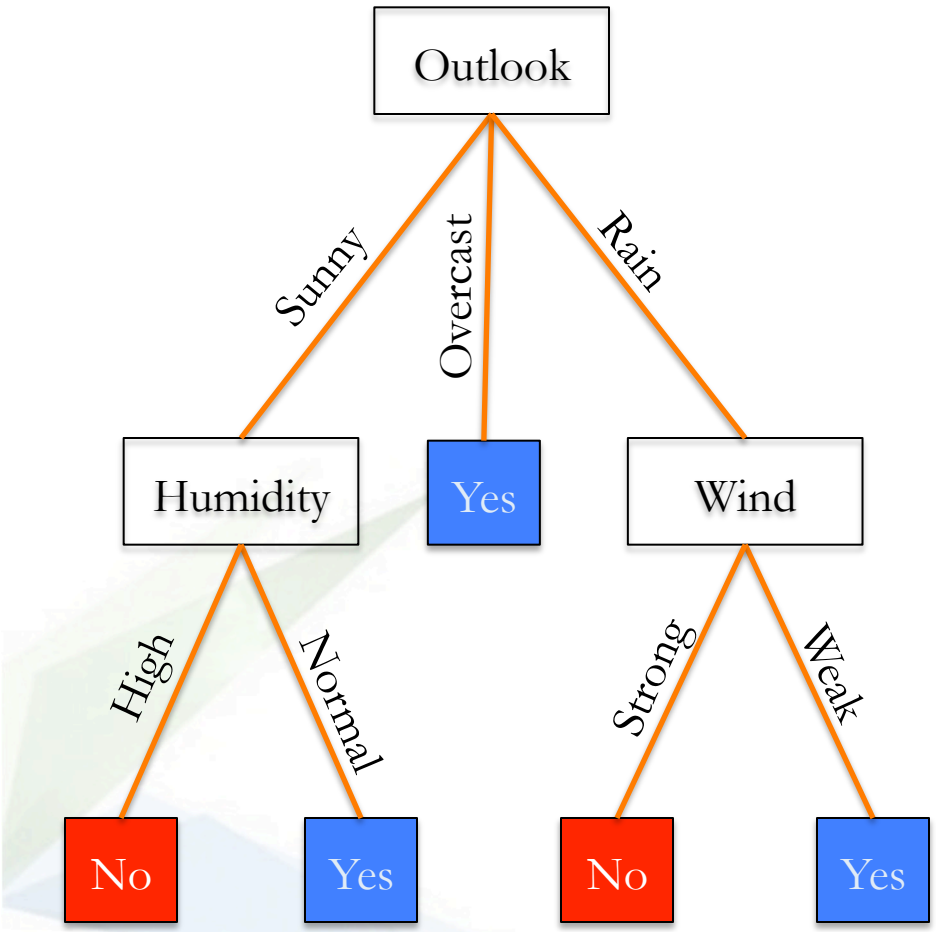
Decision Trees

Will Nadal Play Tennis?

