Data Mining

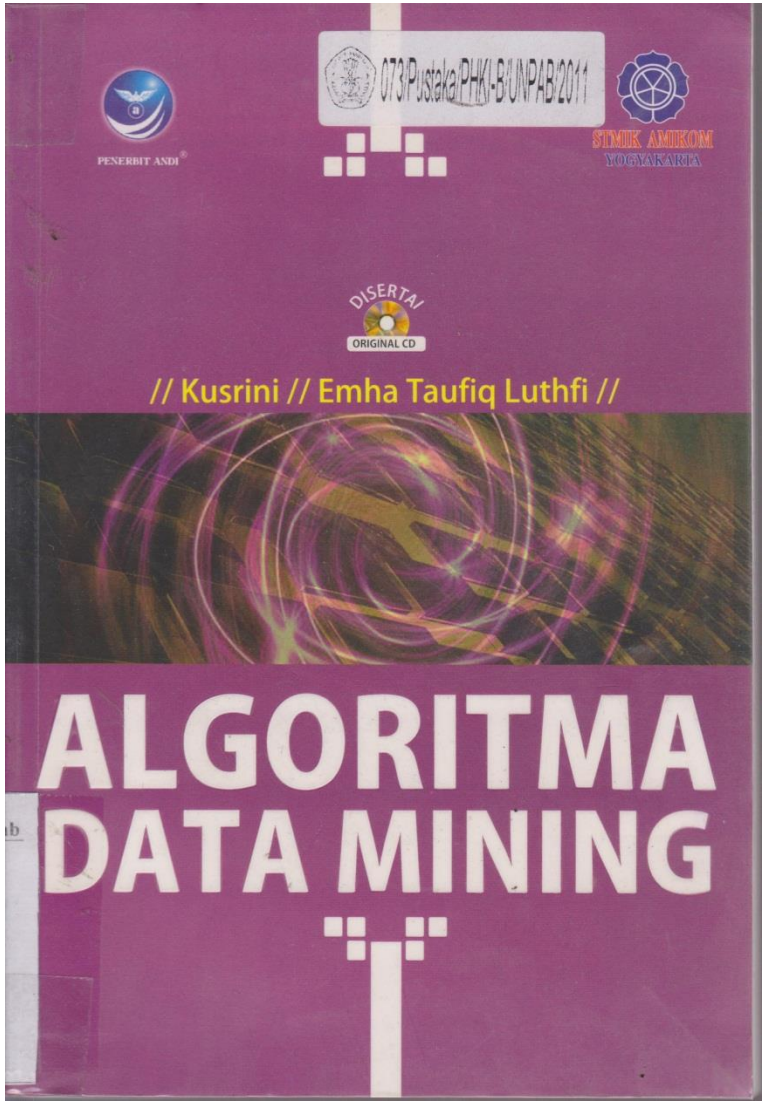
Junta Zeniarja, M.Kom, M.CS
junta@dsn.dinus.ac.id

