COMP9414 Tutorial

Week 7

News

- Assignment 1 submissions have closed
  - Was initially set to auto-submit
    - Check over your marks if you looked at them early
  - Plagiarism checking is currently happening
    - Hope you renamed those variables

- Assignment 2 has been released
  - Due in week 9
  - Should have everything needed to complete it after this week

Background - Entropy

- Measure of the amount of information required to represent something
  - Typically in the form of bits

$$Entropy(S) = -\sum_i p_i log_2 p_i$$

Bits example

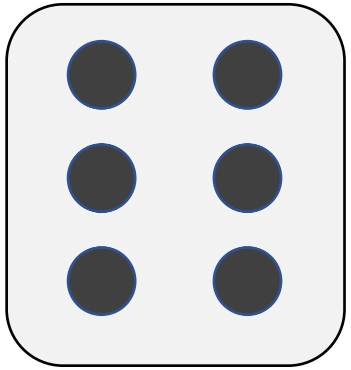
$Entropy(S) = 3$

2^3 = 8 values

- p_i is the probability of some variable having a particular value
- log_2 converts the probability into the number of bits required to represent it

Background - Entropy (Dice Example)

$$Entropy(S) = -\sum_{i} p_i log_2 p_i$$



$$Entropy(S) = \left(\frac{1}{6}\right) log_2 \left(\frac{1}{6}\right) + \left(\frac{1}{6}\right) log_2 \left(\frac{1}{6}\right) + \ldots$$

$$= 6\left(\left(\frac{1}{6}\right) log_2 \left(\frac{1}{6}\right)\right) = 2.585 \text{ (bits)}$$

$2^{2.585} \cong 6.0$ unique values

Background - Entropy (Coin Example)

- Fair coin
  - 50% heads
  - 50% tails
  - Entropy of 1
    - 1 bit required to store all information

$$-(0.5log_2 0.5 + 0.5log_2 0.5) = 1.0$$

- Weighted coin
  - 99% heads
  - 1% tails
  - Entropy of 0.08
    - 0.08 bits required to store all information

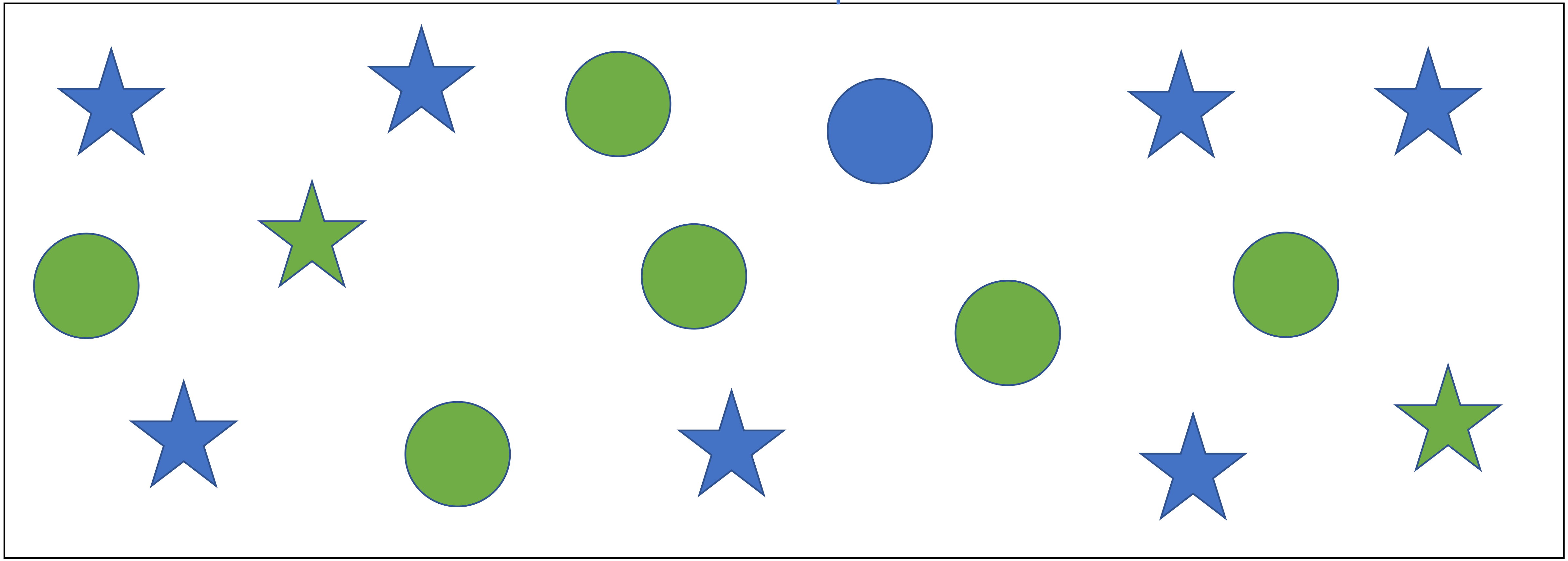
$$-(0.99log_2 0.99 + 0.01log_2 0.01) = 0.08$$