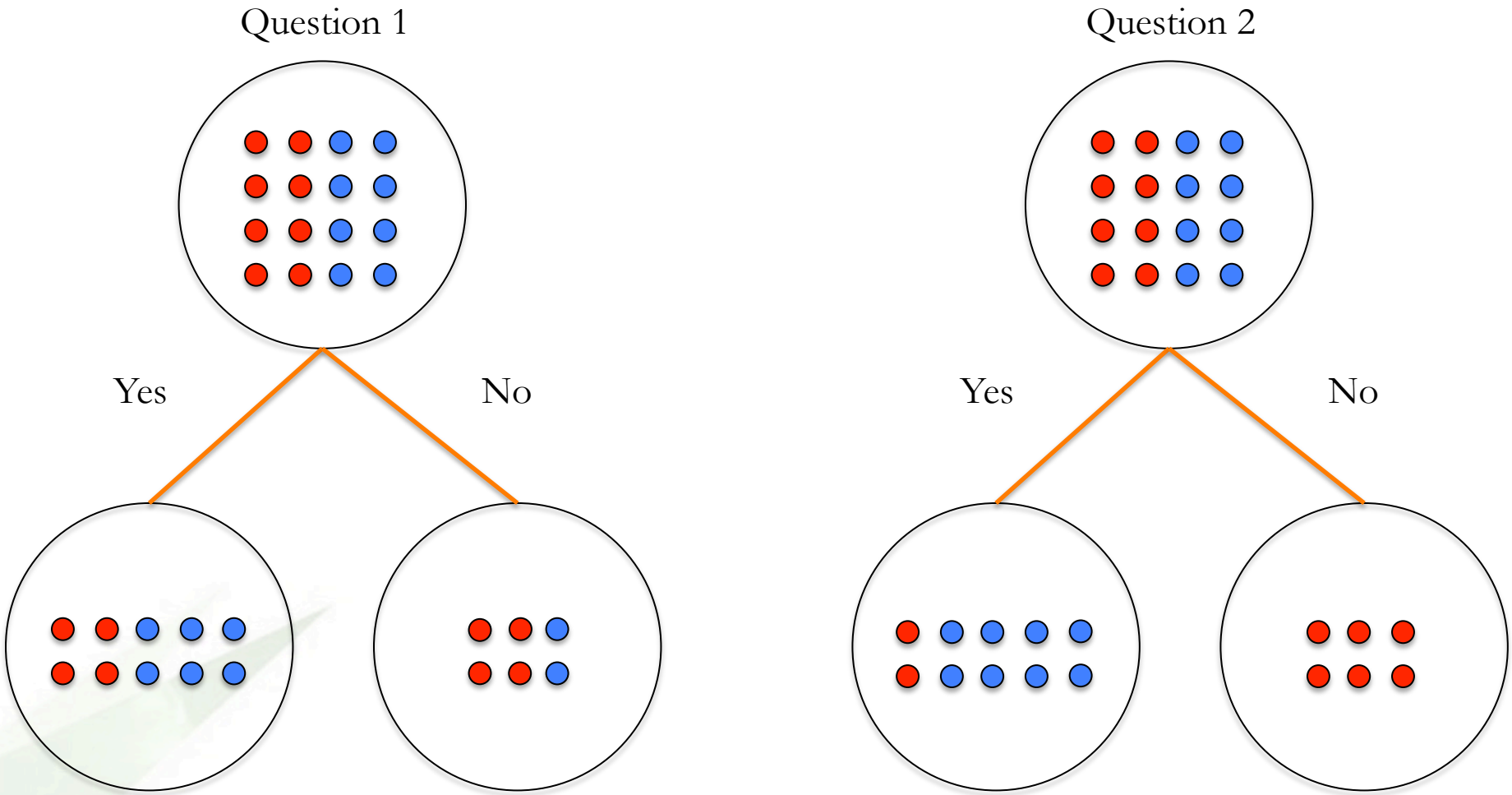
Rafael Nadal

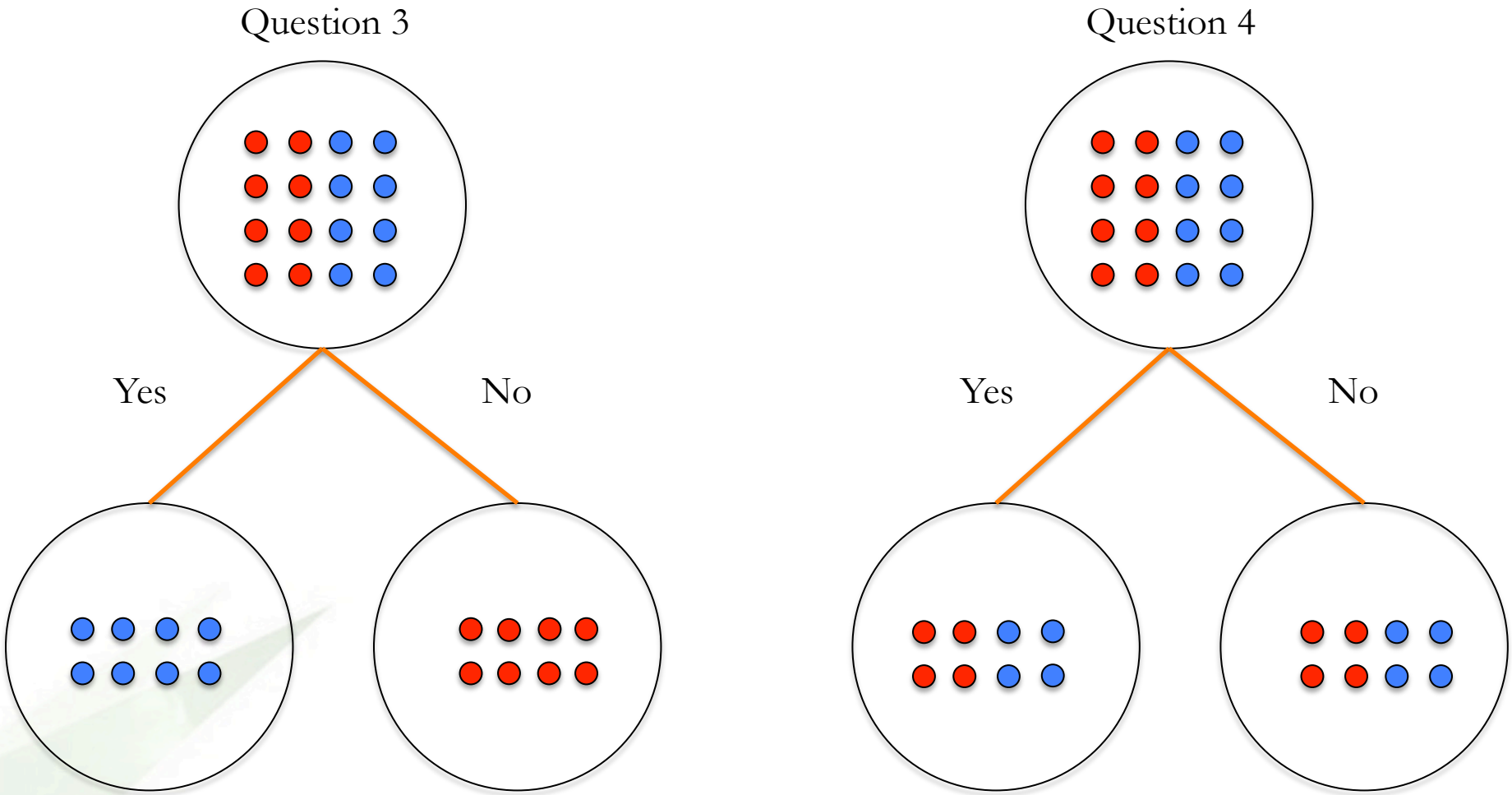
Day	Outlook	Temp	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	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

Will Nadal Play Tennis?

