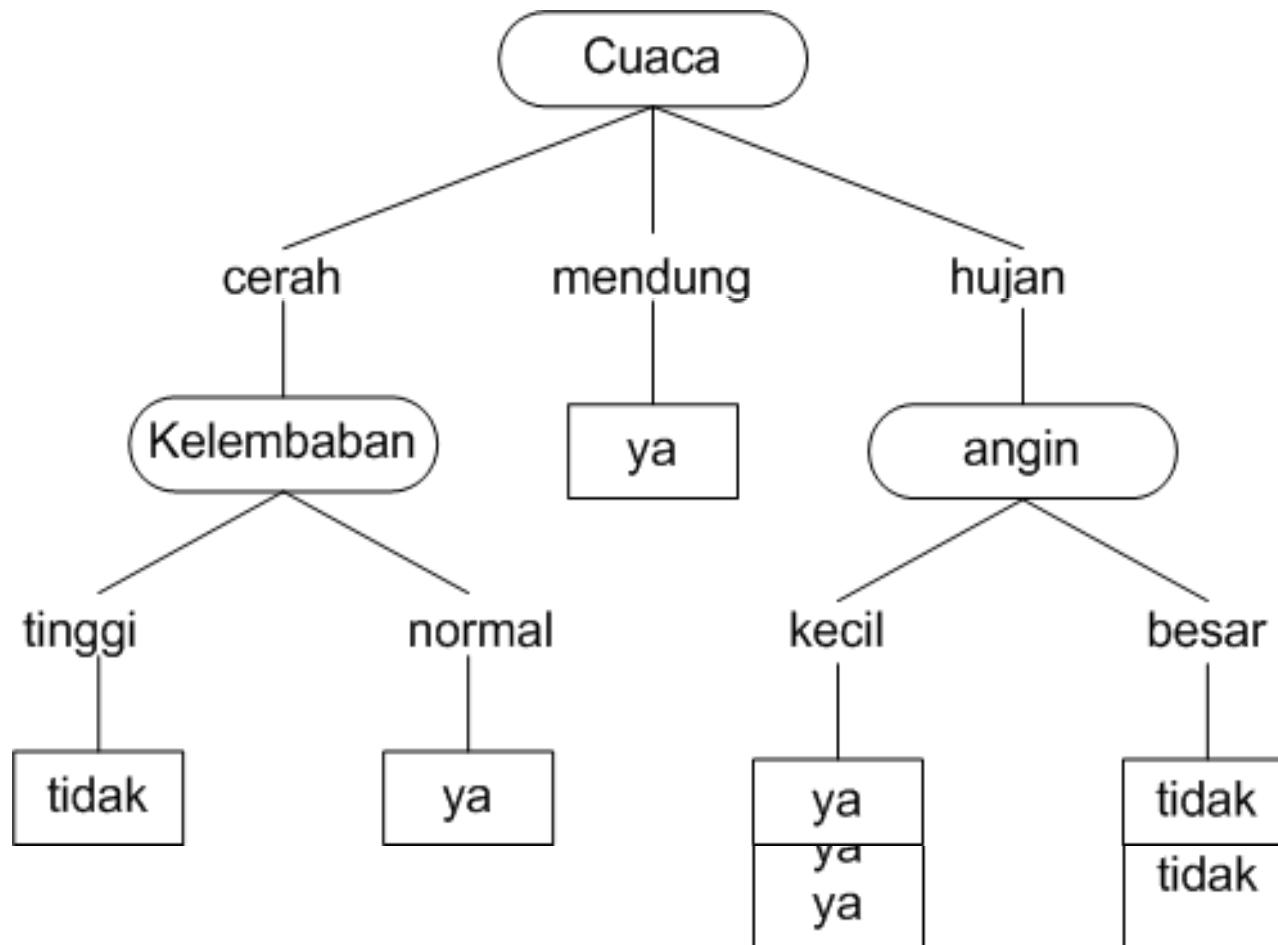
A. INFORMATION GAIN

Dari *information gain* tersebut dapat dilihat bahwa nilai $gain(y, A)$ yang terbesar adalah :