Background - Entropy (Shape Selection Example)

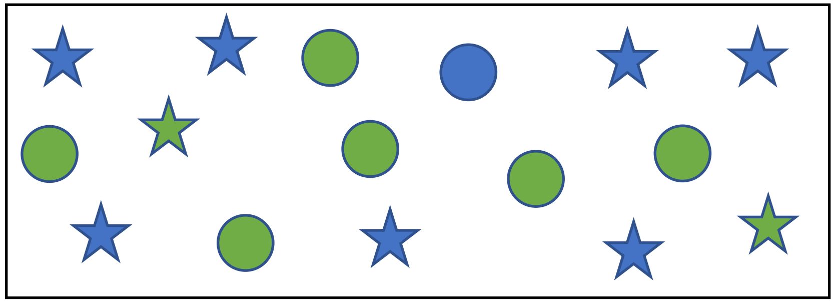
n(star) = n(green) + n(blue) = 2 + 7 = 9

n(circle) = n(green) + n(blue) = 6 + 1 = 7

If a shape is picked at random, will it be a star or a circle?



Background - Entropy (Shape Selection Example)



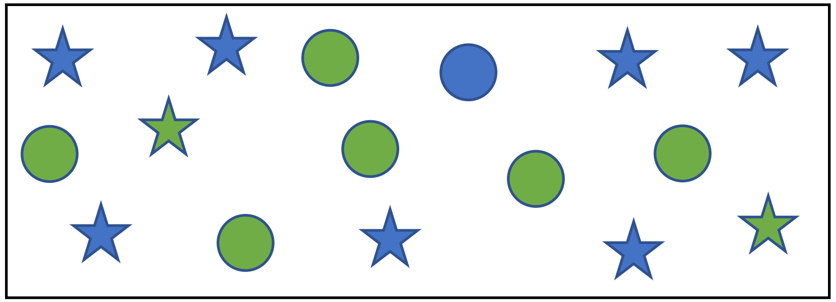
$$E(S)$$

$$= -\left(\left(\frac{9}{16}\right)log_2\left(\frac{9}{16}\right) + \left(\frac{7}{16}\right)log_2\left(\frac{7}{16}\right)\right)$$

$$= -(-0.4669 + -0.5217)$$
$$= 0.9886 \text{ bits}$$

$$2^{0.9886} = 1.9843 \text{ unique values}$$

Background - Entropy (Shape Selection Example)



$$
\begin{aligned}
E(green) \\
&= -(\text{green\_star} + \text{green\_circle}) \\
&= -\left((\tfrac{2}{8}) log_2 (\tfrac{2}{8}) + (\tfrac{6}{8}) log_2 (\tfrac{6}{8}) \right) \\
\\
&= -(-0.5000 + -0.3113) \\
&= 0.8113 \text{ bits}
\end{aligned}
$$

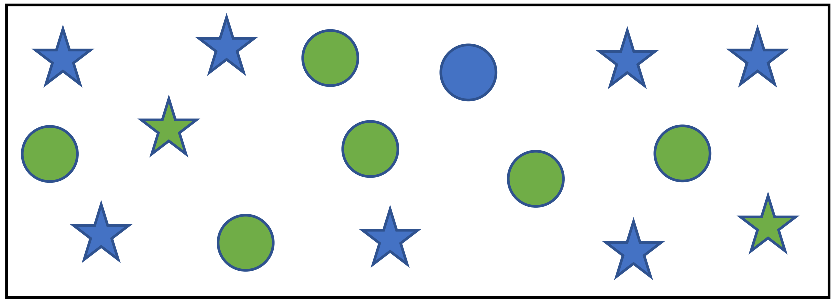
$$2^{0.8113} = 1.7548$$

Background - Entropy (Shape Selection Example)