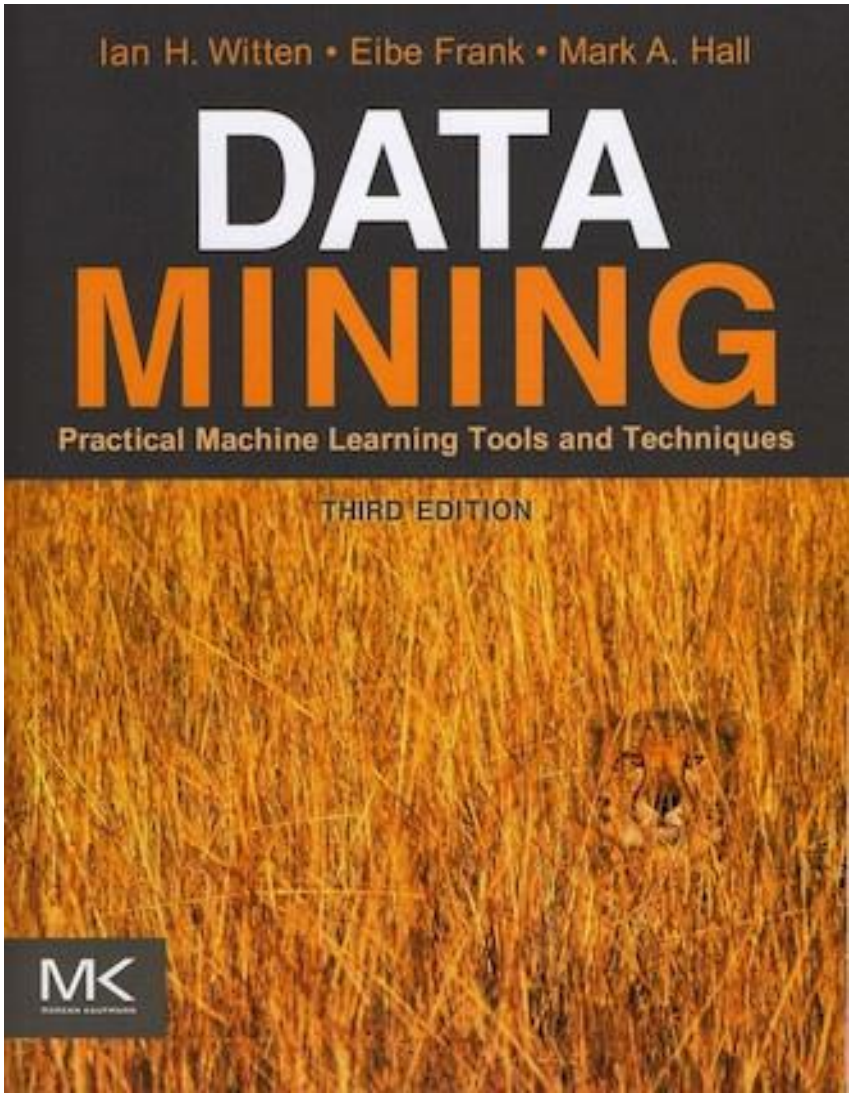
Profil

- Nama : Junta Zeniarja, M.Kom
- Alamat : Semarang
- Kontak
  - Phone : -
  - E-mail : junta@dsn.dinus.ac.id
  - Room : Ruang Dosen TI-S1 (H.2.3)

- Pendidikan
  - S1 => TI – UDINUS
  - S2 => TI – UDINUS
  - S2 => Computer Science UTeM (Universiti Teknikal Malaysia Melaka)
- Konsultasi - Sharing
  - 1:00 pm – 4:00 pm, Senin-Kamis.
  - Appointment via phone or e-mail preferred

Textbooks


PENERBIT ANDI
STMIK AMIKOM YOGYAKARTA
DISERTAI ORIGINAL CD
// Kusrini // Emha Taufiq Luthfi //
ALGORITMA DATA MINING


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Outline

1. Algoritma Data Mining
   - Algoritma C4.5
   - Algoritma Nearest Neighbor
   - Algoritma Apriori
   - Algoritma Fuzzy C Means
   - Bayesian Classification

Algoritma C4.5

Introduction

- Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*).
- Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal.
- Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.

Varian Algoritma Pohon Keputusan

- Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain : ID3, CART, dan C4.5 (Larose, 2005).
- Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Larose, 2005).
- Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule (Basuki & Syarif, 2003).

Contoh Data 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

Algoritma C4.5

- Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :
  1. Pilih atribut sebagai akar.
  2. Buat cabang untuk tiap-tiap nilai.
  3. Bagi kasus dalam cabang.
  4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Konsep Entropy

- Entropy (S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S.
- Entropy dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas.
- Entropy digunakan untuk mengukur ketidakaslian S.

Konsep Entropy (2)

- Untuk perhitungan nilai Entropy sbb :

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

- Keterangan :
  - S : himpunan kasus.
  - A : fitur.
  - n : jumlah partisi S.
  - pi : proporsi dari S_i terhadap S

Konsep Gain

- Gain (S,A) merupakan perolehan informasi dari atribut A relative terhadap output data S.
- Perolehan informasi didapat dari output data atau variable dependent S yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (S,A).

Konsep Gain (2)

- Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada.
- Untuk menghitung *gain* digunakan rumus :

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

- Keterangan :
  - S : himpunan kasus
  - A : atribut
  - n : jumlah partisi atribut A
  - $|S_i|$: jumlah kasus pada partisi ke-i
  - $|S|$: jumlah kasus dalam S

Langkah 1

- 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 perhitungan Gain untuk setiap atribut.
- Hasil perhitungan ditunjukan di bawah ini.

