DATA MINING
METODE KLASIFIKASI

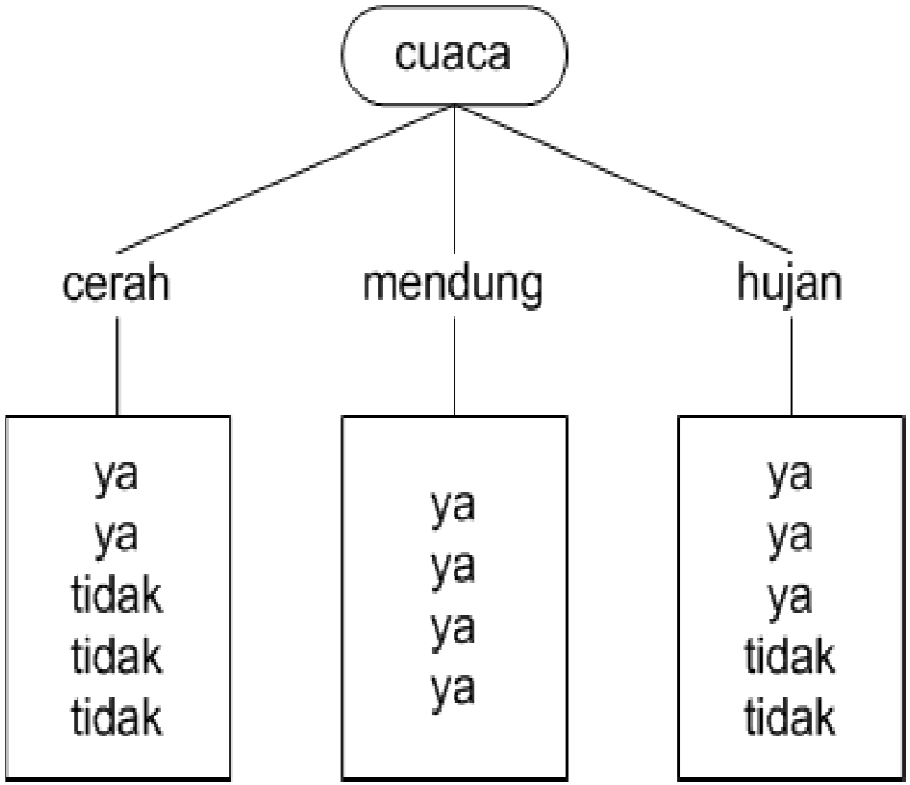
DECISION TREE: PEMILIHAN ATRIBUT BERDASARKAN INDEKS GINI

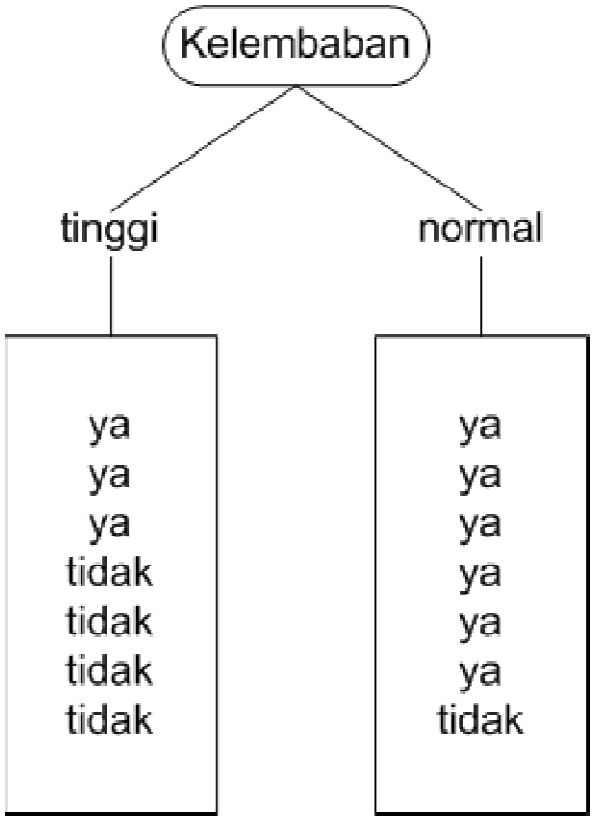
DATA TRAINING

Tabel 1. Data Cuaca dan Keputusan

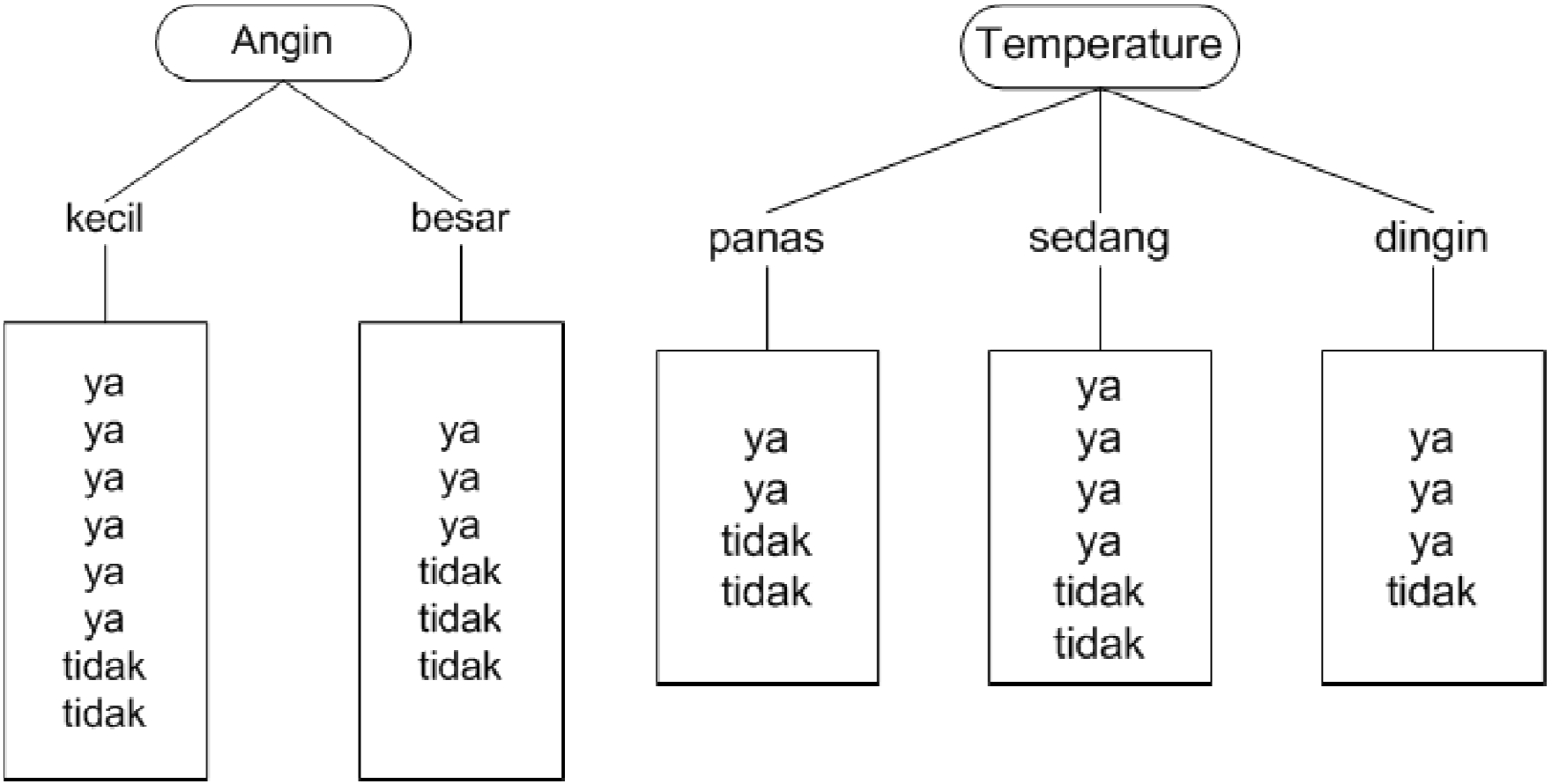
No	Cuaca X1	Temperature X2	Kelembaban X3	Angin X4	Main / Tidak Y
1	Cerah	Panas	Tinggi	Kecil	Tidak
2	Cerah	Panas	Tinggi	Besar	Tidak
3	Mendung	Panas	Tinggi	Kecil	Ya
4	Hujan	Sedang	Tinggi	Kecil	Ya
5	Hujan	Dingin	Normal	Kecil	Ya
6	Hujan	Dingin	Normal	Besar	Tidak
7	Mendung	Dingin	Normal	Besar	Ya
8	Cerah	Sedang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



Indeks Gini – Gini(S)

- Jika kelas obyek dinyatakan dengan k, k – 1, 2, …, C, dimana C adalah jumlah kelas untuk variabel / output dependen y, maka Indeks Gini untuk suatu cabang atau kotak A dihitung menggunakan persamaan :

$$Gini(S) = 1 - \sum_{k=1}^{c} {p_k}^2$$

Indeks Gini – Gini(A)

- Dimana p_k adalah rasio observasi dalam kotak A yang masuk dalam kelas k.
- Jika Gini(A) = 0, maka semua data dalam kotak A berasal dari kelas yang sama.
- Nilai Gini(A) mencapai maksimum jika dalam kelas A proporsi data dari masing-masing kelas yang ada mencapai nilai yang sama.

