DATA MINING
TUGAS PERTEMUAN KE – 7

Dosen Pembimbing

JUNTA ZENIARJA M.Kom.

Disusun Oleh :

Luthfi Kamal Ananda (A11.2020.12586) // A11.4411

PROGRAM STUDI S-1 TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO
SEMARANG
2023

Latihan Soal (Kuis)

1. 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

Jawaban

Langkah 1

Dengan menggunakan Algoritma C4.5 Berikut adalah tabel yang didapat setelah menghitung jumlah kasus, jumlah kasus untuk keputusan **Play** dan **Don't Play** berdasarkan atribut **OUTLOOK, TEMPERATURE, HUMIDITY** dan **WINDY**

	Jml Kasus	Tidak (S1)	Ya (S2)	Entropy	Gain
Total	14	4	10		
Outlook					
Cloudy	4	0	4		
Rainy	5	1	4		
Sunny	5	3	2		
Temp					
Cool	4	0	4		
Hot	4	2	2		
Mild	6	2	4		
Humidity					
High	7	4	3		
Normal	7	0	7		
Windy					
No	8	2	6		
Yes	6	2	4		

Perhitungan Node 1 [1]

- 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

Dengan begitu untuk mencari Entropy dapat menggunakan :

- Entropy(Total) = $-\frac{4}{14}\log_2(\frac{4}{14}) + -\frac{10}{14}\log_2(\frac{10}{14}) = 0.863120568566631$
- Entropy(Cloudy) = $-\frac{0}{4}\log_2(\frac{0}{4}) + -\frac{4}{4}\log_2(\frac{4}{4}) = 0$
- Entropy(Raining) = $-\frac{1}{5}\log_2(\frac{1}{5}) + -\frac{4}{5}\log_2(\frac{4}{5}) = 0.7219280948873623$
- dan seterusnya..

Sehingga ditemukan tabel sebagai berikut :

	Jml Kasus	Tidak (S1)	Ya (S2)	Entropy	Gain
Total	14	4	10	0.863120569	
Outlook					
Cloudy	4	0	4	0	
Rainy	5	1	4	0.721928095	
Sunny	5	3	2	0.970950594	
Temp					
Cool	4	0	4	0	
Hot	4	2	2	1	
Mild	6	2	4	0.918295834	
Humidity					
High	7	4	3	0.985228136	
Normal	7	0	7	0	
Windy					
No	8	2	6	0.811278124	
Yes	6	2	4	0.918295834	

Perhitungan Node 1 [2]

- 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

Dengan begitu untuk mencari Gain sebagai berikut :

- Gain(Total, Outlook) = $0.863120569 - ((\frac{4}{14} \times 0) + (\frac{5}{14} \times 0.721928095) + (\frac{5}{14} \times 0.970950594))$
  ≈ 0.258521037

- Gain(Total, Temp) = $0.863120569 - ((\frac{4}{14} \times 0) + (\frac{4}{14} \times 1) + (\frac{6}{14} \times 0.918295834))$
  ≈ 0.183850926

- Gain(Total, Humid) = $0.863120569 - ((\frac{7}{14} \times 0.985228136) + (\frac{7}{14} \times 0))$
  $= 0.370506501$

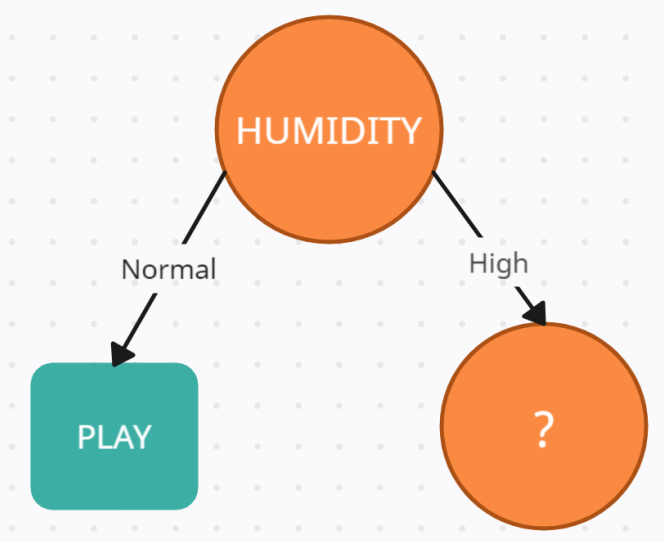
- Gain(Total, Windy) = $0.863120569 - ((\frac{8}{14} \times 0.811278124) + (\frac{6}{14} \times 0.918295834))$
  $= 0.005977712142857139$

