Decision Trees

An Interpretable ML Algorithm

What is it?

- A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).
- It is one way to display an algorithm that only contains conditional control statements.

Types of Nodes on a Decision Tree

Decision/Root Nodes (□)

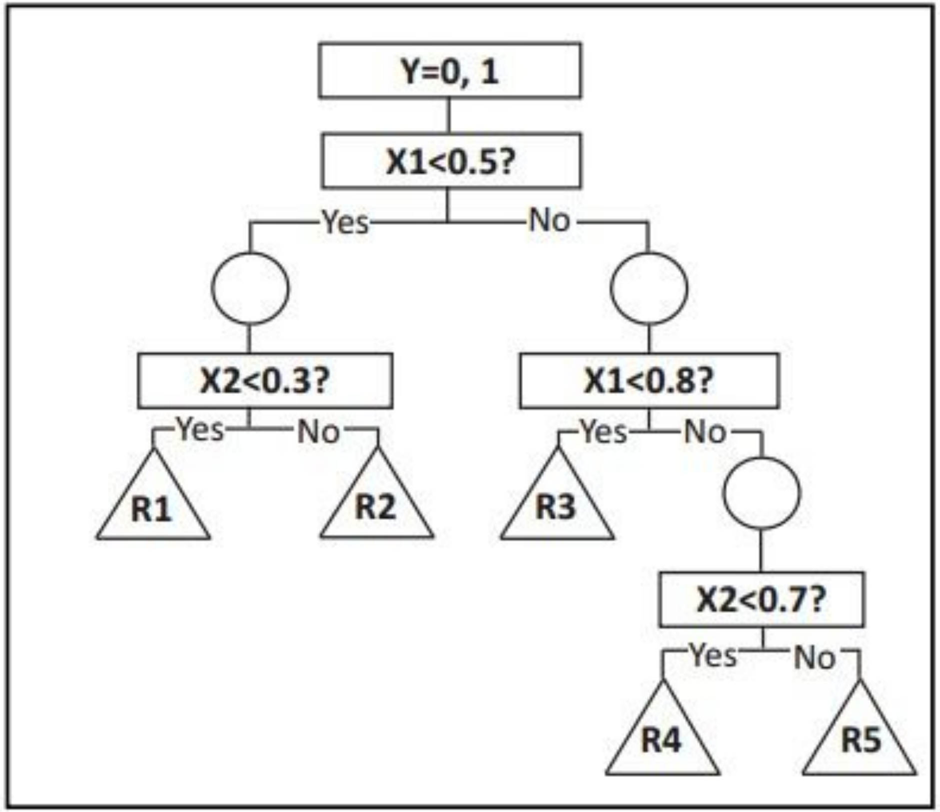
Indicates a decision to be made

Chance/Internal Nodes (○)

Multiple uncertain outcomes

End/Leaf Nodes (△)

Indicates Final Outcome



Decision/Root Nodes (□)

Indicates a decision to be made

Chance/Internal Nodes (○)

Multiple uncertain outcomes

End/Leaf Nodes (△)

Indicates Final Outcome

How to represent a D Tree ?

