CSE422: Artificial intelligence

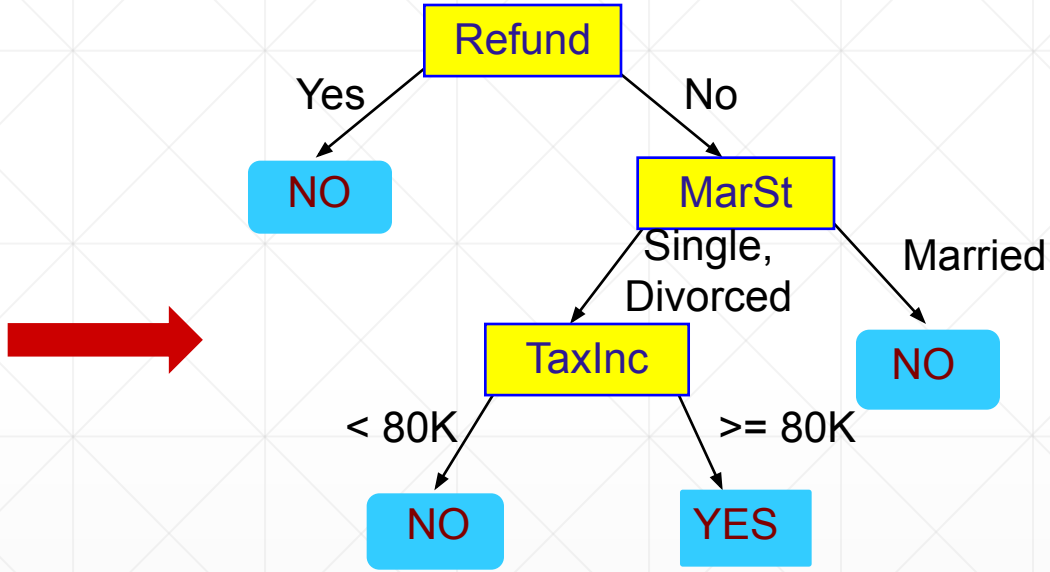
Decision Tree

Asif Shahriar
Lecturer, CSE, BRACU

Example of a Decision Tree

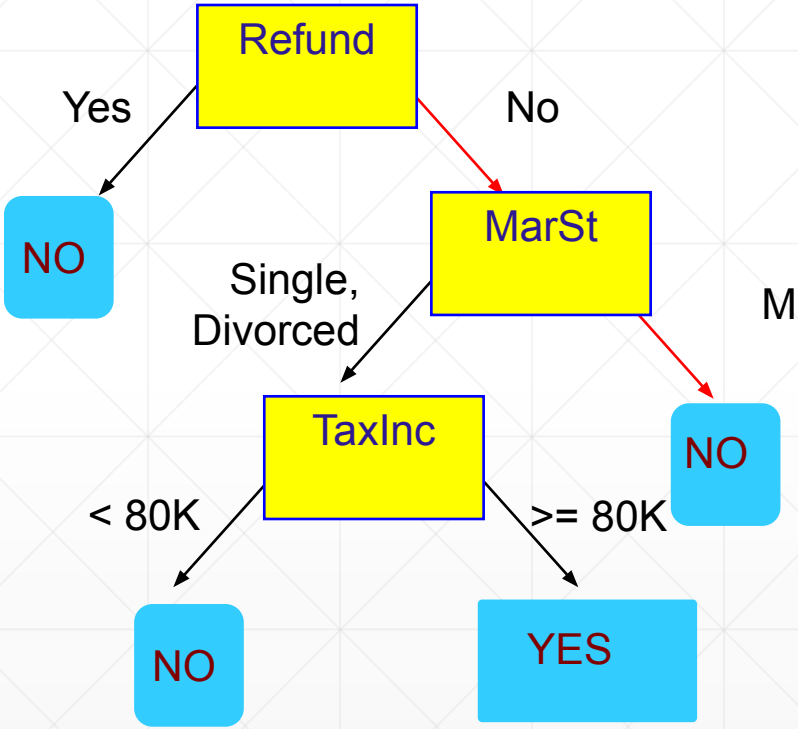
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Example of a Decision Tree