Perhitungan Node 1

		jml kasus(S)	Tidak (S1)	Ya (S2)	Entropy	Gain
total		14	4	10	0.86312	
outlook						0.258521
	cloudy	4	0	4	0	
	rainy	5	1	4	0.72193	
	sunny	5	3	2	0.97095	
temp						0.1838509
	col	4	0	4	0	
	hot	4	2	2	1	
	mild	6	2	4	0.9183	
humidity						0.3705065
	high	7	4	3	0.98523	
	normal	7	0	7	0	
windy						0.0059777
	FALSE	8	2	6	0.81128	
	TRUE	6	4	2	0.9183	

Cara Perhitungan Node 1 (1)

- Baris total kolom Entropy dihitung dengan persamaan:

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

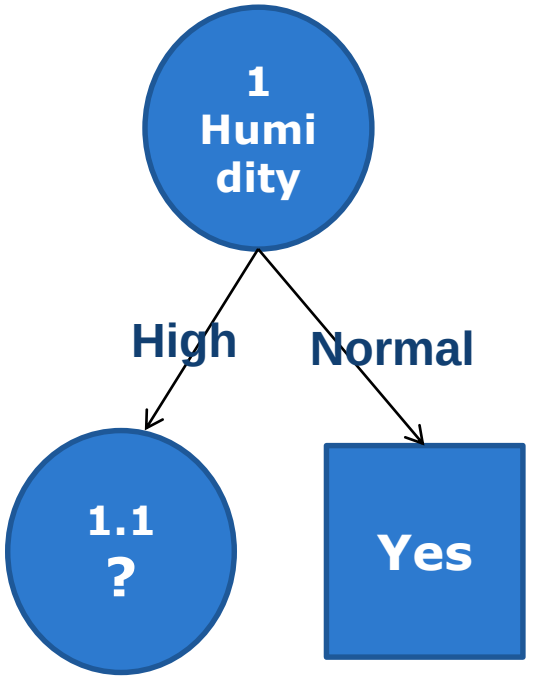
$Entropy(Total)$ =0,863120569

Cara Perhitungan Node 1 (2)

- Nilai gain pada baris **OUTLOOK** dihitung :
- $Gain(Total, Outlook) = Entropy(Total) - \sum_{i=1}^{n} \frac{|Outlook_i|}{|Total|} * Entropy\,(Outlook_i)$
- $Gain(Total, Outlook) = 0.863120569 - \left(\left(\frac{4}{14} * 0\right) + \left(\frac{5}{14} * 0.723\right) + \left(\frac{5}{14} * 0.97\right)\right)$
- $Gain(Total, Outlook) = 0.23$

Cara Perhitungan Node 1 (3)

- Dari hasil diketahui bahwa atribut dengan gain tertinggi adalah **HUMIDITY** yaitu sebesar 0.37. Sehingga **HUMIDITY** dapat menjadi node akar.
- Ada dua nilai atibut dari **HUMIDITY**, yaitu **HIGH** dan **NORMAL**.
- Nilai atribut **NORMAL** sudah mengklasifikasikan kasus menjadi 1, yaitu keputusannya Yes, sehingga tidak perlu dilakukan perhitungan lebih lanjut.
- Tetapi untuk nilai **HIGH** masih perlu dilakukan perhitungan lagi.



Langkah 2

- Menghitung jumlah kasus, jumlah kasus untuk keputusan **Yes**, jumlah kasus untuk keputusan **No**.
- 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 perhitungan Gain, untuk tiap-tiap atribut.