Two ways

1. Tree View
2. Sample Space View(later on this)

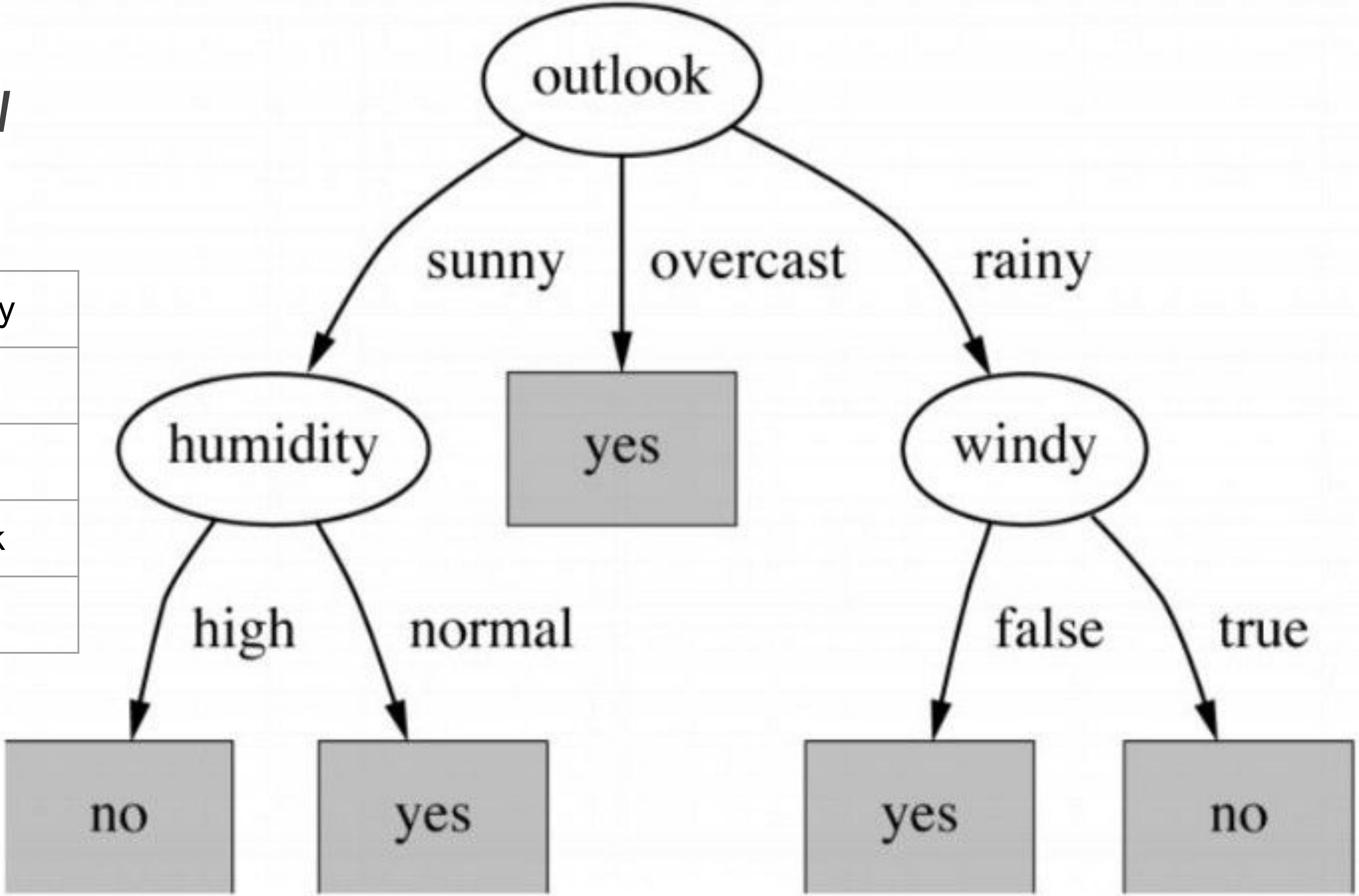
How to interpret an already built Decision Tree?

Day	Outlook	Temp	Humidity	Windy	Play ?
D1	Sunny	Hot	High	FALSE	NO
D2	Sunny	Hot	High	TRUE	NO
D3	Overcast	Hot	High	FALSE	YES
D4	Rainy	Mild	High	FALSE	YES
D5	Rainy	Cool	Normal	FALSE	YES
D6	Rainy	Cool	Normal	TRUE	NO
D7	Overcast	Cool	Normal	TRUE	YES
D8	Sunny	Mild	High	FALSE	NO
D9	Sunny	Cool	Normal	FALSE	YES
D10	Rainy	Mild	Normal	FALSE	YES
D11	Sunny	Mild	Normal	TRUE	YES
D12	Overcast	Mild	High	TRUE	YES
D13	Overcast	Hot	Normal	FALSE	YES
D14	Rainy	Mild	High	TRUE	NO

Weather Data

Tree View

Outlook	Sunny
Temp	Hot
Humidity	High
Wind	Weak
Play?	NO



How to Build a Decision Tree from Data?

On what attribute should the split be made

1. ID3
2. CART

ID3 Algorithm

- Uses Entropy and Information gain as metric to decide the split
- Entropy : Measure of amount of uncertainty in data
- Information Gain : Difference between Entropy before and after the split

