Nama : Nila Farihah

NIM : A11.2022.14667

Kelompok : A11.4509

PERTEMUAN 7

Link repositori pertemuan 7 : https://github.com/Nilapine/data-mining/tree/main/tugas7

Latihan Soal (Kuis)

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

NO	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	Sunny	Hot	High	No	Don't Play
2	Sunny	Hot	High	Yes	Don't Play
3	Cloudy	Hot	High	No	Play
4	Rainy	Mild	High	No	Play
5	Rainy	Cool	Normal	No	Play
6	Rainy	Cool	Normal	Yes	Play
7	Cloudy	Cool	Normal	Yes	Play
8	Sunny	Mild	High	No	Don't Play
9	Sunny	Cool	Normal	No	Play
10	Rainy	Mild	Normal	No	Play
11	Sunny	Mild	Normal	Yes	Play
12	Cloudy	Mild	High	Yes	Play
13	Cloudy	Hot	Normal	No	Play
14	Rainy	Mild	High	Yes	Don't Play

Jawab :

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

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

Perhitungan Node 1

Node 1		jml kasus(S)	Don't Play (S1)	Play (S2)	Entropy	Gain
total		14	4	10	0,86312057	
outlook						0,25852104
	cloudy	4	0	4	0	
	rainy	5	1	4	0,72192809	
	sunny	5	3	2	0,97095059	
temp						0,18385093
	cool	4	0	4	0	
	hot	4	2	2	1	
	mild	6	2	4	0,91829583	
humidity						0,3705065
	high	7	4	3	0,98522814	
	normal	7	0	7	0	
windy						0,005977711
	No	8	2	6	0,811278124	
	Yes	6	4	2	0,91829583	

Cara Perhitungan Node 1 :

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

$$= 0{,}863120569$$

$$\text{Gain}(Total, Outlook) = \text{Entropy}(Total) - \sum_{i=1}^{n} \frac{|Outlook_i|}{|Total|} * \text{Entropy}(Outlook_i)$$

$$= 0.863120569 - \left(\left(\frac{4}{14} * 0\right) + \left(\frac{5}{14} * 0.722\right) + \left(\frac{5}{14} * 0.97\right)\right)$$

$$= 0.2585$$

$$\text{Gain}(Total, Temp) = \text{Entropy}(Total) - \sum_{i=1}^{n} \frac{|Temp_i|}{|Total|} * \text{Entropy}(Temp_i)$$

$$= 0.863120569 - \left(\left(\frac{4}{14} * 0\right) + \left(\frac{4}{14} * 1\right) + \left(\frac{6}{14} * 0.92\right)\right)$$

$$= 0.1839$$

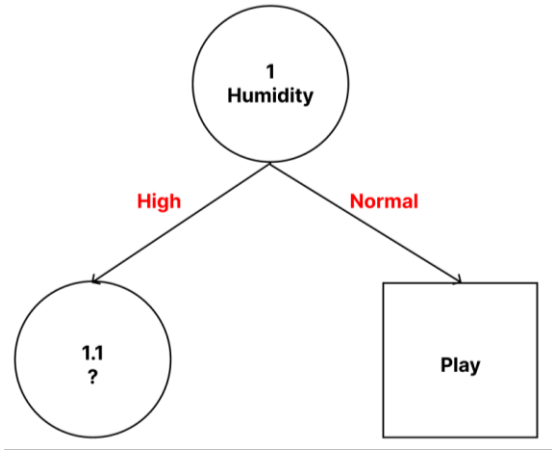
$$\text{Gain}(Total, Humidity) = \text{Entropy}(Total) - \sum_{i=1}^{n} \frac{|Humidity_i|}{|Total|} * \text{Entropy}(Humidity_i)$$

$$= 0.863120569 - \left(\left(\frac{7}{14} * 0.918\right) + \left(\frac{7}{14} * 0\right)\right) = 0.3705$$

$$\text{Gain}(Total, Windy) = \text{Entropy}(Total) - \sum_{i=1}^{n} \frac{|Windy_i|}{|Total|} * \text{Entropy}(Windy_i)$$

$$= 0.863120569 - \left(\left(\frac{8}{14} * 0.811\right) + \left(\frac{6}{14} * 0.918\right)\right) = 0.0059$$

Hasil perhitungan di atas diketahui bahwa atribut dengan gain tertinggi adalah **humidity** dengan nilai 0.3705. Sehingga, **humidity** menjadi node akar dari pohon keputusan. **humidity** memiliki dua nilai atribut, yaitu HIGH dan NORMAL. Untuk cabang dengan nilai atribut NORMAL, sudah bisa langsung diklasifikasikan sebagai keputusan "Yes" (Play), karena hasil perhitungan menunjukkan bahwa semua instance dengan **humidity** NORMAL mengarah pada keputusan yang sama. Namun, untuk cabang dengan nilai atribut HIGH, diperlukan perhitungan lebih lanjut untuk menentukan keputusan akhir. Hal ini menunjukkan bahwa data dengan **humidity** HIGH masih memiliki variabilitas yang memerlukan eksplorasi lebih mendalam sebelum dapat diklasifikasikan dengan pasti.