Sehingga ditemukan tabel sebagai berikut :

	Jml Kasus	Tidak (S1)	Ya (S2)	Entropy	Gain
Total	14	4	10	0.863120569	
Outlook					0.258521037
Cloudy	4	0	4	0	
Rainy	5	1	4	0.721928095	
Sunny	5	3	2	0.970950594	
Temp					0.183850926
Cool	4	0	4	0	
Hot	4	2	2	1	
Mild	6	2	4	0.918295834	
Humidity					0.370506501
High	7	4	3	0.985228136	
Normal	7	0	7	0	
Windy					0.005977712
No	8	2	6	0.811278124	
Yes	6	2	4	0.918295834	

Perhitungan Node 1 [3]

- Dapat dilihat bahwa atribut yang memiliki **Gain** tertinggi adalah **HUMIDITY** sehingga **HUMIDITY** dapat menjadi node akar
- Terdapat 2 nilai atribut pada **HUMIDITY** yaitu **HIGH** dan **NORMAL**
- Nilai atribut **NORMAL** sudah mengklasifikasikan kasus menjadi 1 yaitu hanya memiliki keputusan **Play**, sehingga tidak perlu dilakukan perhitungan lebih lanjut
- Tetapi **HIGH** masih perlu dilakukan perhitungan lagi



Perhitungan Node 1.1

- Setelah itu, melakukan perhitungan **Gain** lagi pada tiap – tiap atribut yang dapat menjadi node akar dari nilai **HUMIDITY = HIGH**

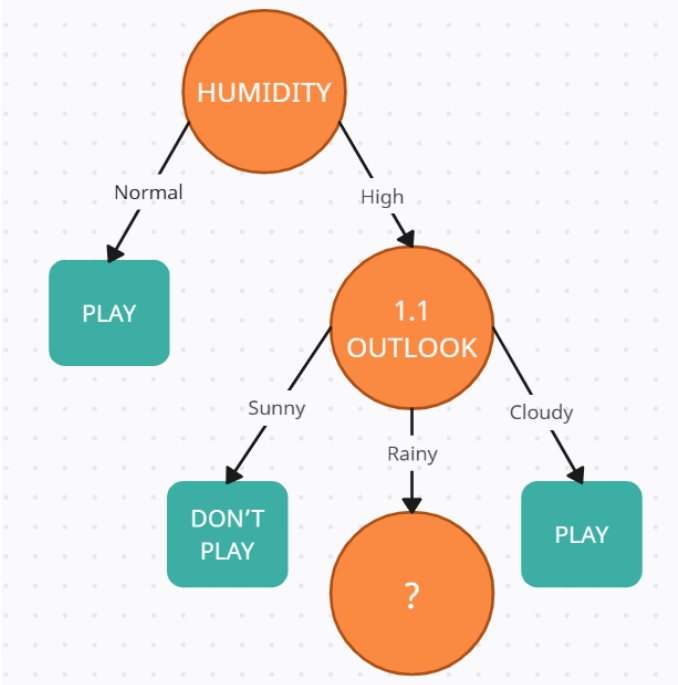
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	No	Don't Play
Sunny	Hot	High	Yes	Don't Play
Cloudy	Hot	High	No	Play
Rainy	Mild	High	No	Play
Rainy	Cool	Normal	No	Play
Rainy	Cool	Normal	Yes	Play
Cloudy	Cool	Normal	Yes	Play
Sunny	Mild	High	No	Don't Play
Sunny	Cool	Normal	No	Play
Rainy	Mild	Normal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