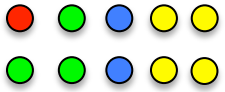


Day	Outlook	Temp	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	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


Question 1
Yes
No
Question 2
Yes
No



$$E(S) = -p_+ log(p_+) - p_- log(p_-)$$

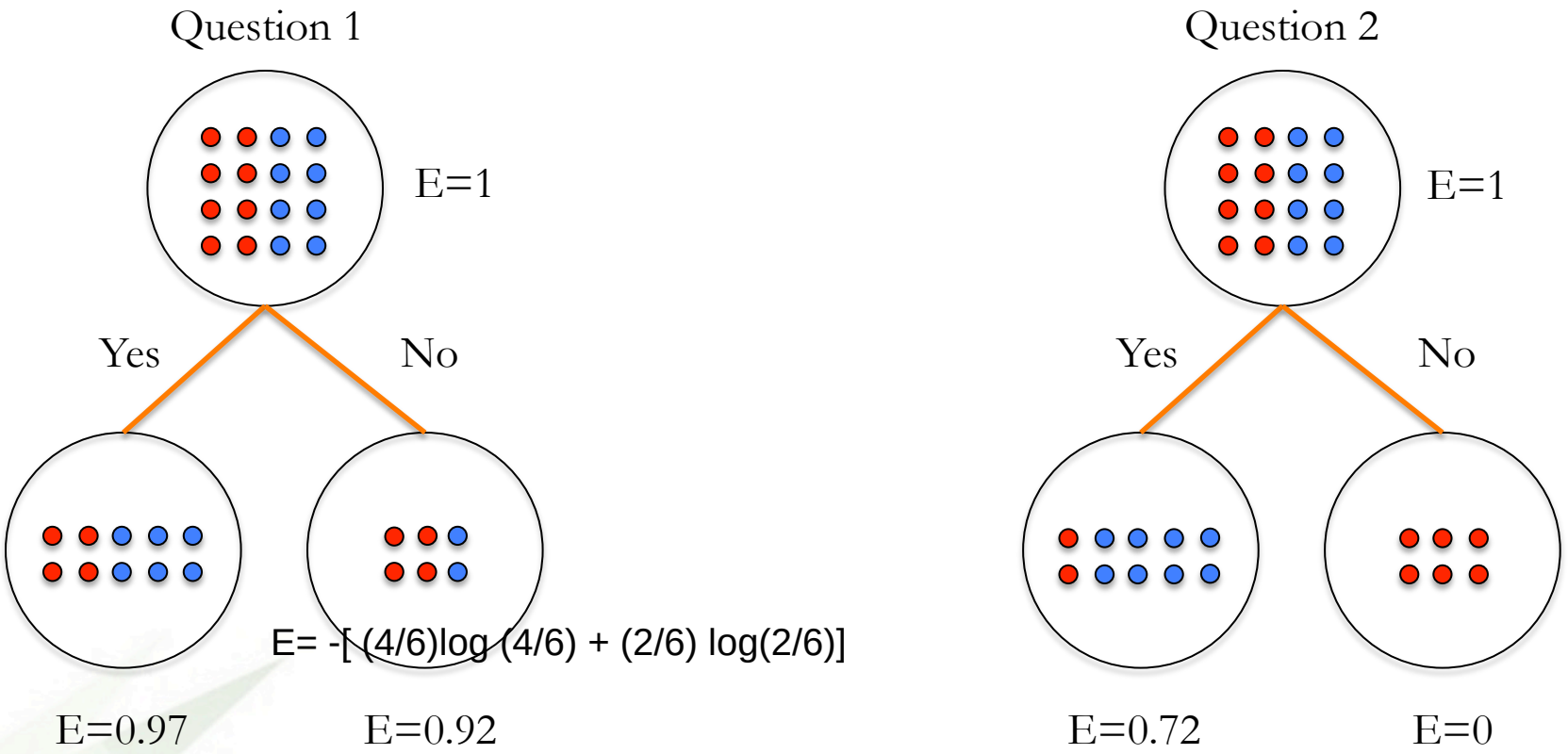


$$y = -\sum_{i=1}^{k} p_i log_k(p_i)$$

$$y = -\underbrace{\left[\left(\frac{1}{10}\right) log_4 \left(\frac{1}{10}\right)\right]}_{\text{Red}} - \underbrace{\left[\left(\frac{3}{10}\right) log_4 \left(\frac{3}{10}\right)\right]}_{\text{Green}} - \underbrace{\left[\left(\frac{2}{10}\right) log_4 \left(\frac{2}{10}\right)\right]}_{\text{Blue}} - \underbrace{\left[\left(\frac{4}{10}\right) log_4 \left(\frac{4}{10}\right)\right]}_{\text{Yellow}}$$

الإنتروبيا: هي مقياس لعدم اليقين أو الفوضى داخل مجموعة بيانات. لنفترض أن لدينا مجموعة بيانات تتعلق بالطقس، وتحتوي على 100 يوم، نصفها مشمس والنصف الآخر غائم. إذا لم نكن نعلم حالة الطقس لأي يوم، فإن الإنتروبيا ستكون عالية لأن هناك عدم يقين كبير.

مكسب المعلومات: هو تقليل الإنتروبيا أو عدم اليقين الذي نحصل عليه عندما نقسم مجموعة البيانات بناءً على سمة معينة. باستخدام مثال الطقس، إذا استخدمنا سمة "درجة الحرارة" لتقسيم الأيام المشمسة والغائمة، ووجدنا أن الأيام المشمسة دائمًا دافئة والأيام الغائمة دائمًا باردة، فإن هذا التقسيم يقلل من عدم اليقين ويزيد من مكسب المعلومات، مما يجعل "درجة الحرارة" سمة فعالة في تصنيف الطقس.