Perhitungan untuk Node 1.1 (Jika humidity = High)

Menghitung Gain untuk atribut lain yang tersisa: OUTLOOK, TEMPERATURE, dan WINDY, berdasarkan subset data di mana humidity = High.

Node 1.1		jml kasus(S)	Don't Play (S1)	Play (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
	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	

Cara perhitungan Node 1.1 :

$$\text{Entropy}(S) = \left(-\frac{3}{7} \cdot \log_2\left(\frac{3}{7}\right)\right) + \left(-\frac{4}{7} \cdot \log_2\left(\frac{4}{7}\right)\right) = 0{,}9852$$

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

$$= 0.98522814 - \left(\left(\frac{2}{7} * 0\right) + \left(\frac{2}{7} * 1\right) + \left(\frac{3}{7} * 0\right)\right) = 0.6995$$

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

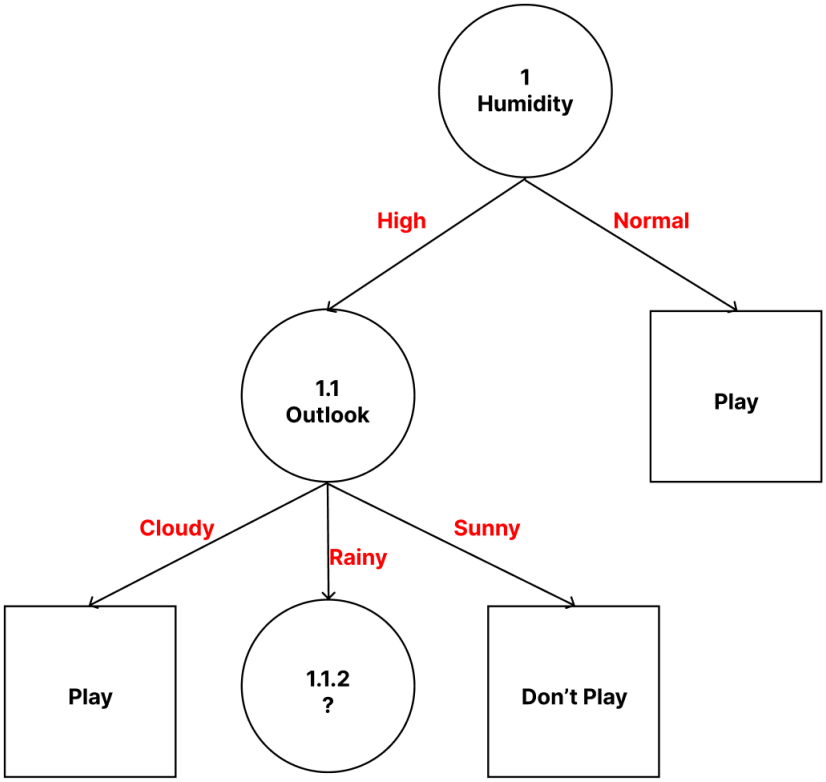
$$= 0.98522814 - \left(\left(\frac{0}{7} * 0\right) + \left(\frac{3}{7} * 0.9182\right) + \left(\frac{4}{7} * 1\right)\right) = 0.0202$$

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

$$= 0.98522814 - \left(\left(\frac{4}{7} * 1\right) + \left(\frac{3}{7} * 0.9182\right)\right) = 0.0202$$

OUTLOOK dipilih sebagai node cabang dari nilai humidity = High karena Gain tertingginya sebesar 0.6995. OUTLOOK memiliki tiga nilai:

- CLOUDY: Semua kasus di cabang ini diklasifikasikan sebagai Play.
- SUNNY: Semua kasus di cabang ini diklasifikasikan sebagai Don't Play.
- RAINY: Masih ada variasi antara Play dan Don't Play, sehingga perlu dilakukan perhitungan lebih lanjut untuk atribut lain agar bisa mengklasifikasikan kasus di cabang ini.



Perhitungan untuk Node 1.1.2 (Jika humidity = High, OUTLOOK = Rainy)

Node 1.1.2		jml kasus(S)	Don't Play (S1)	Play (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
	No	1	0	1	0	
	Yes	1	1	0	0	

Untuk detail perhitungannya sebagai berikut:

$$\text{Entropy}(S) = \left(-\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right) + \left(-\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right) = 1$$

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

$$= 1 - \left(\left(\frac{0}{2} * 0\right) + \left(\frac{3}{0} * 0\right) + \left(\frac{2}{2} * 1\right)\right) = 0$$

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

$$= 1 - \left(\left(\frac{1}{2} * 0\right) + \left(\frac{1}{2} * 0\right)\right) = 1$$

Atribut WINDY dengan Gain tertinggi (1) menjadi node untuk OUTLOOK = RAINY. Jika WINDY = No, semua kasus adalah Play, dan jika WINDY = Yes, maka semua kasus adalah Don't Play. Karena kedua nilai sudah mengklasifikasikan kasus, maka tidak perlu perhitungan lebih lanjut.