Indeks Gini – Gini(A)

- Hasil Perhitungan
  - Cuaca = cerah

    IG[2,3] = 0.480 bits
  - Cuaca = mendung

    IG[4,0] = 0 bits
  - Cuaca = hujan

    IG[3,2] = 0.480 bits

INDEKS GINI SPLIT

- Jika dataset A dibelah ke dalam dua subset, A1 dan A2, dengan size N1 dan N2, secara berurutan, Gini index dari data yang terbelah berisi contoh dari kelas n, dan *Gini index* didefinisikan dengan:

$$Gini_{split}(A) = \frac{N_1}{N} \cdot Gini(A_1) + \frac{N_2}{N} \cdot Gini(A_2) + \cdots + \frac{N_k}{N} \cdot Gini(A_k)$$

INDEKS GINI SPLIT

Dari semua atribut, jika dihitung *indeks gini split*-nya adalah:

- Cuaca = 0.343 bits
- Temperatur = 0.440 bits
- Kelembaban = 0.367 bits
- Angin = 0.429 bits

INDEKS GINI SPLIT

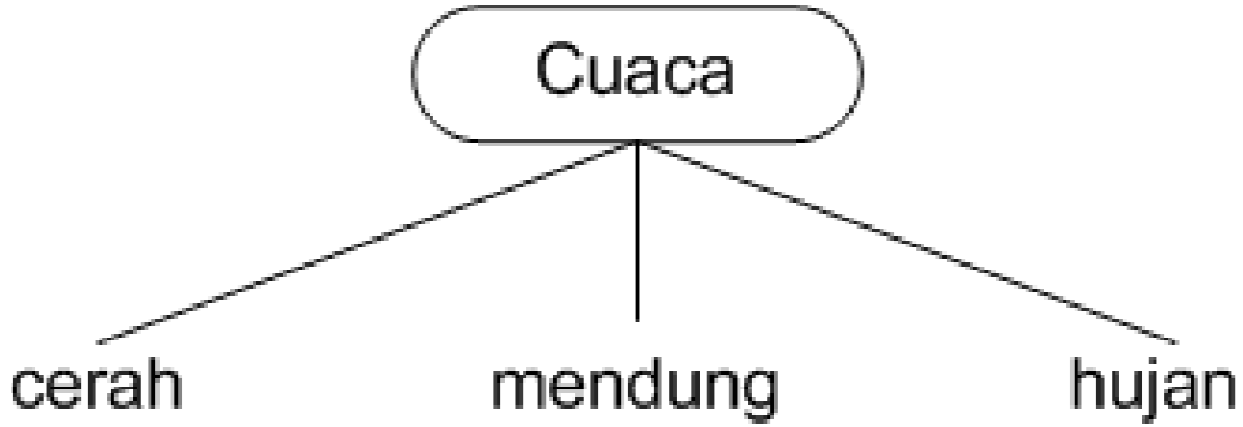
Dari *indeks gini split* semua atribut dapat dilihat bahwa nilai GiniSplit(A) yang **terkecil** adalah :

- Cuaca = 0.343 bits

- Sehingga atribut cuaca dapat dipilih sebagai atribut pemecahan pertama dalam *Decision Tree*

INDEKS GINI

- Pemecahan pertama pada *Decision Tree*



INDEKS GINI

- Iterasi dilakukan kembali pada setiap cabang cuaca.
- Pada cabang cuaca cerah, kita hitung *GiniSplit(A)* setiap atribut x_2, x_3, dan x_4.
- Diperoleh indeks gini, Gini(A)
  - GiniSplit(temperatur) = 0.200 bits
  - GiniSplit(kelembaban) = 0.000 bits
  - GiniSplit(angin) = 0.467 bits