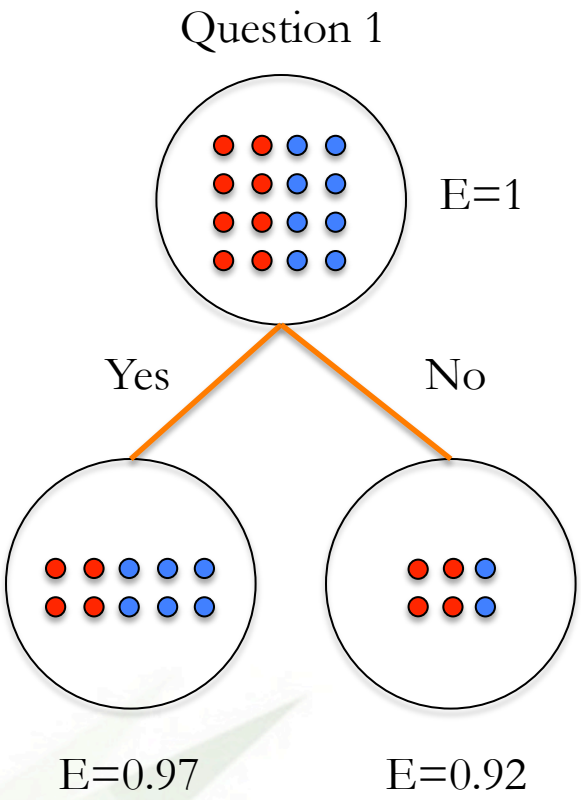


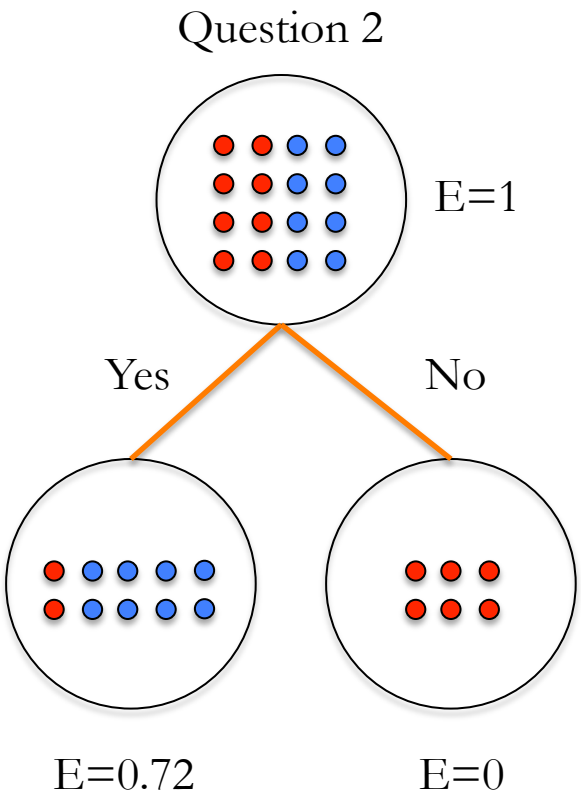
E= -[(4/6)log (4/6) + (2/6) log(2/6)]

E= - [(4/10)log(4/10) + (6/10)log (6/10)]

Information Gain

$$G(S,Q) = E(S) - \sum_{i=1}^{k} p_i E(S,Q_i)$$





$$G(S, Q_1) = 1 - \left(\frac{10}{16}\right) \times 0.97 - \left(\frac{6}{16}\right) \times 0.92$$

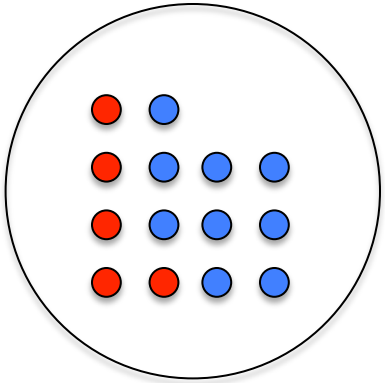
$$G(S, Q_1) = 0.049$$

$$G(S, Q_2) = 1 - \left(\frac{10}{16}\right) \times 0.72 - \left(\frac{6}{16}\right) \times 0$$

