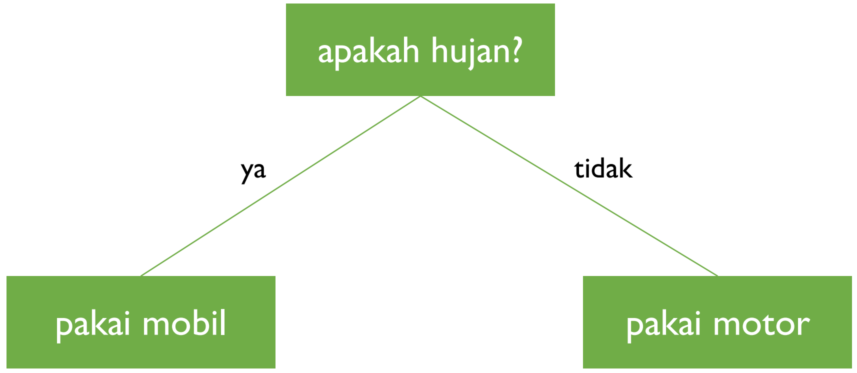
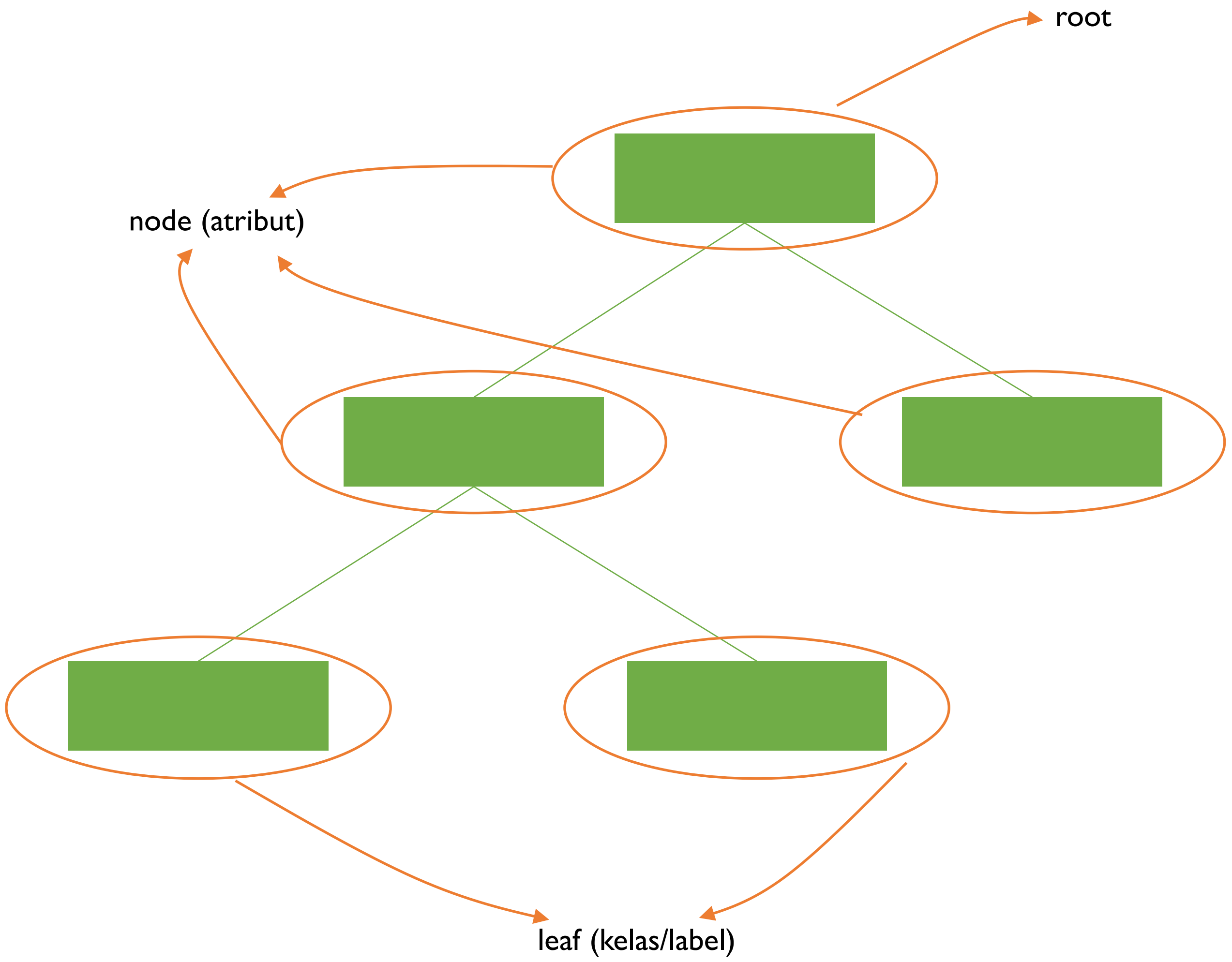
Decision Tree