- Sehingga dilakukan lagi perhitungan dengan data yang memiliki **HUMIDITY = HIGH** dengan atribut **OUTLOOK, TEMPERATURE** dan **WINDY**

- Dengan menggunakan rumus sebelumnya untuk mencari **Entropy** dan **Gain** dengan data yang memiliki **HUMIDITY = HIGH** dapat ditemukan tabel baru sebagai berikut

	Jml Kasus	Tidak (S1)	Ya (S2)	Entropy	Gain
Humidity = HIGH	7	4	3		
Outlook					0.69951385
Cloudy	2	0	2	0	
Rainy	2	1	1	1	
Sunny	3	3	0	0	
Temp					0.02024421
Cool	0	0	0	0	
Hot	3	2	1	0.91829583	
Mild	4	2	2	1	
Windy					0.02024421
No	4	2	2	1	
Yes	3	2	1	0.91829583	

- Dapat dilihat bahwa atribut **OUTLOOK** adalah atribut yang memiliki **GAIN** tertinggi sebesar 0.69951385
- Sehingga **OUTLOOK** dapat dijadikan menjadi node cabang dari nilai atribut **HIGH**
- Ada tiga nilai dari atribut **OUTLOOK** yaitu :
    - **CLODY** ➔ Klasifikasi Kasus 1 (**PLAY**)
    - **SUNNY** ➔ Klasifikasi Kasus 2 (**DON'T PLAY**)
    - **RAINY** ➔ Perlu perhitungan lagi



Perhitungan Node 1.2

- Setelah itu lakukan perhitungan **GAIN** lagi pada tiap – tiap atribut yang dapat menjadi node akar dari nilai **HUMIDITY = HIGH DAN OUTLOOK = RAINY**