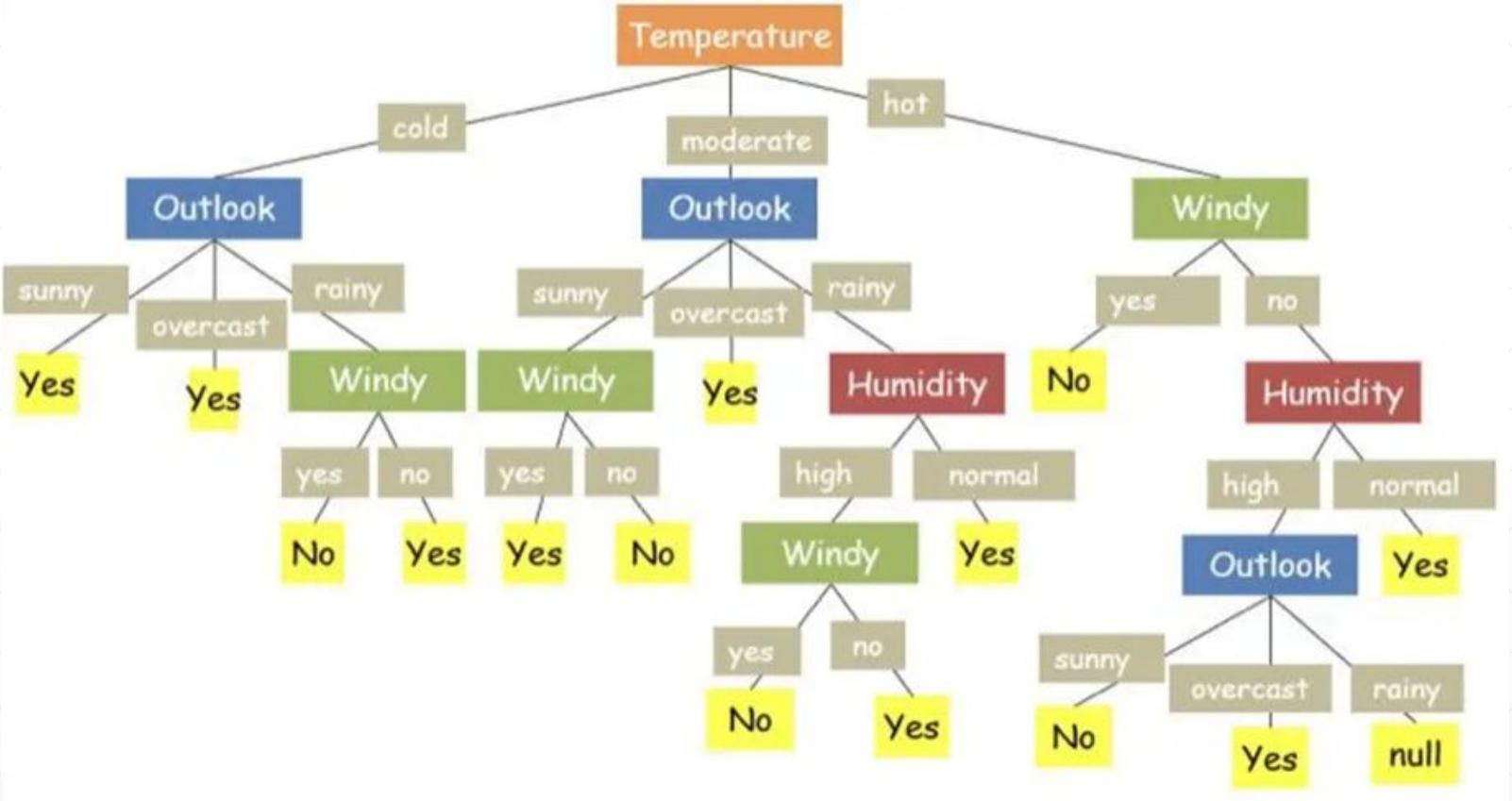
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