Made Satria Wibawa, M.Eng.
2020

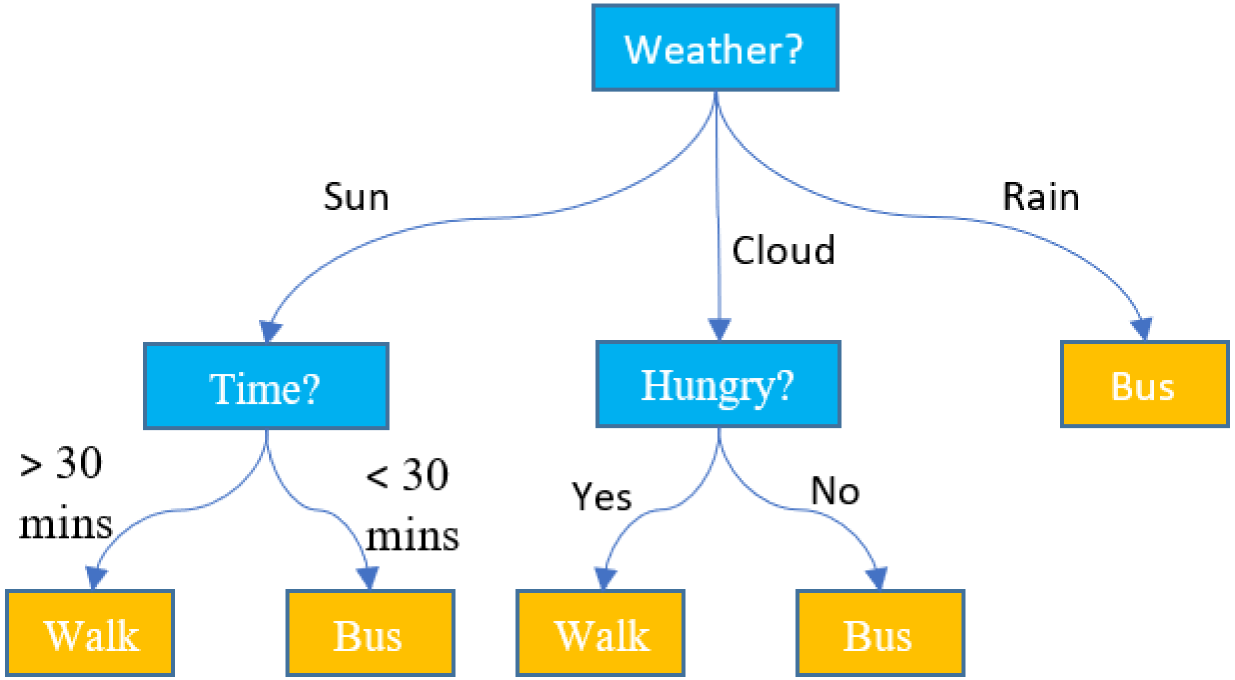
PENDAHULUAN

Tree Structure





Decision Tree



- Sekitar tahun 1970-1980 seorang peneliti bernama J. Ross Quinlan mengembangkan algoritma yang dikenal dengan nama ID3 (Iterative Dichotomiser)
- Kemudian Quinlan memperkenalkan algoritma C4.5 (pengembangan dari ID3)
- Algoritma inilah yang menjadi dasar dari decision tree saat ini.

KONSEP DECISION TREE

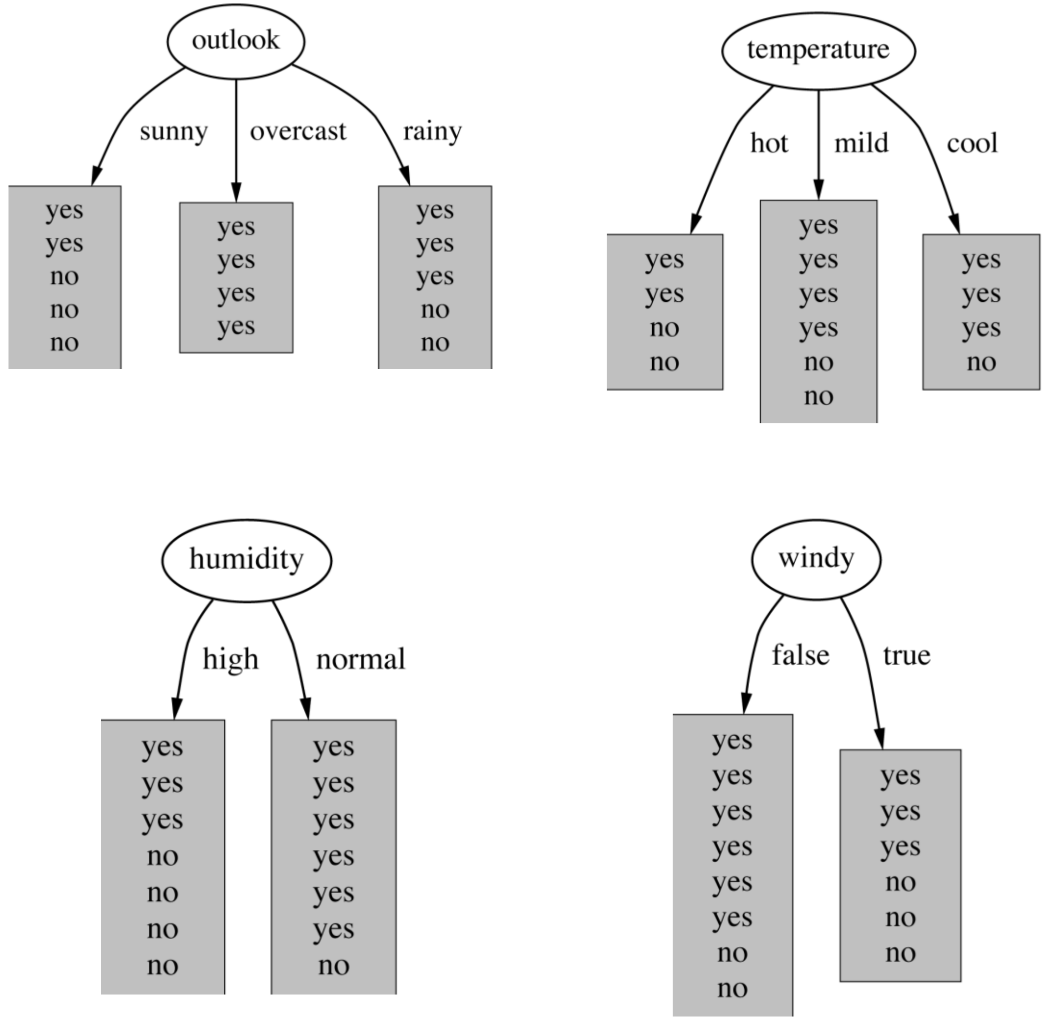
Parameter

- Dataset
- Daftar atribut
- Metode seleksi atribut

Metode seleksi atribut merupakan prosedur untuk menentukan splitting criterion yang dapat membagi data dengan cara 'terbaik' ke masing-masing kelas.