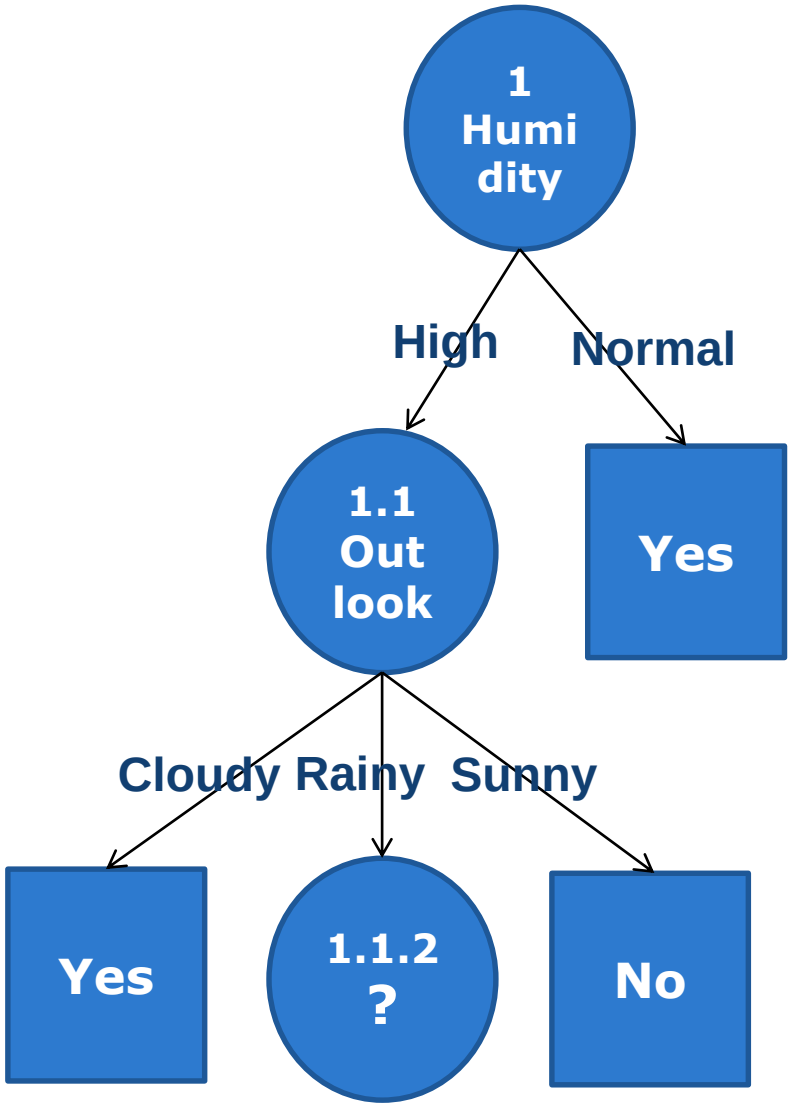
Perhitungan Node 1.1

		jml kasus(S)	Tidak (S1)	Ya (S2)	Entropy	Gain
Humidity High		7	4	3	0.98522814	
outlook						0.69951385
	cloudy	2	0	2	0	
	rainy	2	1	1	1	
	sunny	3	3	0	0	
temp						0.02024421
	col	0	0	0	0	
	hot	3	2	1	0.91829583	
	mild	4	2	2	1	
windy						0.02024421
	FALSE	4	2	2	1	
	TRUE	3	2	1	0.91829583	

Cara Perhitungan Node 1.1 (1)

- Atribut dengan Gain tertinggi adalah OUTLOOK, yaitu sebesar 0.6995.
- Sehingga OUTLOOK dapat menjadi node cabang dari nilai atribut HIGH.
- Ada tiga nilai dari atribut OUTLOOK yaitu CLOUDY, RAINY dan SUNNY.
  - CLOUDY => klasifikasi kasus 1 (Yes)
  - SUNNY => klasifikasi kasus 1 (No)
  - RAINY => masih perlu perhitungan lagi.

Cara Perhitungan Node 1.1 (2)



Langkah 3

- Menghitung jumlah kasus, jumlah kasus untuk keputusan **Yes**, jumlah kasus untuk keputusan **No**.
- 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 perhitungan Gain, untuk tiap-tiap atribut.