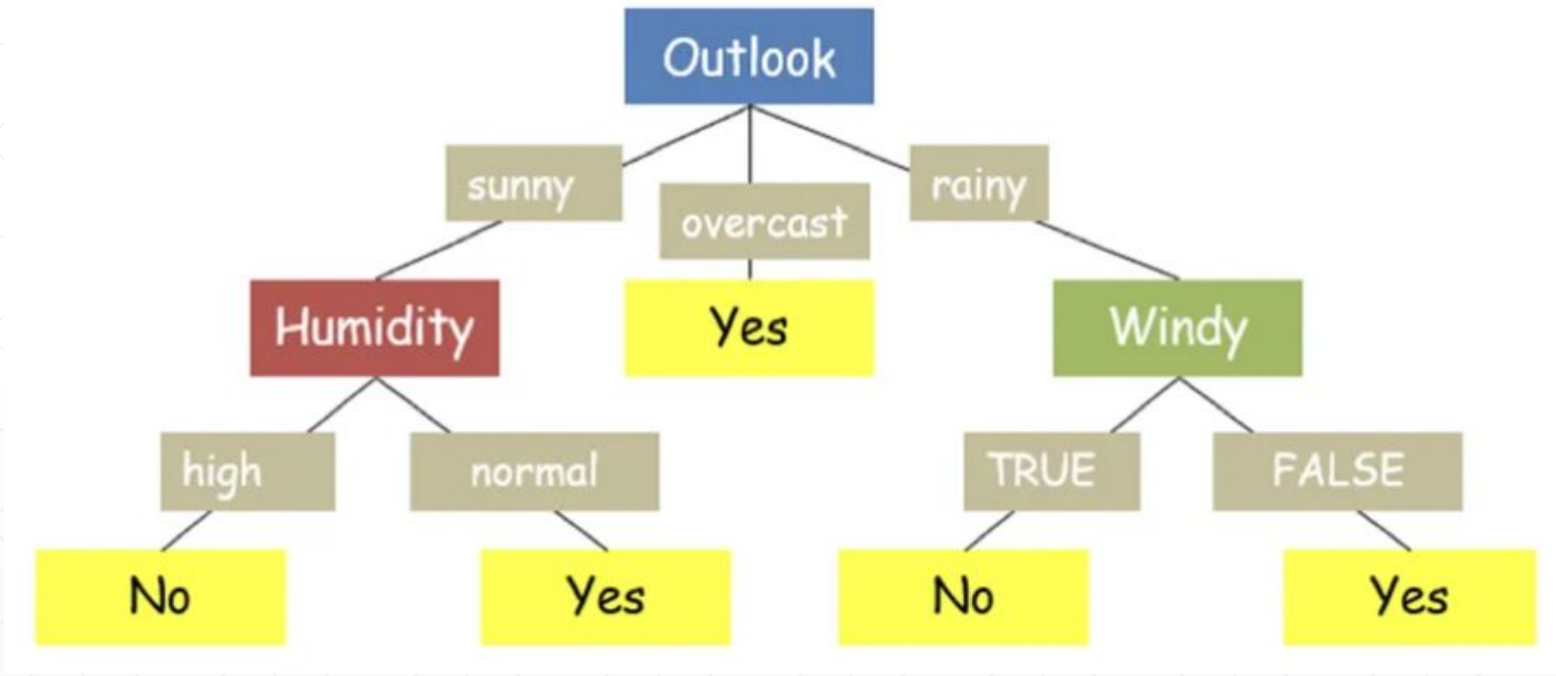
How to Build a Decision Tree

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

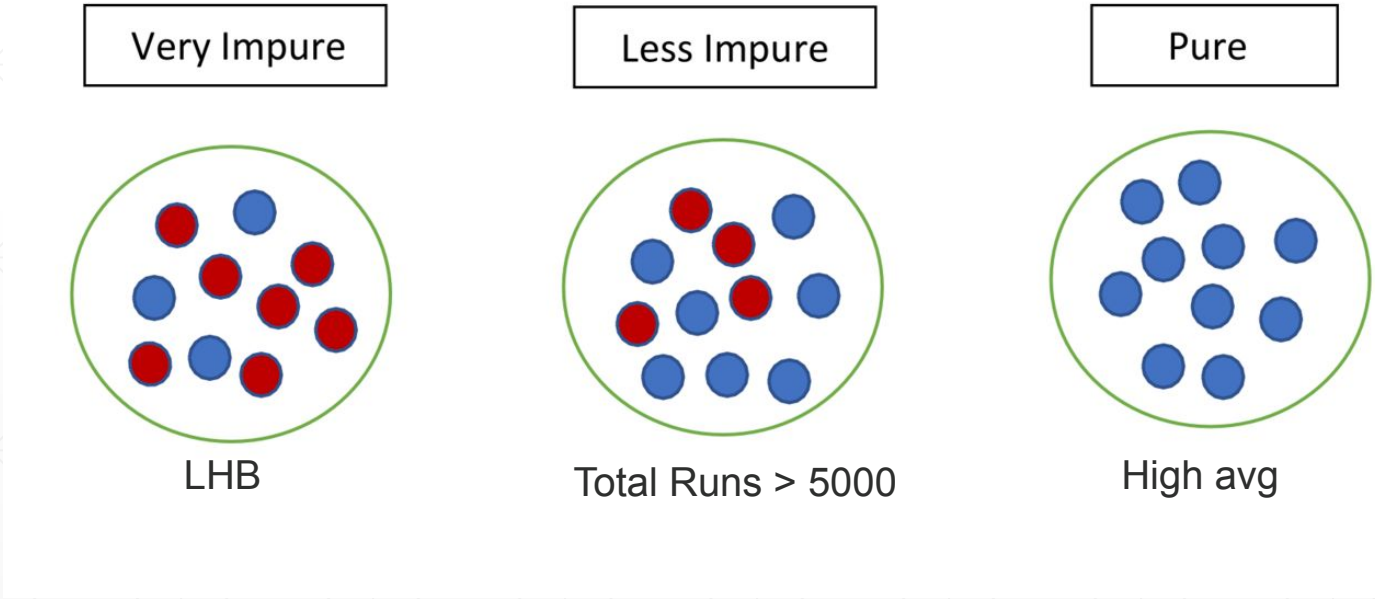
Decision Tree 1



Decision Tree 2



Which Feature is Better?



Entropy

- Consider a Feature **S** with N classes: $\{c_1, c_2, \dots, c_N\}$
- Each class has an associated probability: $\{p_1, p_2, \dots, p_N\}$
- Entropy of the feature **S,** $$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$
- Consider a dataset of 100 batsmen. Consider the feather “batting hand”, where 55 are right handed and rest are left handed

$$H(batting\_hand) = -0.55 \log_2 0.55 \; - 0.45 \log_2 0.45 = 0.99$$

- Consider their average, where 15 has high average (>45), rest low average (<35)

$$H(average) = 0.61$$

- Lower entropy -> higher discrimination -> easier to decide -> better feature

How to Choose a Feature?

- Let, output is S (boolean), there are THREE features A, B, C
- Calculate entropy of S, H(S)
- Feature A has N classes: $a_1, a_2, \ldots, a_N$
- Calculate Remainder of A: $Remainder\ (A) = \sum_{i=1}^{N} weight(a_i) \times H(a_i)$
- Information gain of A for S, ***Gain(S, A) = H(S) - Remainder(A)***
- Similarly calculate Gain(S, B), Gain(S, C), Gain(S, D)
- Choose the feature with **highest gain**

How to Build a Decision Tree

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Here output is **Decision**
- There are four features
- We need to choose a feature first that will make our decision easier (reduce the tree size)
- Calculate H(decision)
- Calculate Gain(Decision, Outlook), Gain(Decision, Temp), Gain(Decision, Humidity), Gain(Decision, Wind)
- Choose the feature w/ most gain