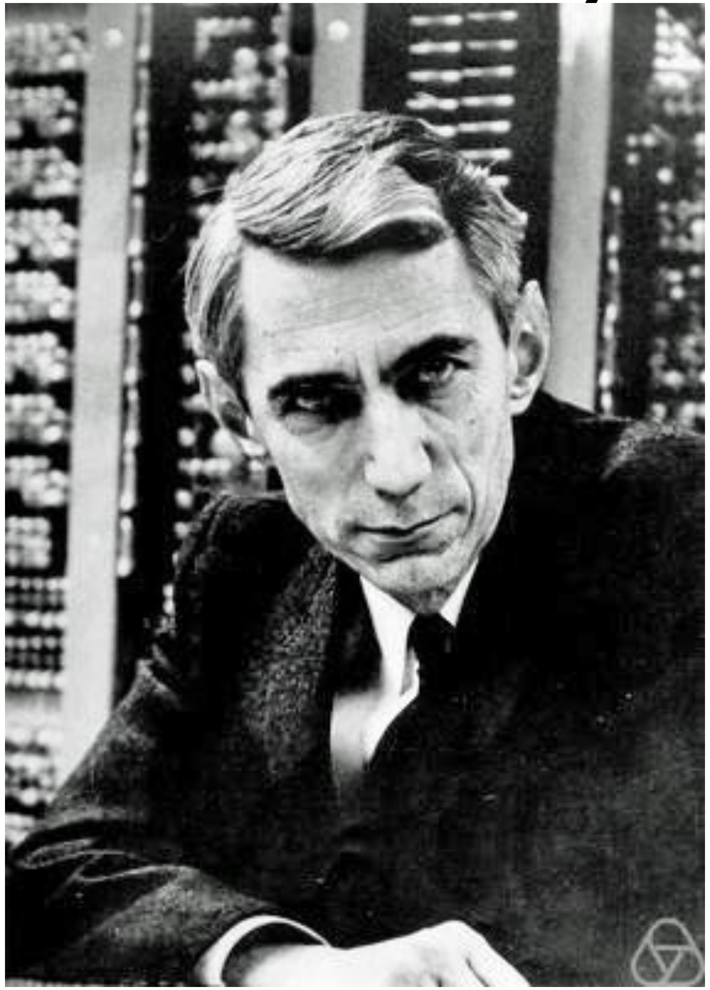
Splitting Criterion

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	high	false	yes
overcast	cool	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
rainy	mild	normal	false	yes
rainy	mild	high	true	no
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no
sunny	cool	normal	false	yes
sunny	mild	normal	true	yes



bagaimana caranya menentukan atribut yang dijadikan node?

Information Theory

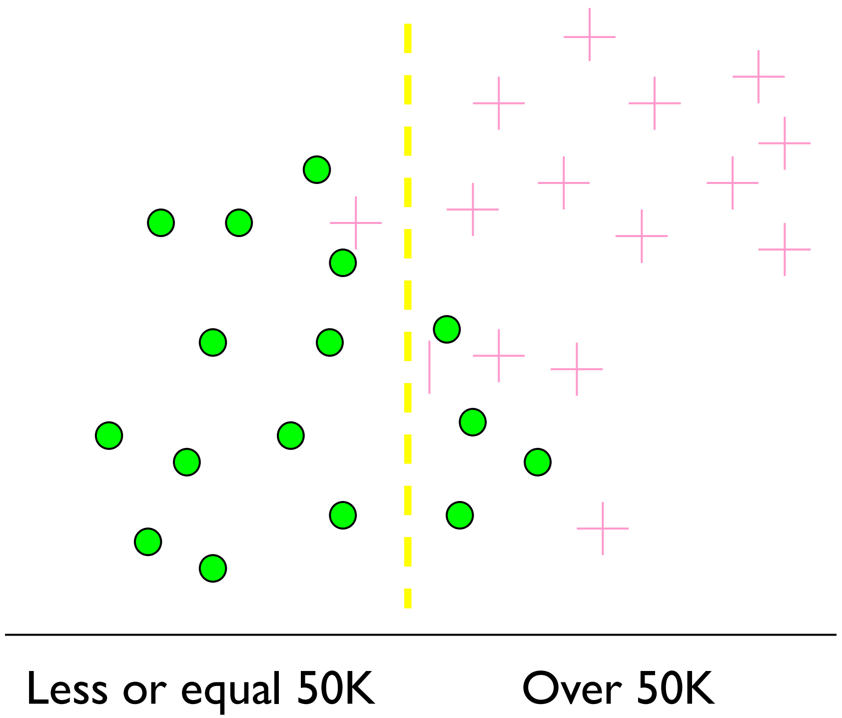


Claude Elwood Shannon
Father of Information Theory

"Information is the resolution of uncertainty."