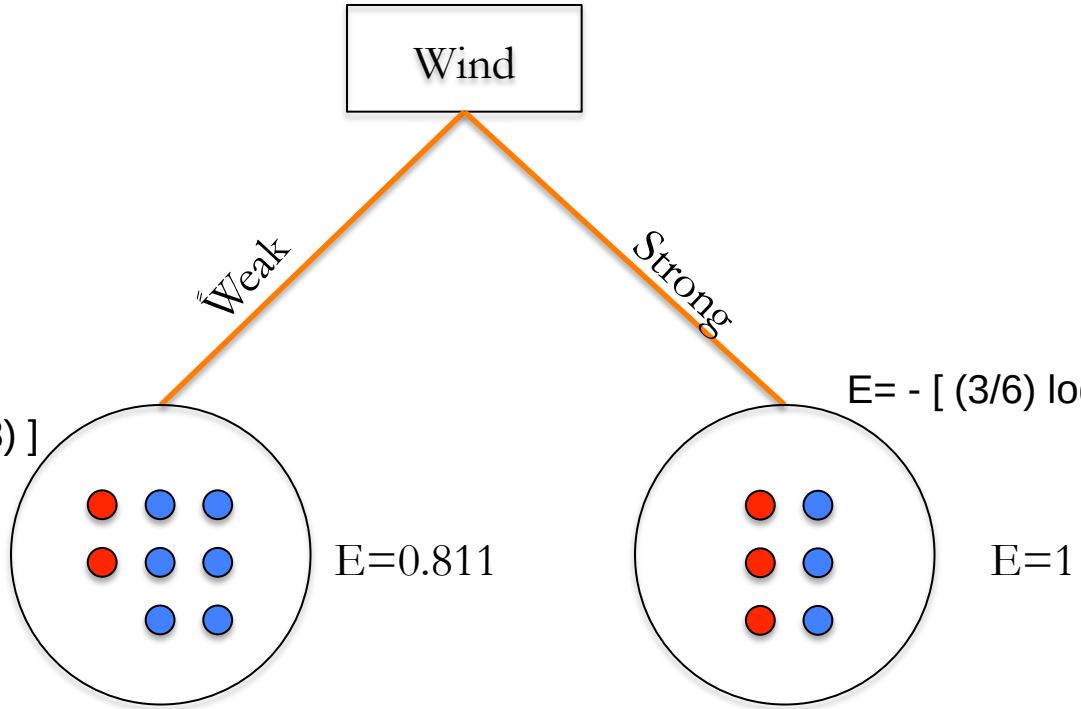
$$G(S, Q_2) = 0.55$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Strong	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

E= - [(5/14) log (5/14) + (9/14)log(9/14)]



E=0.954



E=- [(2/8)log(2/8) + (6/8)log(6/8)]

E=0.811

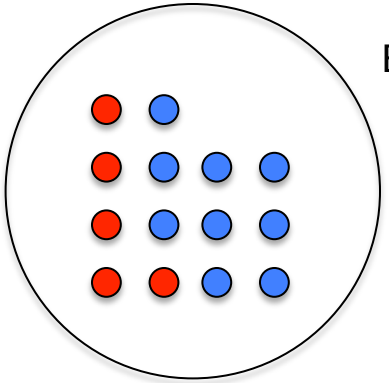
E= - [(3/6) log(3/6) + (3/6) log(3/6)]

E=1

$$G(S, Wind) = 0.954 - \frac{8}{14}0.811 - \frac{6}{14}1$$

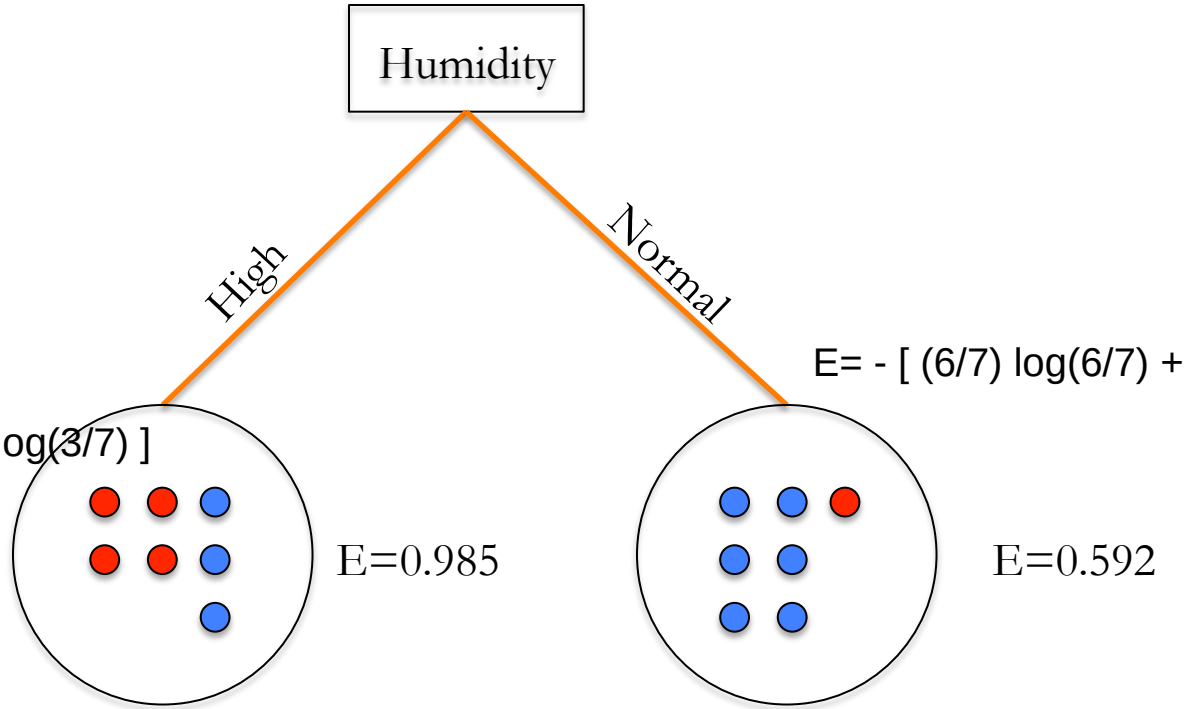
$$G(S, Wind) = 0.048$$

$G(S, Wind) = 0.048$



E= - [(5/14)log(5/14) + (9/14)log(9/14)]