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$H(decision) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$Gain(Decision, Wind) = H(Decision) - Remainder(Wind)$$

$$Remainder(Wind) = weight(strong) \times H(strong) + weight(weak) \times H(weak)$$

$$H(strong) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 0.94$$

$$Weight(strong) = \frac{6}{14}$$

$$H(weak) = -\frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

$$Weight(weak) = \frac{8}{14}$$

$$Remainder(Wind) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.811 = 0.892$$

$$Gain(Decision, Wind) = 0.94 - 0.892 = 0.048$$

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$Gain(Decision, Wind) = 0.048$$

$$Gain(Decision, Outlook) = 0.246$$

$$Gain(Decision, Temp) = 0.029$$

$$Gain(Decision, Humidity) = 0.151$$

Choose OUTLOOK as root of Tree

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Note: If Outlook = Overcast, Decision is always YES
- What if Outlook is Sunny or Rain?
- Now we have to choose a second feature to look at: Temp or Humidity or Wind
- Calculate Gain(Sunny, Temp), Gain(Sunny, Humidity), Gain(Sunny, Wind) -> choose the feature with highest gain
- Calculate Gain(Rain, Temp), Gain(Rain, Humidity), Gain(Rain, Wind) -> choose the feature with highest gain

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Let's calculate Gain(Sunny, Temp)

$$H(sunny) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$$

$$H(Hot) = -1\log_2 1 - 0 = 0$$

$$H(Cool) = -0 - 1\log_2 1 = 0$$

$$H(Mild) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Remainder(Temp) = \frac{2}{5} \times 1 + 0 + 0 = 0.4$$

$$Gain(Sunny, Temp) = 0.97 - 0.4 = 0.57$$

How to Build a Decision Tree: ID3 Algorithm

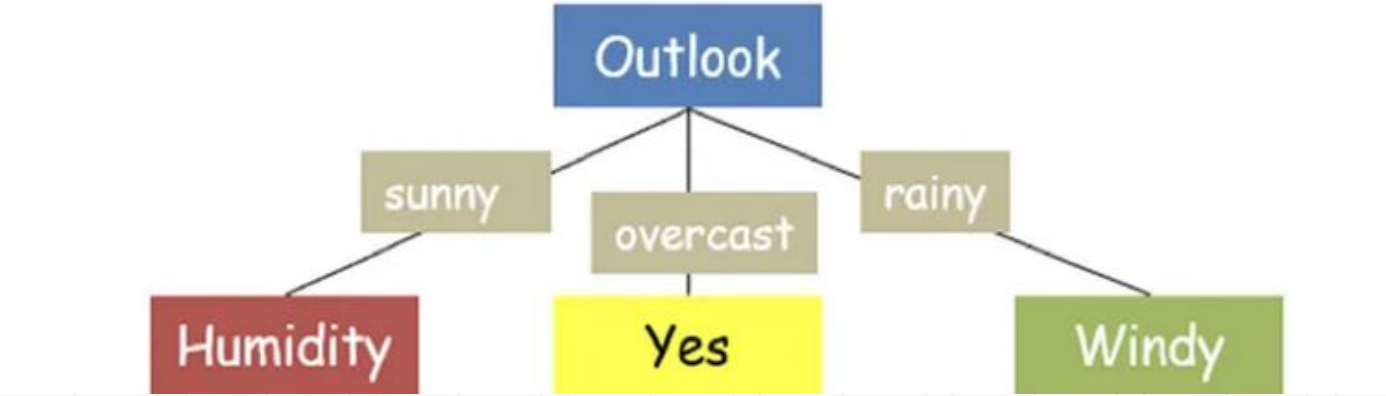
Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

- Gain(Sunny, Temp) = 0.57
- Gain(Sunny, Humidity) = 0.97
- Gain(Sunny, Wind) = 0.019
- Choose Humidity for sunny

How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Similarly, Calculate Gain(Rain, Temp), Gain(Rain, Humidity), Gain(Rain, Wind)
- Gain(Rain, Wind) is highest, choose Wind for Rain



How to Build a Decision Tree: ID3 Algorithm

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Notice that, if Outlook = Sunny and Humidity = High, Decision is always NO
- Similarly if Outlook = Sunny and Humidity = Normal, Decision is always Yes
- Moreover, if Outlook = Rain and Wind = Weak, Decision is always Yes
- If Outlook = Rain and Wind = Strong, Decision is always No
- At this point, the tree construction is over

Final Decision Tree