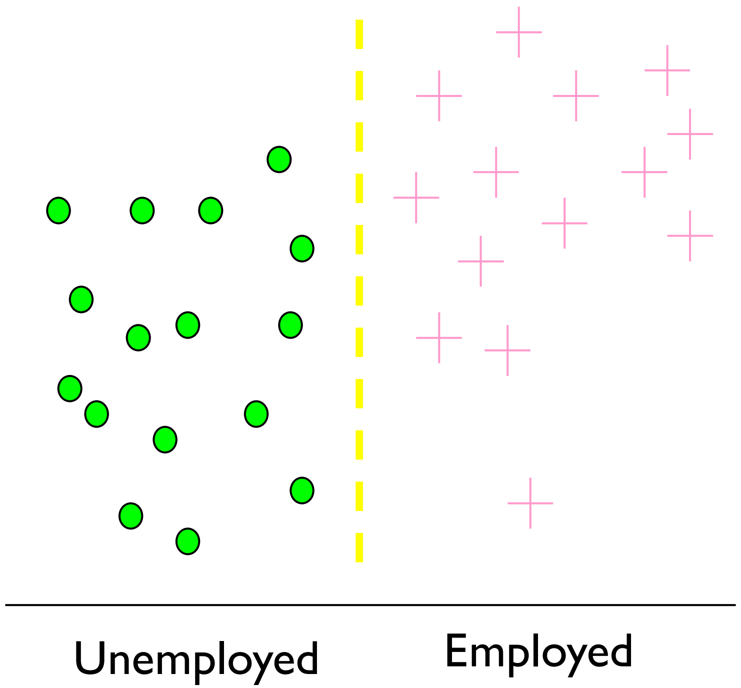
Information Gain

split dengan atribut GAJI



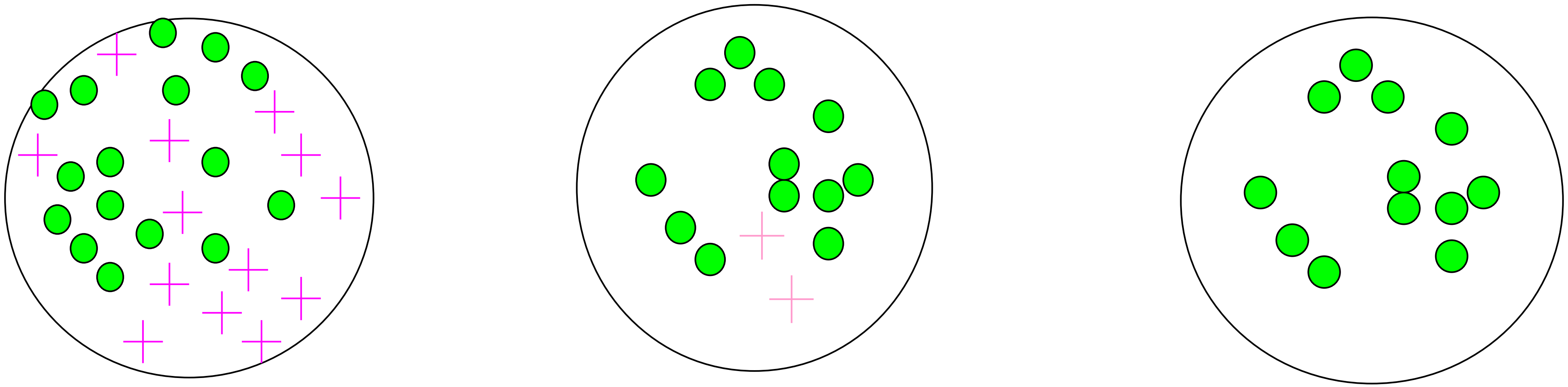
Atribut mana yang lebih **Informatif**?

split dengan atribut BEKERJA



Entropy

Entropy is measure of disorder



high entropy medium entropy low entropy

ALGORITMA DECISION TREE

Pengukuran Entropy: ID3

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j),$$

$$Gain(A) = Info(D) - Info_A(D)$$

- $Info(D)$ = Information gain dari kelas D
- $Info_A(D)$ = Information gain dari atribut A kelas D
- p_i = probabilitas objek ke-i
- m = jumlah kelas label
- p_i = probabilitas objek ke-i
- D = jumlah instance
- D_j = jumlah instance pada atribut ke-j

Pengukuran Entropy: C4.5

$$Split\ Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right),$$

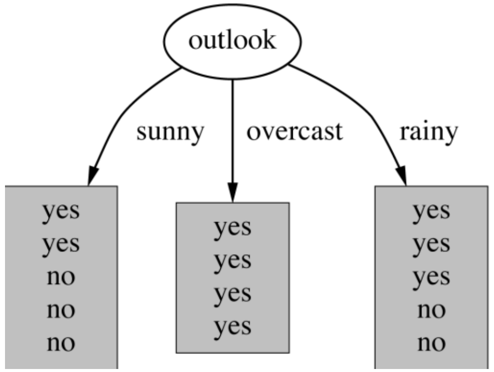
$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info_A(D)}$$

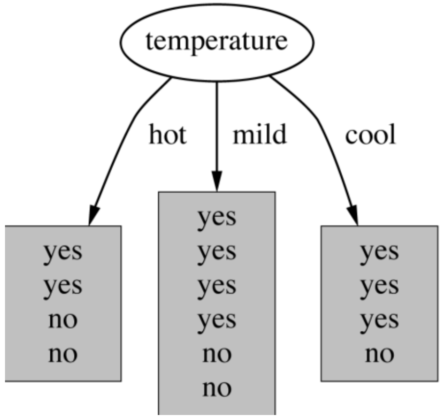
STUDI KASUS

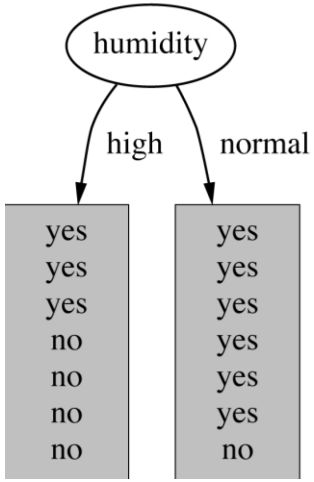
Dataset

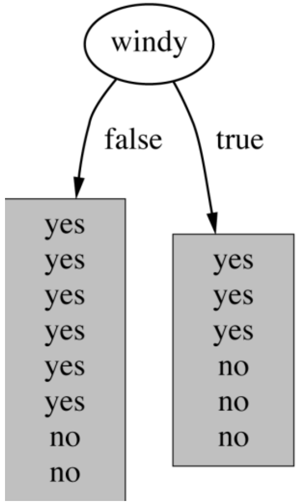
Outlook	Temperature	Humidity	Windy	Play
overcast	hot	high	false	yes
overcast	cool	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
rainy	mild	normal	false	yes
rainy	mild	high	true	no
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no
sunny	cool	normal	false	yes
sunny	mild	normal	true	yes

Dataset









Outlook (o)		
	yes	***no***
overcast	4	0
rainy	3	2
sunny	2	3

Temperature (t)		
	yes	***no***
hot	2	2
mild	4	2
cool	3	1

Humidity (h)		
	yes	***no***
high	3	4
normal	6	1

Windy (w)		
	yes	***no***
false	6	2
true	3	3

Play	
yes	***no***
9	5

ID3

Training-Node 1 Level 0 (Root)

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

Training-Node 1 Level 0 (Root)

Information Gain Tiap Atribut :

$$Info_o(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) = 0.694$$

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{4}{14} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{4}{14} \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{14} \times \left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) = 0.911$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{7}{14} \times \left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) + \frac{7}{14} \times \left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right) = 0.788$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{8}{14} \times \left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{6}{14} \times \left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{1}{7}\log_2\frac{1}{7}\right) = 0.892$$

Training-Node 1 Level 0 (Root)

Gain Tiap Atribut :

$Gain(o) = Info(D) - Info_o(D) = 0.940 - 0.694 = 0.246$ → ***splitting attribute***

$Gain(t) = Info(D) - Info_t(D) = 0.940 - 0.911 = 0.029$

$Gain(h) = Info(D) - Info_h(D) = 0.940 - 0.788 = 0.152$

$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0.892 = 0.048$

Pembentukan Tree

Outlook ?

overcast

Temperature	Humidity	Windy	Play
Hot	High	F	Yes
Cool	Normal	T	Yes
Mild	High	T	Yes
Hot	Normal	F	Yes

dalam satu kelompok STOP

rainy

Temperature	Humidity	Windy	Play
Mild	High	F	Yes
Cool	Normal	F	Yes
Cool	Normal	T	No
Mild	Normal	F	Yes
Mild	High	T	No

sunny

Temperature	Humidity	Windy	Play
Hot	High	F	No
Hot	High	T	No
Mild	High	F	No
Cool	Normal	F	Yes
Mild	Normal	T	Yes

lanjutkan split tree

Training-Node 2 Level 1

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.970$$

Training-Node 2 Level 1

Information Gain Tiap Atribut :