E=0.954



E= - [(4/7) log(4/7) + (3/7)log(3/7)]

E=0.985

E= - [(6/7) log(6/7) + (1/7) log(1/7)]

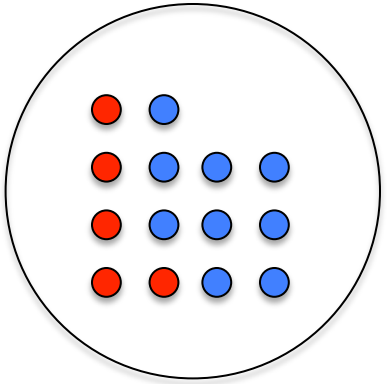
E=0.592

$$G(S, Humidity) = 0.954 - \frac{7}{14}0.985 - \frac{7}{14}0.592$$

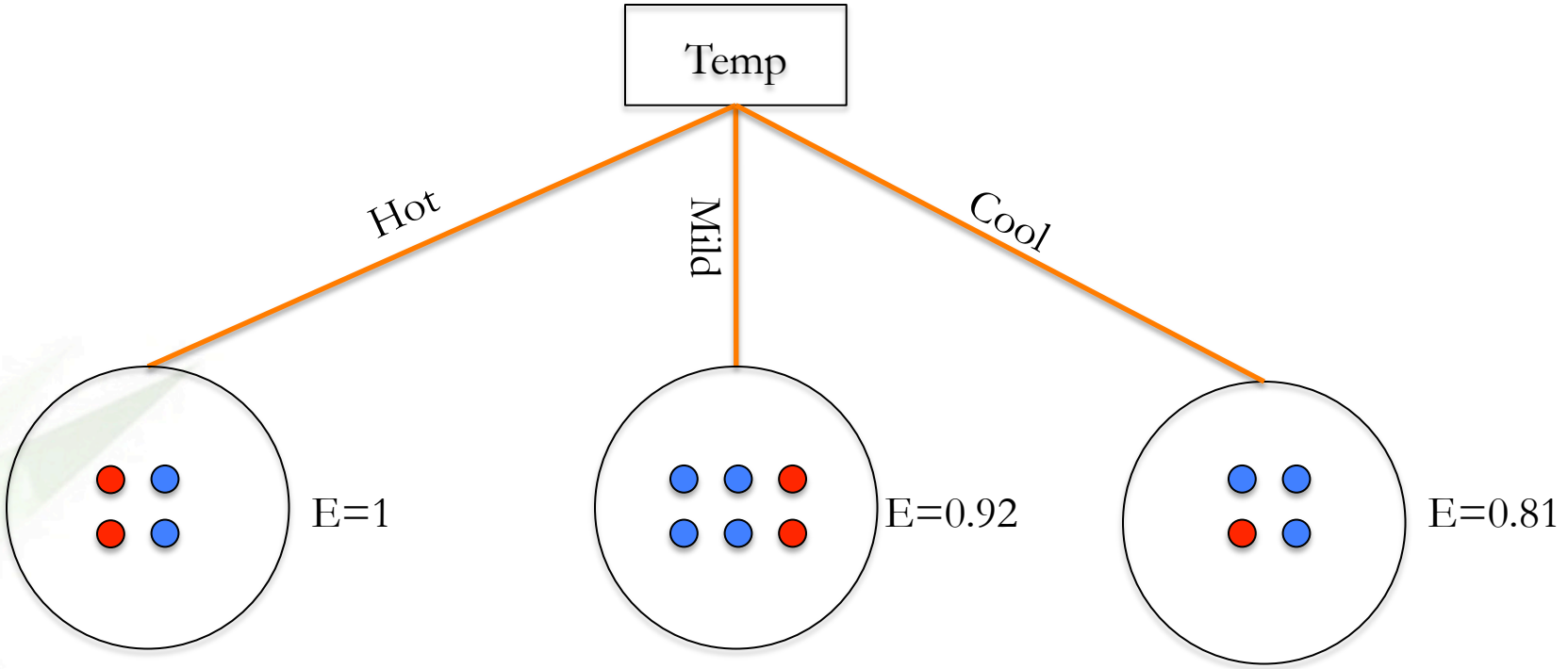
$$G(S, Humidity) = 0.151$$

$G(S, Wind) = 0.048$

$G(S, Humidity) = 0.151$



E=0.954



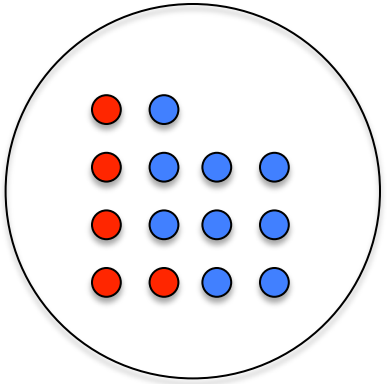
$$G(S, Temp) = 0.954 - \frac{4}{14}1 - \frac{6}{14}0.92 - \frac{4}{14}0.81$$