Entropy

$$H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

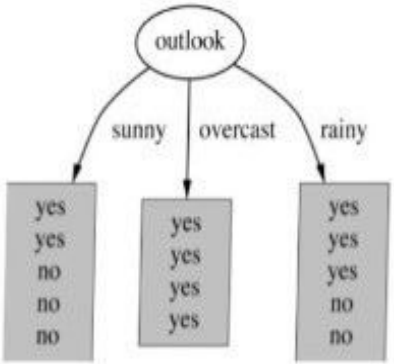
C ={yes,no}

Out of 14 instances, 9 are classified as yes, and 5 as no

pyes = −(9/14)*log2(9/14) = 0.41

pno = −(5/14)*log2(5/14) = 0.53

H (S) = pyes + pno = 0.94



E (Outlook=sunny) = $-\frac{2}{5}\log\left(\frac{2}{5}\right)-\frac{3}{5}\log\left(\frac{3}{5}\right)=0.971$

E (Outlook=overcast) = $-1\log(1)-0\log(0)=0$

E (Outlook=rainy) = $-\frac{3}{5}\log\left(\frac{3}{5}\right)-\frac{2}{5}\log\left(\frac{2}{5}\right)=0.971$

H(S,Outlook)

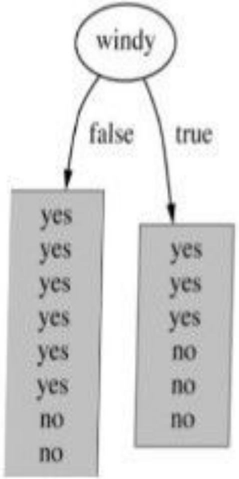
Average Entropy information for Outlook

I (Outlook) = $\frac{5}{14}*0.971+\frac{4}{14}*0+\frac{5}{14}*0.971=0.693$

$$\sum_{t\in T} p(t)H(t)$$

Gain (Outlook) = E(S) – I (outlook) = 0.94-.693 = 0.247

$$IG(A,S)=H(S)-\sum_{t\in T} p(t)H(t)$$



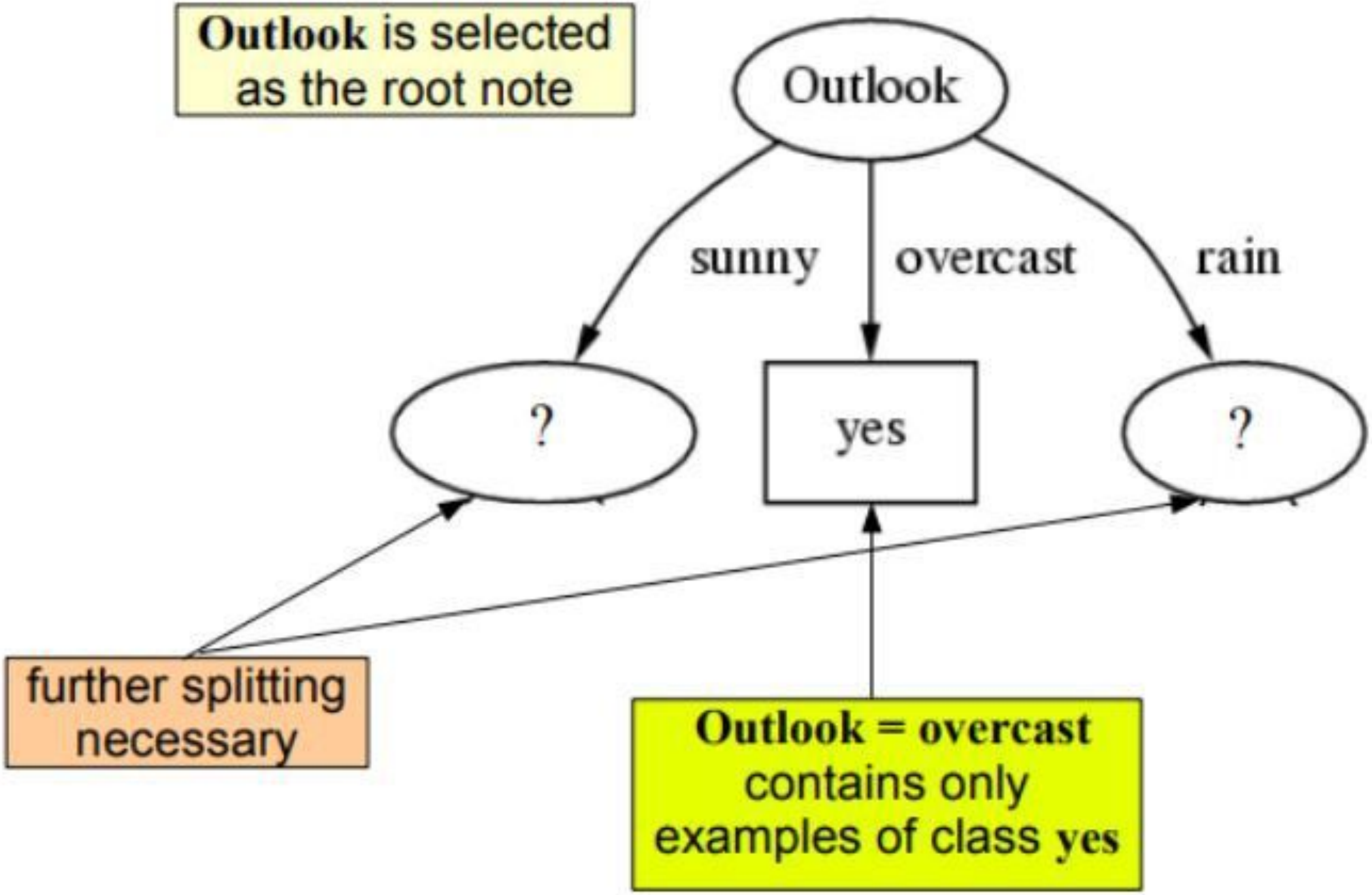
E (Windy=false) = $-\frac{6}{8}\log\left(\frac{6}{8}\right)-\frac{2}{8}\log\left(\frac{2}{8}\right)=0.811$