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 0.951$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{3}{3}\log_2\frac{3}{3}\right) + \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

Training-Node 2 Level 1

Gain Tiap Atribut :

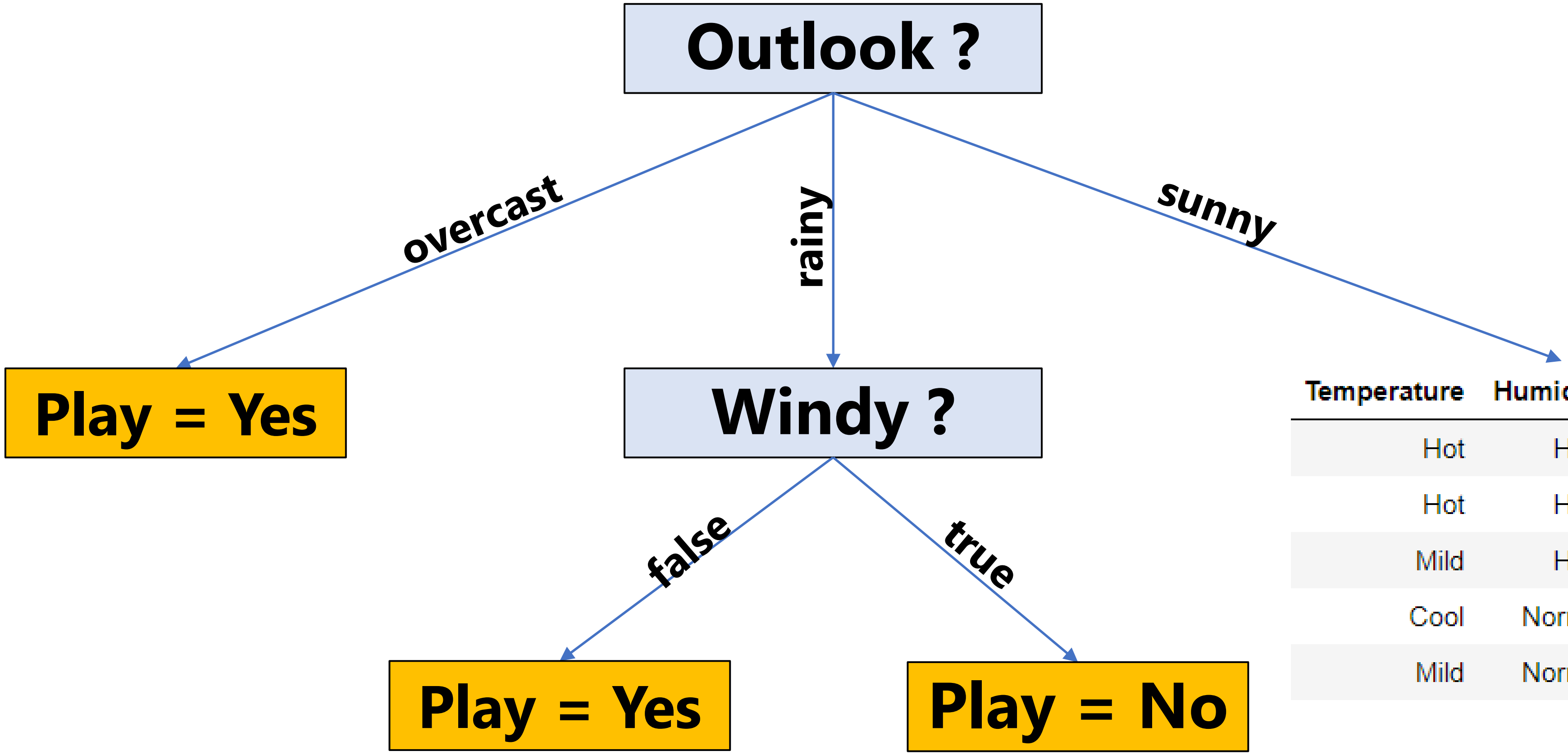
$$Gain(t) = Info(D) - Info_t(D) = 0.970 - 0.951 = 0.019$$

$$Gain(h) = Info(D) - Info_h(D) = 0.970 - 0.951 = 0.019$$

$$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0 = 0.940$$

splitting attribute

Pembentukan Tree



Temperature	Humidity	Windy	Play
Hot	High	F	No
Hot	High	T	No
Mild	High	F	No
Cool	Normal	F	Yes
Mild	Normal	T	Yes

lanjutkan split tree

Training-Node 3 Level 1

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.970$$

Training-Node 3 Level 1

Information Gain Tiap Atribut :

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2}\right) = 0.4$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{3}{3}\log_2\frac{3}{3}\right) + \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

Training-Node 3 Level 1

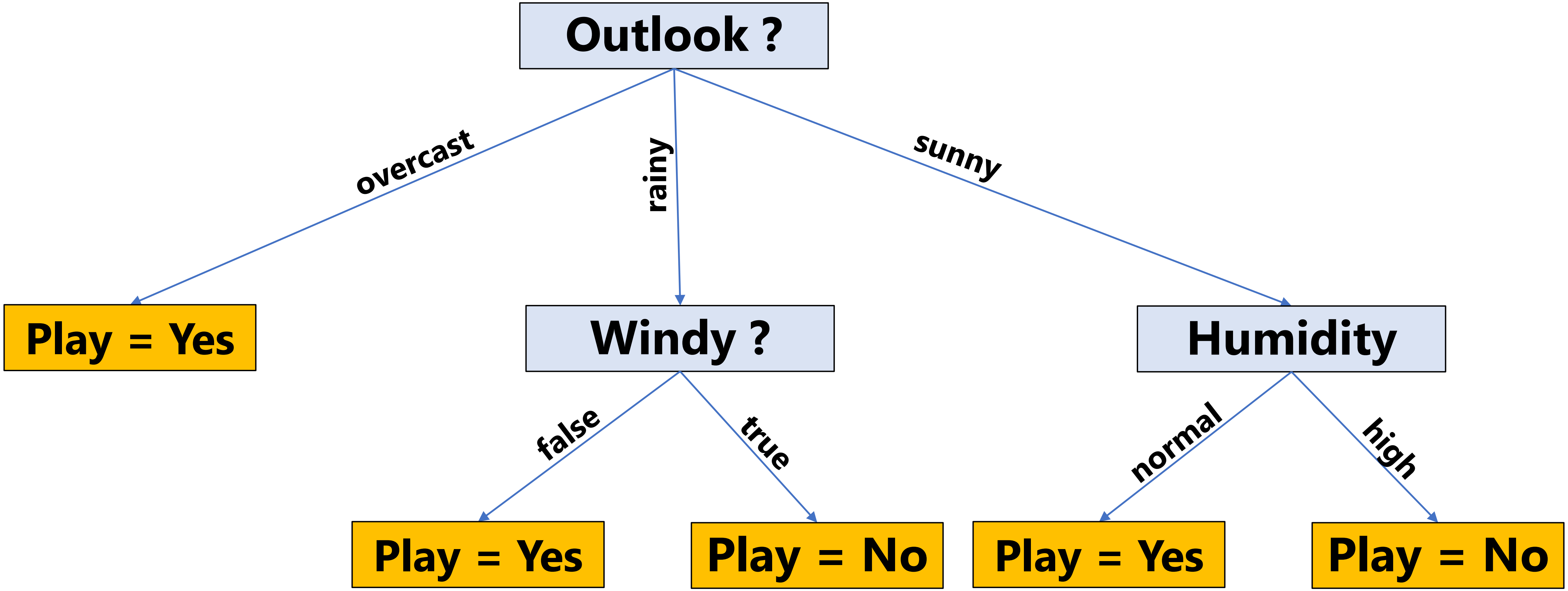
Gain Tiap Atribut :