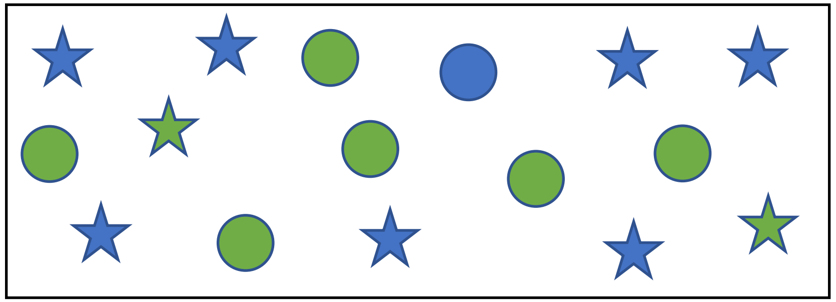
$$E(blue)$$
$$= -(\text{blue\_star} + \text{blue\_circle})$$
$$= -\left((\tfrac{7}{8}) log_2 (\tfrac{7}{8}) + (\tfrac{1}{8}) log_2 (\tfrac{1}{8}) \right)$$

$$= -(-0.1686 + -0.3750)$$
$$= 0.5436 \text{ bits}$$

$$2^{0.5436} = 1.4576$$



Background - Entropy (Shape Selection Example)



$$\begin{aligned}
Combined\ E(colour) \\
&= -(\text{green} + blue) \\
&= -\left(\left(\frac{8}{16}\right)E(green)+\left(\frac{8}{16}\right)E(blue)\right) \\
&= 0.4056 + 0.2718 \\
&= 0.6774 \text{ bits}
\end{aligned}$$

$$2^{0.6774} = 1.59$$

Entropy - Information Gain

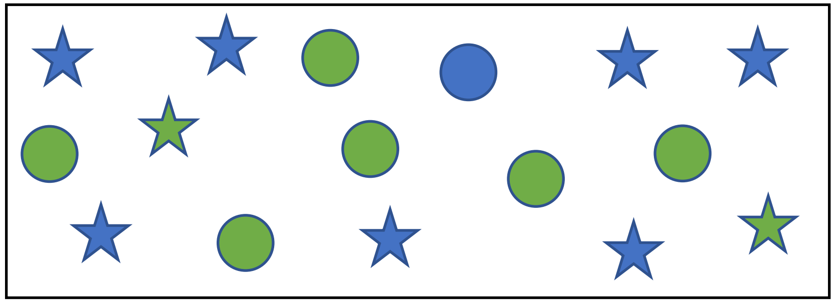
- Details how much information some properties tells you
  - High information gain is an informative property

- High information gain properties should come first in decision trees
  - Splits the tree to a larger degree earlier
  - Resultant tree will be more succinct and compact

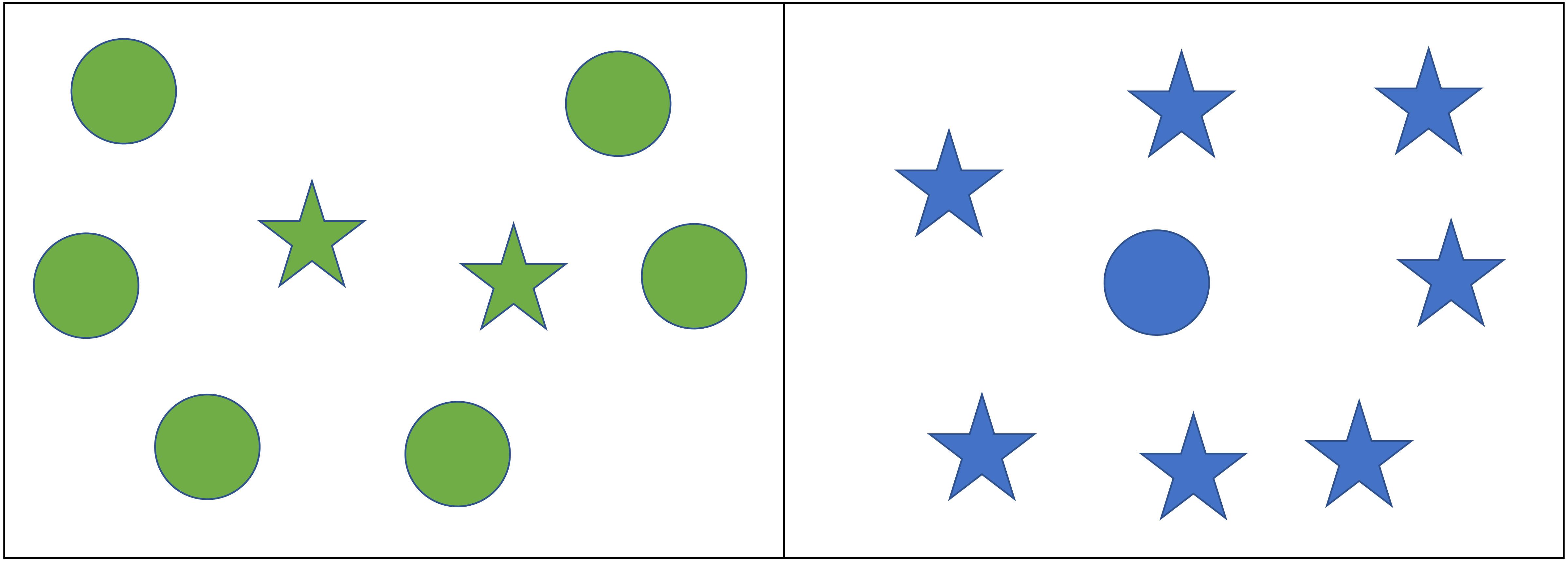
Background - Entropy (Shape Selection Example)