INDEKS GINI

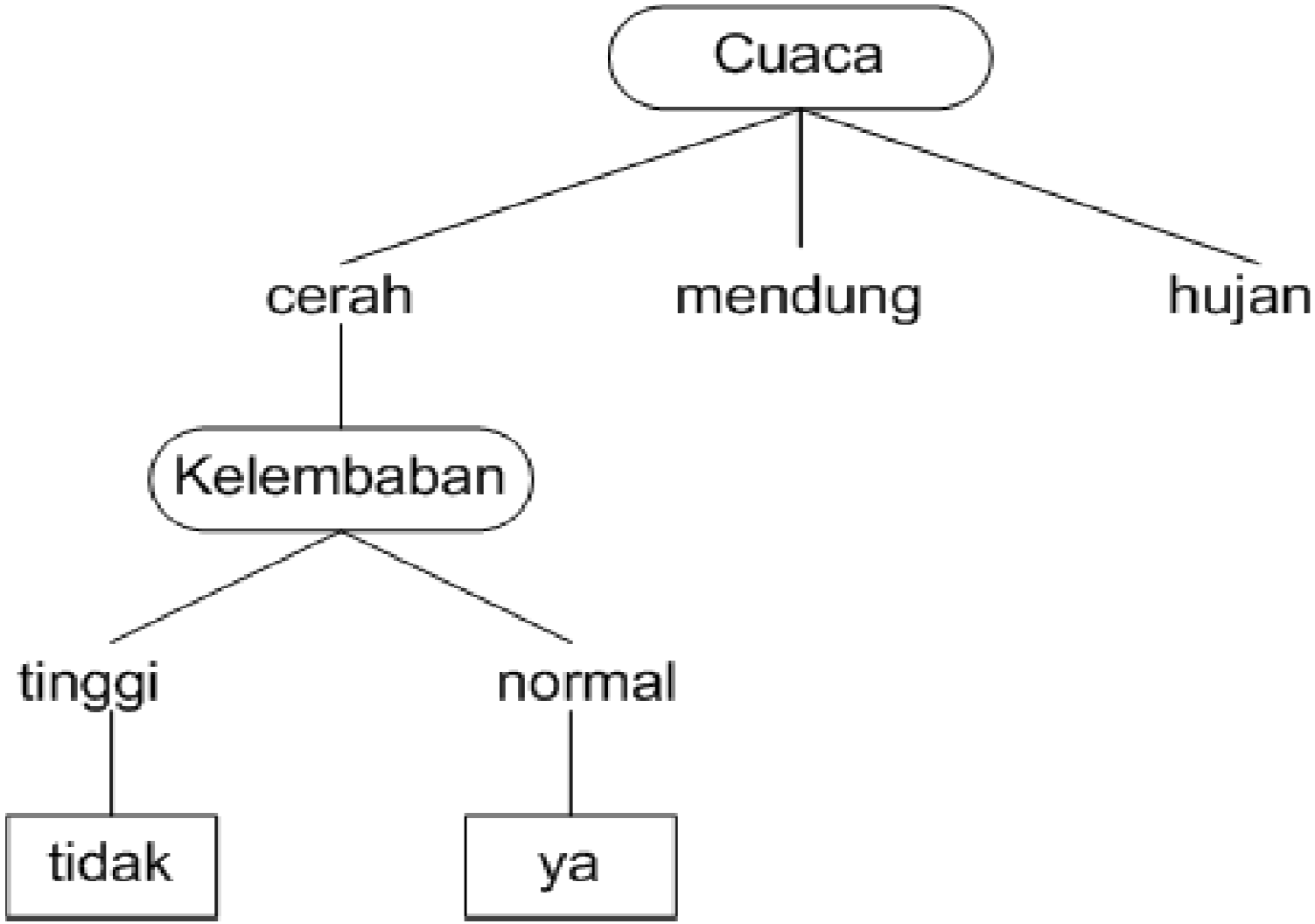
Dari *indeks gini* tersebut dapat dilihat bahwa nilai GiniSplit yang terkecil adalah :

- GiniSplit(Kelembaban) = 0.000 bits

- Sehingga atribut **kelembaban** dapat dipilih sebagai atribut pemecahan kedua pada cabang Cuaca = cerah dalam *Decision Tree*