$Gain(t) = Info(D) - Info_t(D) = 0.970 - 0.4 = 0.57$

$Gain(h) = Info(D) - Info_h(D) = 0.970 - 0 = 0.970$ → ***splitting attribute***

$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0.951 = 0.019$

Pembentukan Tree



Kekurangan ID3

- Pengukuran menggunakan parameter information gain akan bias terhadap atribut yang memiliki banyak nilai. ID3 akan lebih memilih atribut yang memiliki banyak nilai.

- Misalkan terdapat atribut yang berperan sebagai unique identifier seperti ID_produk. Setiap pembagian (split) akan menghasilkan jumlah yang partisi yang sangat banyak sejumlah instance. Karena setiap partisi memiliki low entropy maka atribut ini akan digunakan sebagai split atribut. Partisi semacam ini jelas tidak berguna untuk klasifikasi.

- Untuk itulah digunakan perhitungan rasio dari Gain pada algoritma C4.5

C4.5

Training-Node 1 Level 0 (Root)

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

Training-Node 1 Level 0 (Root)

Information Gain Tiap Atribut :

$$Info_o(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) = 0.694$$

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{4}{14} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{4}{14} \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{14} \times \left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) = 0.911$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{7}{14} \times \left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right) + \frac{7}{14} \times \left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right) = 0.788$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{8}{14} \times \left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{6}{14} \times \left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{1}{7}\log_2\frac{1}{7}\right) = 0.892$$

Training-Node 1 Level 0 (Root)

Gain Tiap Atribut :

$$Gain(o) = Info(D) - Info_o(D) = 0.940 - 0.694 = 0.246$$

$$Gain(t) = Info(D) - Info_t(D) = 0.940 - 0.911 = 0.029$$

$$Gain(h) = Info(D) - Info_h(D) = 0.940 - 0.788 = 0.152$$

$$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0.892 = 0.048$$

Training-Node 1 Level 0 (Root)

Split Info Tiap Atribut:

$$SplitInfo_o(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{5}{14} \times \log_2\left(\frac{5}{14}\right) - \frac{5}{14} \times \log_2\left(\frac{5}{14}\right) = 1.578$$

$$SplitInfo_t(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) = 1.556$$

$$SplitInfo_h(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{7}{14} \times \log_2\left(\frac{7}{14}\right) - \frac{7}{14} \times \log_2\left(\frac{7}{14}\right) = 1$$

$$SplitInfo_t(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{8}{14} \times \log_2\left(\frac{8}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) = 0.985$$

Training-Node 1 Level 0 (Root)

Gain Ration Tiap Atribut:

$$GainRatio(o) = \frac{Gain(o)}{SplitInfo_o(D)} = \frac{0.246}{1.578} = 0.156$$

splitting attribute

$$GainRatio(t) = \frac{Gain(t)}{SplitInfo_t(D)} = \frac{0.029}{1.556} = 0.019$$

$$GainRatio(h) = \frac{Gain(h)}{SplitInfo_h(D)} = \frac{0.152}{1} = 0.152$$

$$GainRatio(w) = \frac{Gain(w)}{SplitInfo_w(D)} = \frac{0.048}{0.985} = 0.049$$

Pembentukan Tree

Outlook ?

overcast

Temperature	Humidity	Windy	Play
Hot	High	F	Yes
Cool	Normal	T	Yes
Mild	High	T	Yes
Hot	Normal	F	Yes

dalam satu kelompok STOP

rainy

Temperature	Humidity	Windy	Play
Mild	High	F	Yes
Cool	Normal	F	Yes
Cool	Normal	T	No
Mild	Normal	F	Yes
Mild	High	T	No

sunny

Temperature	Humidity	Windy	Play
Hot	High	F	No
Hot	High	T	No
Mild	High	F	No
Cool	Normal	F	Yes
Mild	Normal	T	Yes

lanjutkan split tree

Training-Node 2 Level 1

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.970$$

Training-Node 2 Level 1

Information Gain Tiap Atribut :

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \right) = 0.951$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \right) + \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \right) = 0.951$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{3}{3}\log_2\frac{3}{3} \right) + \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2} \right) = 0$$

Training-Node 2 Level 1

Gain Tiap Atribut :

$$Gain(t) = Info(D) - Info_t(D) = 0.970 - 0.951 = 0.019$$

$$Gain(h) = Info(D) - Info_h(D) = 0.970 - 0.951 = 0.019$$

$$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0 = 0.940$$

Training-Node 2 Level 1

Split Info Tiap Atribut:

$$SplitInfo_t(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = 0.971$$

$$SplitInfo_h(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times \log_2\left(\frac{3}{5}\right) = 0.971$$

$$SplitInfo_w(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = 0.971$$

Training-Node 2 Level 1

Gain Ration Tiap Atribut:

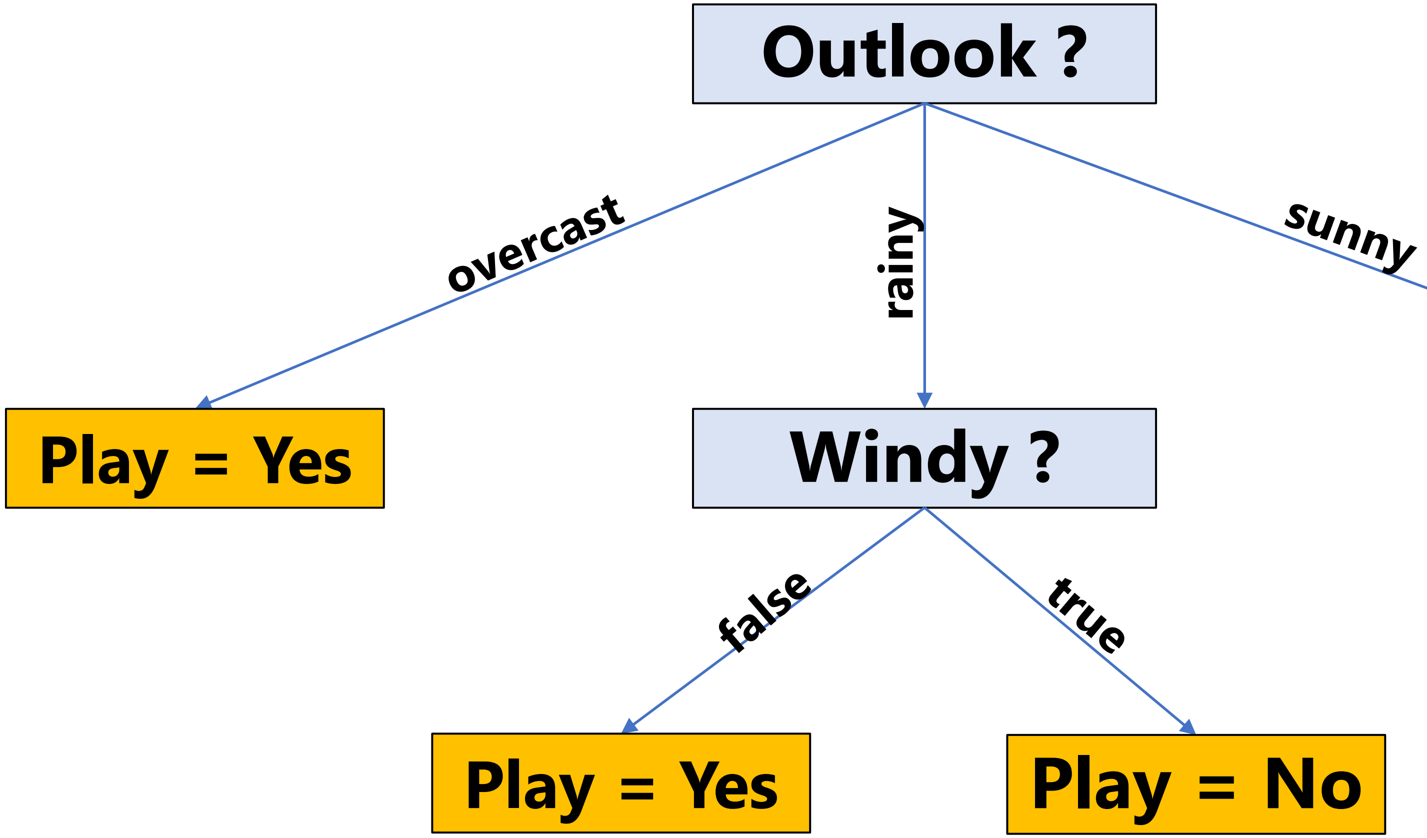
$$GainRatio(t) = \frac{Gain(t)}{SplitInfo_t(D)} = \frac{0.019}{0.971} = 0.02$$

$$GainRatio(h) = \frac{Gain(h)}{SplitInfo_h(D)} = \frac{0.019}{0.971} = 0.02$$

$$\boxed{GainRatio(w) = \frac{Gain(w)}{SplitInfo_w(D)} = \frac{0.940}{0.971} = 0.968} \Rightarrow$$

splitting attribute

Pembentukan Tree



Temperature	Humidity	Windy	Play
Hot	High	F	No
Hot	High	T	No
Mild	High	F	No
Cool	Normal	F	Yes
Mild	Normal	T	Yes

lanjutkan split tree

Training-Node 3 Level 1

Information Gain Kelas:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.970$$

Training-Node 3 Level 1

Information Gain Tiap Atribut :

$$Info_t(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2}\right) = 0.4$$

$$Info_h(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{3}{3}\log_2\frac{3}{3}\right) + \frac{2}{5} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

$$Info_w(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j), = \frac{3}{5} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{5} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.951$$

Training-Node 3 Level 1

Gain Tiap Atribut :

$Gain(t) = Info(D) - Info_t(D) = 0.970 - 0.4 = 0.57$

$Gain(h) = Info(D) - Info_h(D) = 0.970 - 0 = 0.970$ → ***splitting attribute***

$Gain(\mathrm{w}) = Info(D) - Info_w(D) = 0.940 - 0.951 = 0.019$

Training-Node 3 Level 1

Split Info Tiap Atribut:

$$SplitInfo_t(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) - \frac{1}{5} \times \log_2\left(\frac{1}{5}\right) = 1.522$$

$$SplitInfo_h(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = 0.971$$

$$SplitInfo_w(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right), = -\frac{3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right) = 0.971$$

Training-Node 3 Level 1

Gain Ration Tiap Atribut:

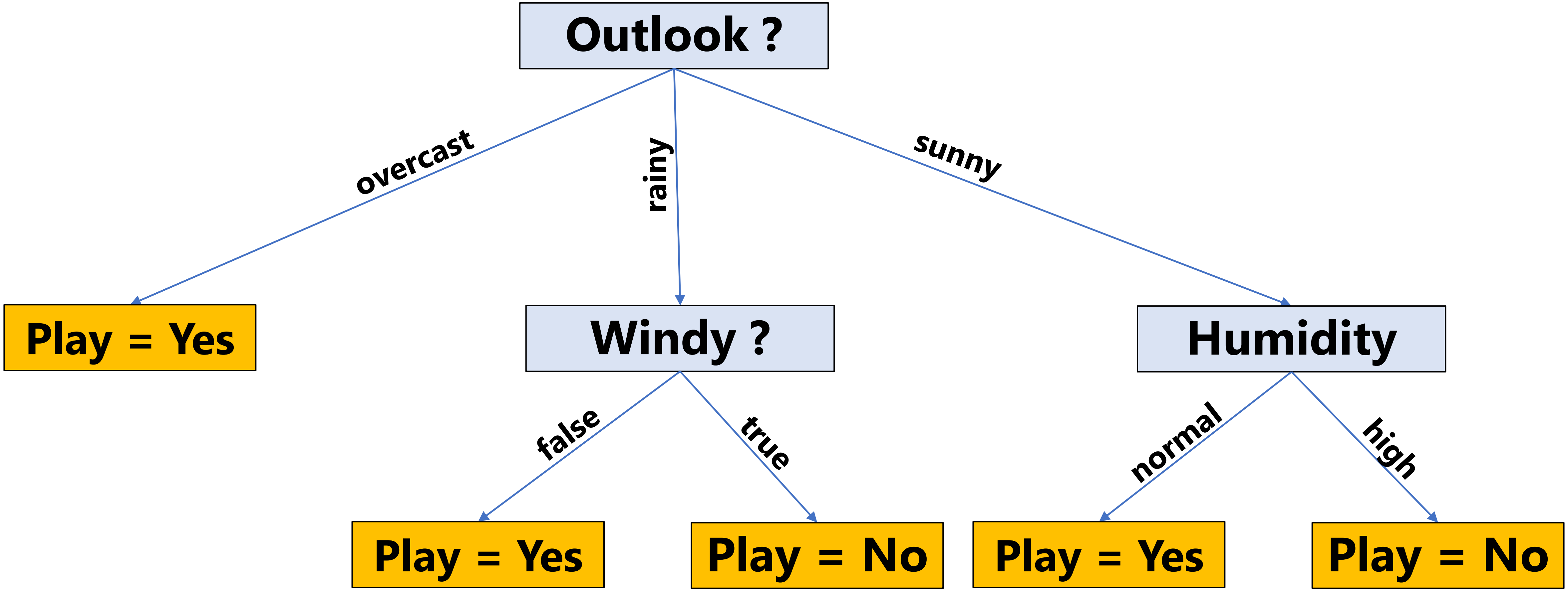
$$GainRatio(t) = \frac{Gain(t)}{SplitInfo_t(D)} = \frac{0.57}{1.522} = 0.375$$

$$GainRatio(h) = \frac{Gain(h)}{SplitInfo_h(D)} = \frac{0.970}{0.971} = 0.999$$

splitting attribute

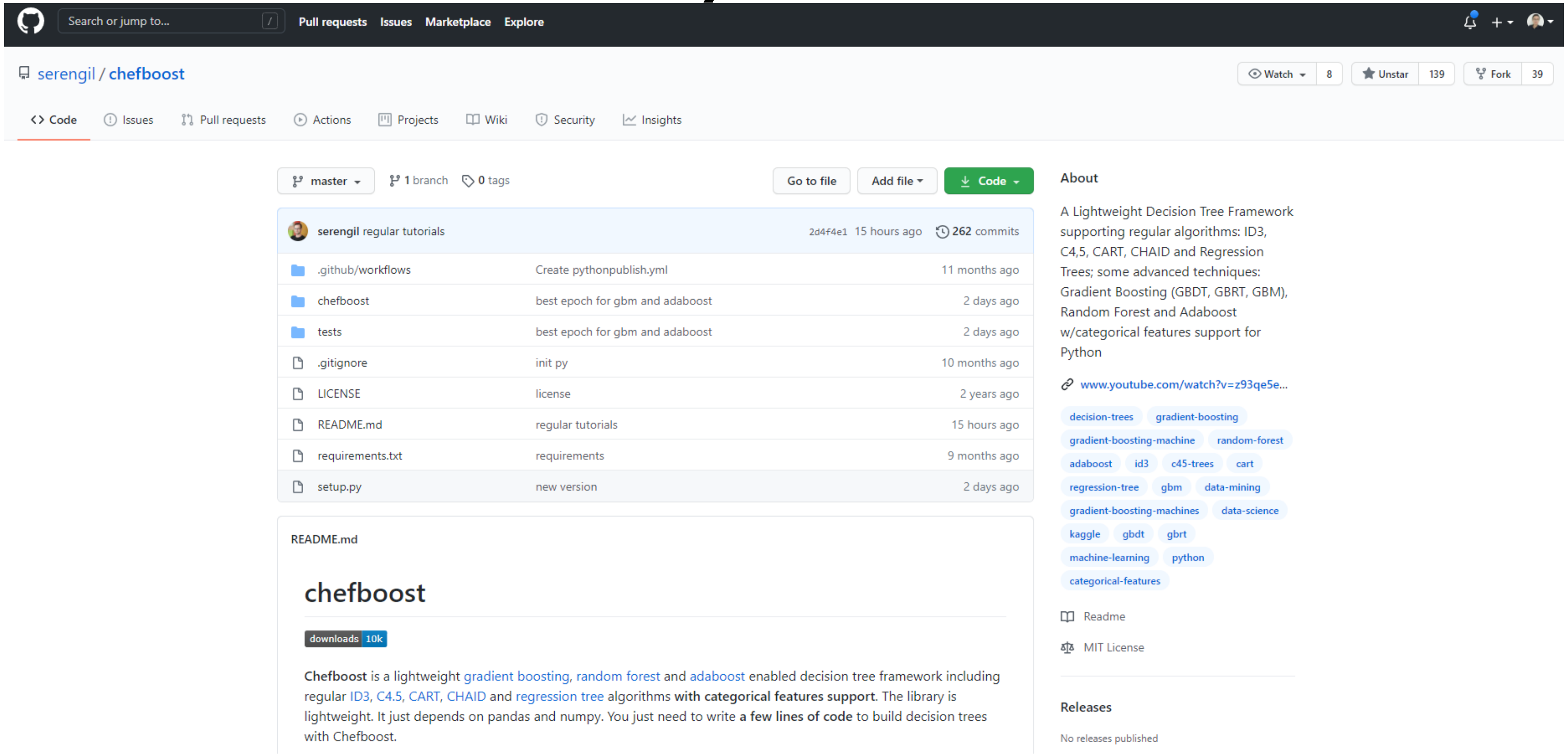
$$GainRatio(w) = \frac{Gain(w)}{SplitInfo_w(D)} = \frac{0.019}{0.971} = 0.02$$

Pembentukan Tree



IMPLEMENTASI PYTHON

Decision Tree Library



pip install chefboost

pertanyaan/troubleshooting silahkan buat di channel Diskusi Teams