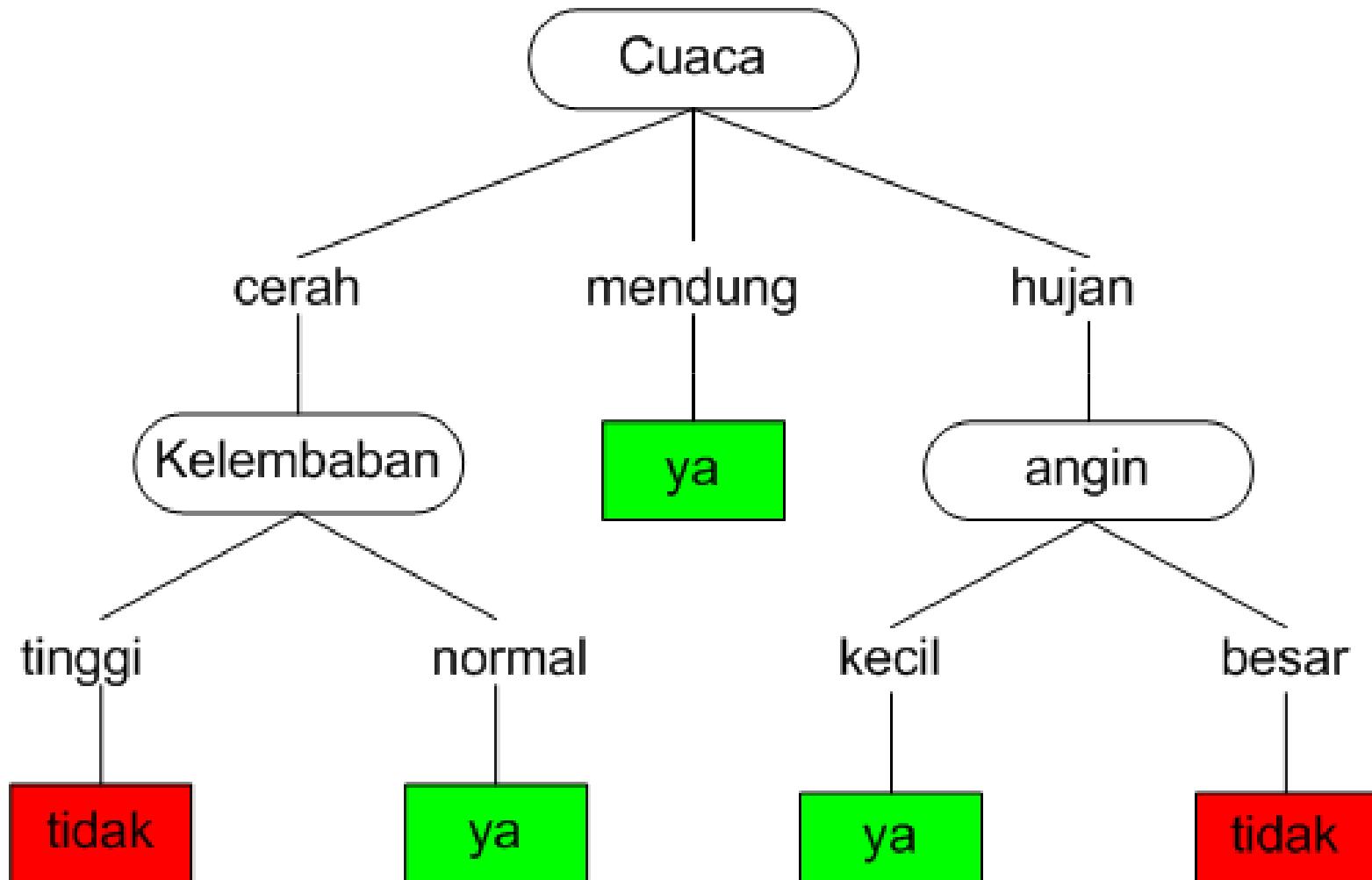
- Gain(angin) = 0.971 bits
- Sehingga atribut **angin** dapat dipilih sebagai atribut pemecahan pada cabang Cuaca = hujan dalam *Decision Tree*

A. INFORMATION GAIN

- Pemecahan kedua pada *Decision Tree*



Decision Tree – Information Gain



Cukup mudahkan?



Pemilihan Kriteria Atribut