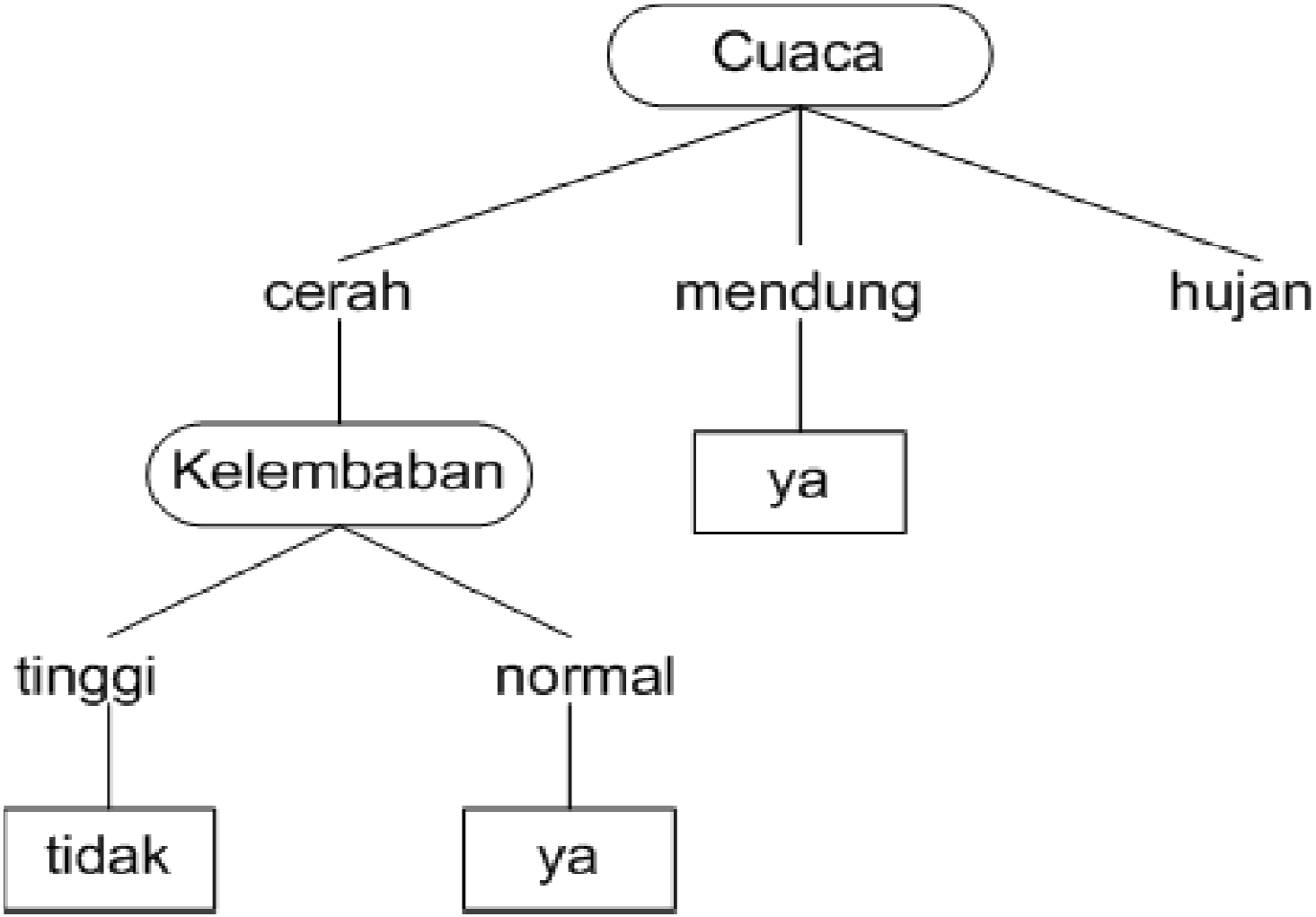
INDEKS GINI

- Pemecahan kedua pada *Decision Tree*



INDEKS GINI

- Pemecahan kedua pada *Decision Tree*



INDEKS GINI

- Iterasi dilakukan kembali pada cabang cuaca hujan.
- Pada cabang cuaca hujan, kita hitung *GiniSplit(A)* atribut x_2, dan x_4.
- Diperoleh indeks gini, GiniSplit(A):
  - GiniSplit(temperatur) = 0.467 bits
  - GiniSplit(angin) = 0.000 bits

INDEKS GINI

Dari *indeks gini* tersebut dapat dilihat bahwa nilai GiniSplit(A) yang terkecil adalah :

- GiniSplit(angin) = 0.000 bits

- Sehingga atribut **angin** dapat dipilih sebagai atribut pemecahan pada cabang Cuaca = hujan dalam *Decision Tree*

Decision Tree – INDEKS GINI