$$G(S, Temp) = 0.042$$

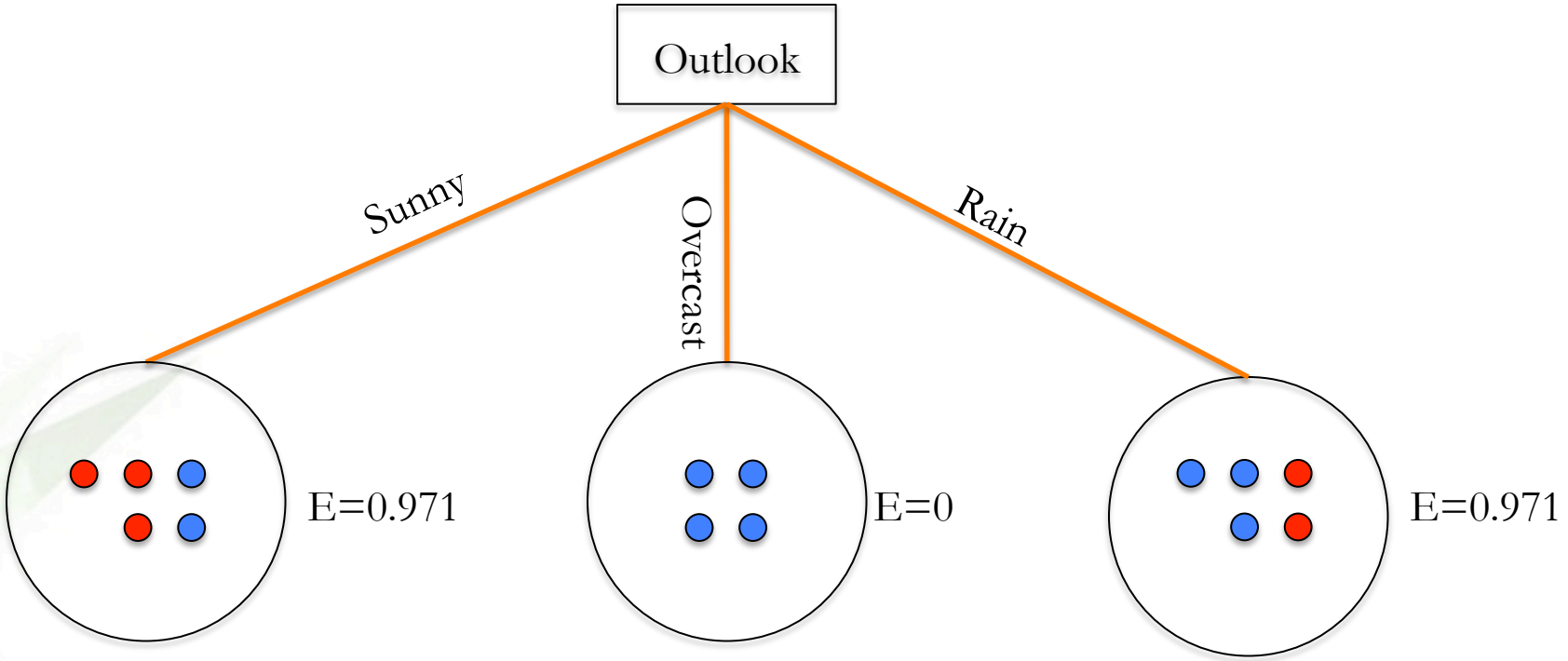
$G(S, Wind) = 0.048$

$G(S, Humidity) = 0.151$

$G(S, Temp) = 0.042$

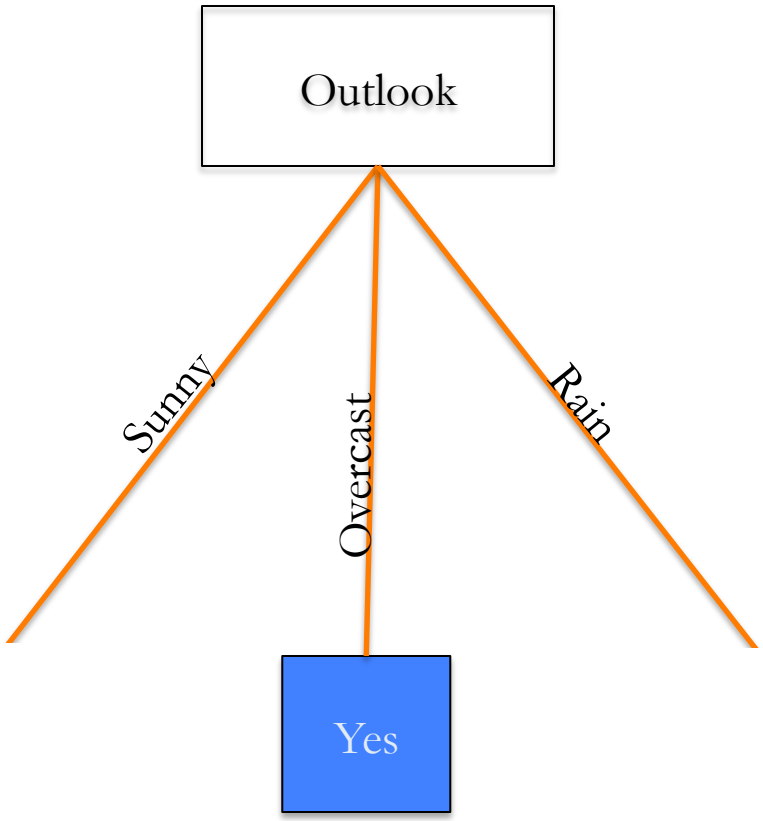


E=0.954



$$G(S, Outlook) = 0.954 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971$$