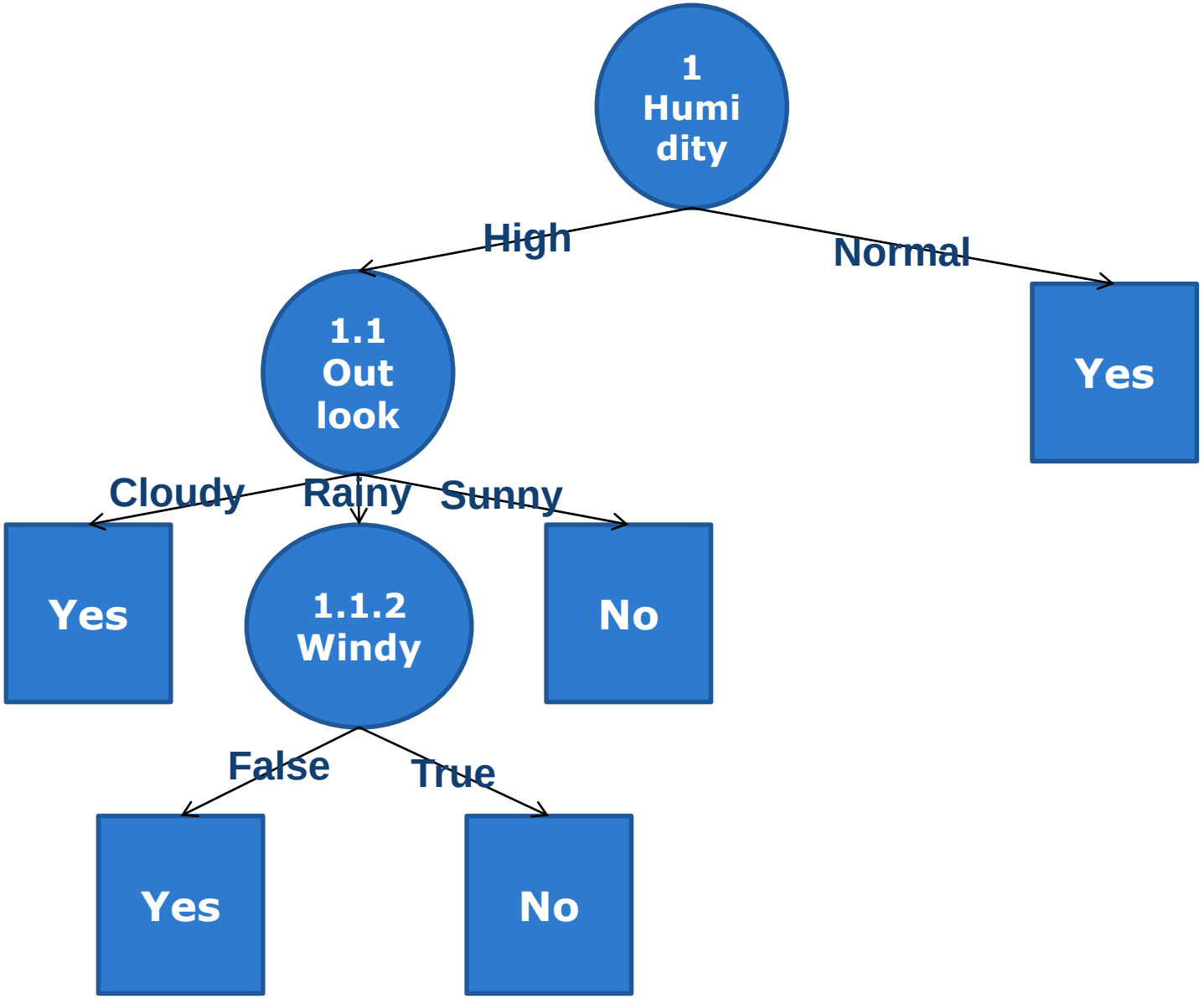
Perhitungan Node 1.1.2

Node 1.1.2		jml kasus(S)	Tidak (S1)	Ya (S2)	Entropy	Gain
Humidity High and Outlook Rainy		2	1	1	1	
temp						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	

Cara Perhitungan Node 1.1.2 (1)

- Atribut dengan Gain tertinggi adalah WINDY, yaitu sebesar 1.
- Sehingga WINDY dapat menjadi node cabang dari nilai atribut RAINY.
- Ada dua nilai dari atribut WINDY, yaitu FALSE dan TRUE.
  - Nilai atribut FALSE sudah mengklasifikasikan kasus menjadi 1 (**Yes**).
  - Nilai atribut TRUE sudah mengklasifikasikan kasus menjadi 1 (**No**).
  - Sehingga tidak perlu dilakukan perhitungan lagi.

Cara Perhitungan Node 1.1.2 (2)



Latihan

- Hitung Entropy dan Gain serta tentukan pohon keputusan yang terbentuk dari contoh kasus keputusan bermain tenis dibawah ini :

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	No	Don't Play
Sunny	Hot	High	Yes	Don't Play
Cloudy	Hot	High	No	Play
Rainy	Mild	High	No	Play
Rainy	Cool	Normal	No	Play
Rainy	Cool	Normal	Yes	Play
Cloudy	Cool	Normal	Yes	Play
Sunny	Mild	High	No	Don't Play
Sunny	Cool	Normal	No	Play
Rainy	Mild	Normal	No	Play
Sunny	Mild	Normal	Yes	Play
Cloudy	Mild	High	Yes	Play
Cloudy	Hot	Normal	No	Play
Rainy	Mild	High	Yes	Don't Play

Referensi

1. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 3rd Edition, *Elsevier*, 2011
2. Kusrini, Taufiq Emha, Algoritma Data Mining, *Penerbit Andi*, 2009


Thank You!