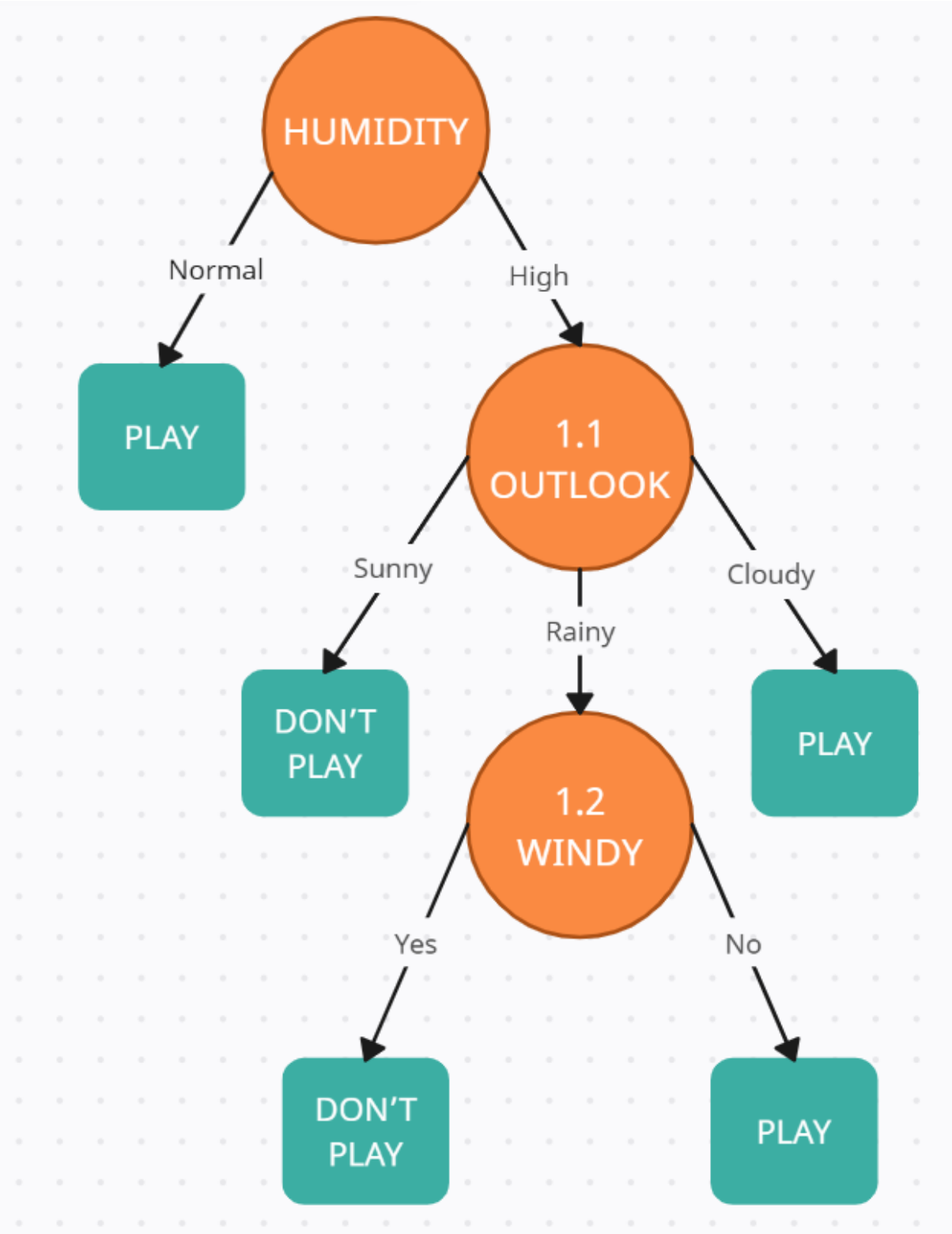
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	No	Don't Play
Sunny	Hot	High	Yes	Don't Play
Cloudy	Hot	High	No	Play
Rainy	Mild	High	No	Play
Rainy	Cool	Normal	No	Play
Rainy	Cool	Normal	Yes	Play
Cloudy	Cool	Normal	Yes	Play
Sunny	Mild	High	No	Don't Play
Sunny	Cool	Normal	No	Play
Rainy	Mild	Normal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

- Sehingga dilakukan lagi perhitungan dengan data yang memiliki **HUMIDITY = HIGH DAN OUTLOOK = RAINY** dengan atribut **TEMPERATURE** dan **WINDY**
- Dengan menggunakan rumus sebelumnya untuk mencari **Entropy** dan **Gain** dengan data yang memiliki **HUMIDITY = HIGH** dan **OUTLOOK = RAINY** dapat ditemukan tabel baru sebagai berikut

	Jml Kasus	Tidak (S1)	Ya (S2)	Entropy	Gain
Humidity = HIGH & 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
No	1	0	1	0	
Yes	1	1	0	0	

- Atribut dengan **GAIN** tertinggi adalah **WINDY** sebesar 1
- Sehingga **WINDY** dapat menjadi node cabang dari nilai atribut **RAINY**
- Ada dua nilai dari atribut **WINDY** yaitu **NO** dan **YES**
    - Nilai **NO** sudah mengklasifikasikan kasus menjadi 1 (**PLAY)**
    - Nilai **YES** sudah mengklasifikasikan kasus menjadi 1(**DON'T PLAY)**
    - Sehingga tidak perlu dilakukan perhitungan lagi

- Sehingga berikut ini adalah decision tree akhir



Latihan Soal (Kuis)

2. Kerjakan latihan tahapan klasifikasi dengan Decision Tree pada latihan sebelumnya, dataset bisa diganti / dimodifikasi, simpan dalam ***decisiontree.py*** atau ***decisiontree.ipynb***, repositorikan file pada **github.com** dan kirimkan URL github melalui Assignment pada kulino (Pada blok Minggu ke-7).

Jawaban :

LINK GITHUB : https://github.com/luthfikamalananda/Decision-Tree