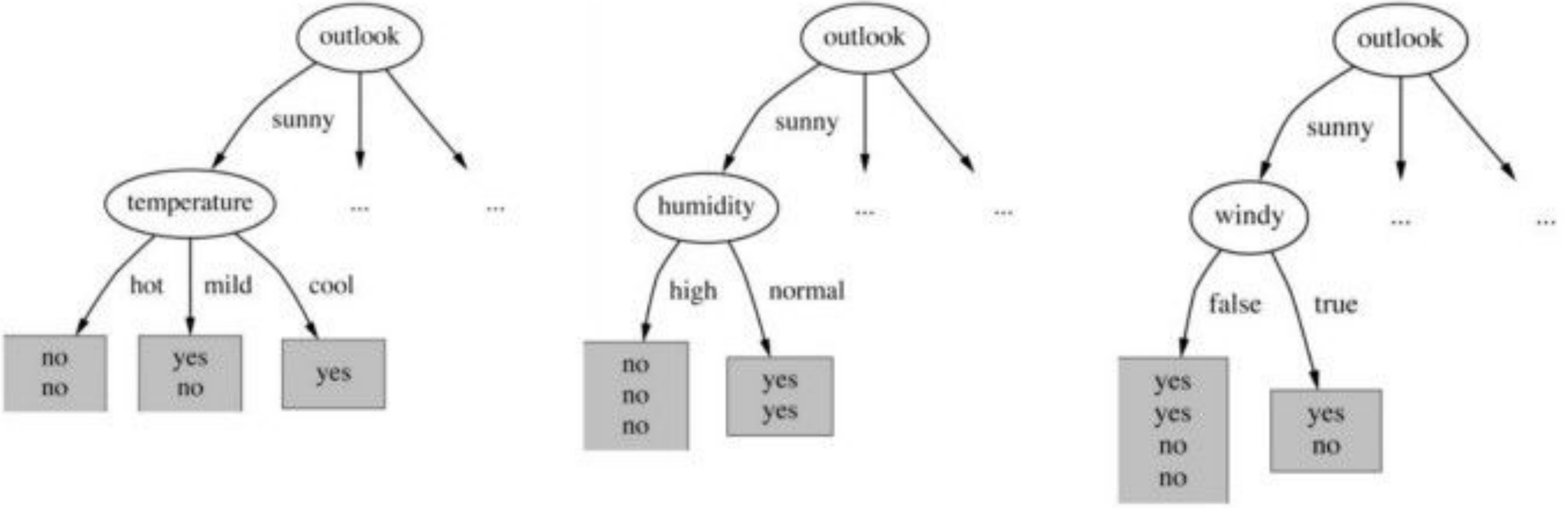
E (Windy=true) = $-\frac{3}{6}\log\left(\frac{3}{6}\right)-\frac{3}{6}\log\left(\frac{3}{6}\right)=1$