$$G(S, Outlook) = 0.247$$


Outlook
Sunny
Overcast
Rain
Yes

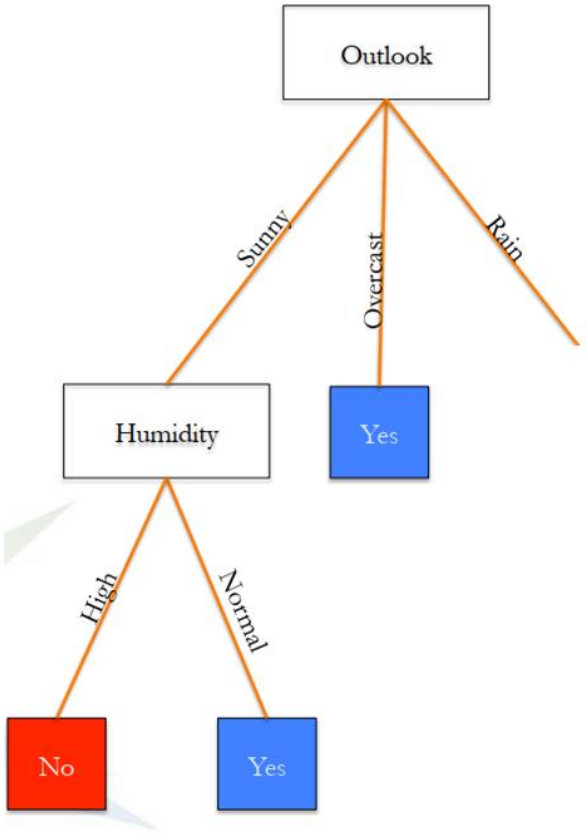
Sunny Outlook data

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Rainy Outlook data

Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Attributes Gain Data

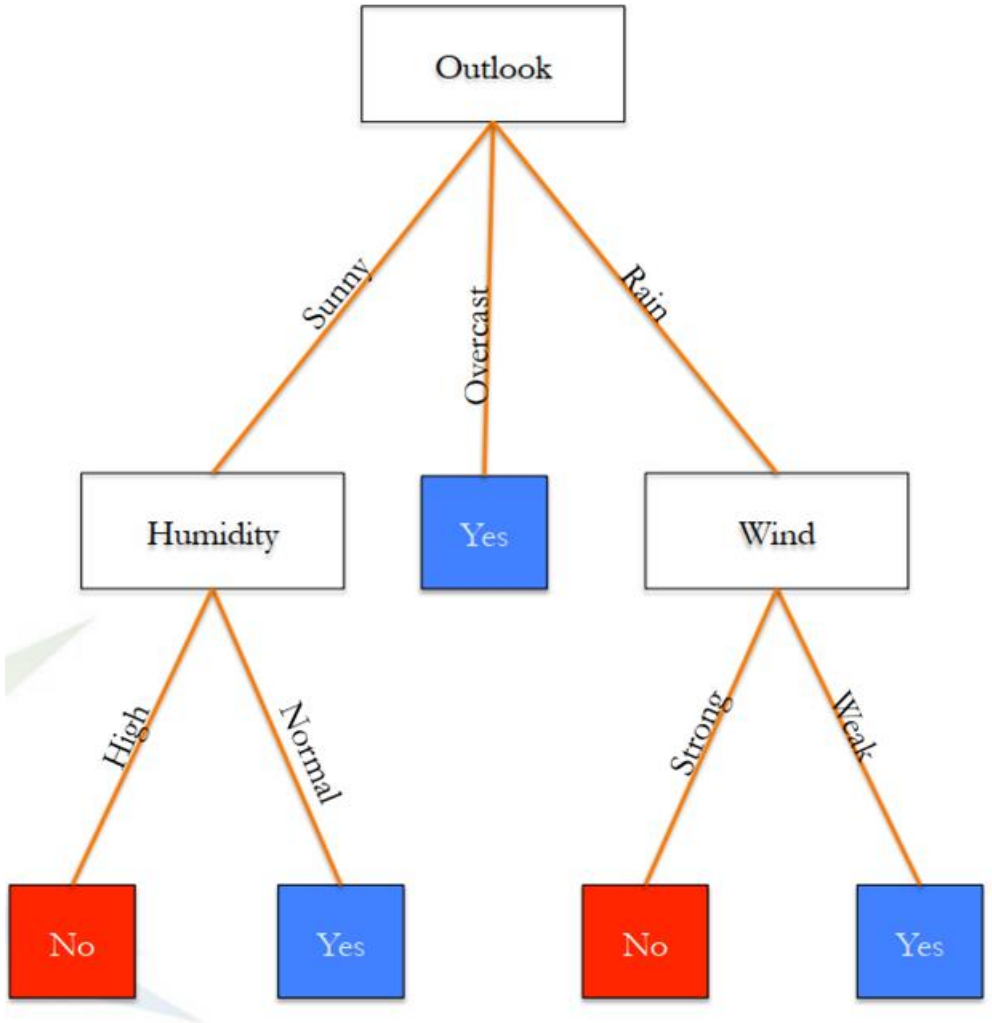


The next node is the humidity

Attributes	Gain
Temperature	0.571
Humidity	0.971
Windy	0.020

Outlook	Humidity	PlayTennis
Sunny	High	No
Sunny	High	No
Sunny	High	No
Sunny	Normal	Yes
Sunny	Normal	Yes

Highest Gain Attributes Data



Attributes	Gain
Humidity	0.020
Windy	0.971
Temperature	0.020

Outlook	Windy	PlayTennis
Rainy	Strong	No
Rainy	Strong	No
Rainy	Weak	Yes
Rainy	Weak	Yes
Rainy	Weak	Yes