Average entropy information for Windy

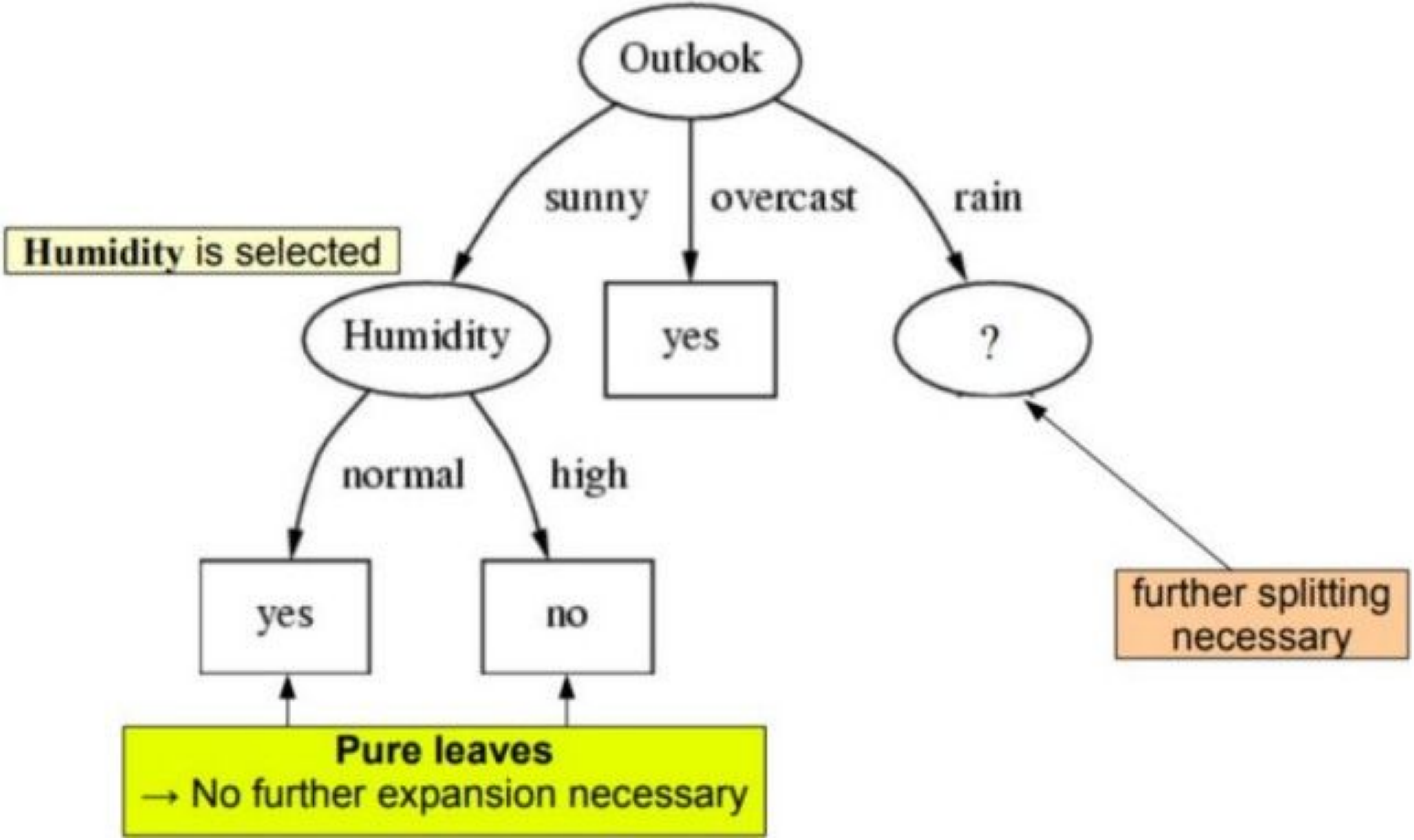
I (Windy) = $\frac{8}{14}*0.811+\frac{6}{14}*1=0.892$

Gain (Windy) = E(S) – I (Windy) = 0.94-0.892=0.048


Outlook is selected as the root note
Outlook
sunny
overcast
rain
?
yes
?
further splitting necessary
Outlook = overcast contains only examples of class **yes**



$$\left.\begin{aligned} \text{Gain}(\textit{Temperature}) &= 0.571 \text{ bits}\\ \text{Gain}(\textit{Humidity}) &= 0.971 \text{ bits}\\ \text{Gain}(\textit{Windy}) &= 0.020 \text{ bits} \end{aligned}\right\} \boxed{\textbf{Humidity}\text{ is selected}}$$



Outlook
sunny
overcast
rain
Humidity is selected
Humidity
yes
?
normal
high
yes
no
further splitting necessary
Pure leaves
→ No further expansion necessary