$$Gain(S, colour) = E(S) - E(colour)$$
$$= 0.9886 - 0.6774$$
$$= 0.3112 \text{ bits}$$

$$2^{0.3112} = 1.24$$



Background - Entropy (Shape Selection Example)



Question 1

Day	Outlook	Temperature	Humidity	Wind	Play_tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Question 1 – Parent Entropy

$$\mathrm{E(S)} = -\left(\left(\frac{\text{yes}}{\text{tennis}}\right)\log_2\left(\frac{\text{yes}}{\text{tennis}}\right) + \left(\frac{\text{no}}{\text{tennis}}\right)\log_2\left(\frac{\text{no}}{\text{tennis}}\right)\right)$$

$$= -\left(\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) + \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)\right)$$

$$= -(-0.4098 + -0.5305)$$

$$= 0.940$$

Question 1 – Outlook Entropy

$$\mathrm{E(Sunny)} \quad = -\left(\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right) = 0.971$$

$$\mathrm{E(Overcast)} = -\left(\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) + \left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right)\right) = 0.000$$

$$\mathrm{E(Rain)} \quad = -\left(\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right) = 0.971$$

Question 1 – Outlook Entropy

$\text{Gain}(\mathrm{S}, \mathrm{Outlook})$

$$= \mathrm{E(S)} - \left(\begin{array}{c} (\frac{5}{14})\mathrm{E(Sunny)} + (\frac{4}{14})\mathrm{E(Overcast)} + \\ (\frac{5}{14})\mathrm{E(Rain)} \end{array} \right)$$

$$= 0.940 - (0.3467 + 0 + 0.3467) = 0.2470$$

Question 1 – Temperature Entropy

$$\mathrm{E(Hot)} \;\; = -\left(\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) +- \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right) = 1.000$$

$$\mathrm{E(Mild)} = -\left(\left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) + \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right)\right) = 0.918$$

$$\mathrm{E(Cool)} = -\left(\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)\right) = 0.811$$

Question 1 – Temperature Entropy

$$\text{Gain}(S, \text{Temperature})$$

$$= \mathrm{E(S)} - \left((\frac{4}{14})\mathrm{E(Hot)} + (\frac{6}{14})\mathrm{E(Mild)} + (\frac{4}{14})\mathrm{E(Cool)} \right)$$

$$= 0.940 - (0.2857 + 0.3934 + 0.2317) = 0.0292$$

Question 1 – Humidity Entropy

$$\text{E(High)} \quad = -\left(\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) + - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right)\right) = 0.985$$

$$\text{E(Normal)} = -\left(\left(\frac{6}{7}\right)\log_2\left(\frac{6}{7}\right) + \left(\frac{1}{7}\right)\log_2\left(\frac{1}{7}\right)\right) = 0.592$$

Question 1 – Humidity Entropy

$$\text{Gain}(S, \text{Humidity})$$
$$= \text{E(S)} - \left((\frac{7}{14})\text{E(High)} + (\frac{7}{14})\text{E(Normal)} \right)$$

$$= 0.940 - (0.4925 + 0.296) = 0.1515$$

Question 1 – Wind Entropy

$$\text{E(Weak)} \quad = -\left(\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) + -\left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right)\right) = 0.811$$

$$\text{E(Strong)} \quad = -\left(\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) + \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right)\right) = 1.000$$

Question 1 – Wind Entropy

$$\text{Gain}(\mathrm{S}, \mathrm{Wind})$$
$$= \mathrm{E(S)} - \left(\left(\frac{8}{14}\right)\mathrm{E(Weak)} + \left(\frac{6}{14}\right)\mathrm{E(Strong)} \right)$$

$$= 0.940 - (0.4634 + 0.4286) = 0.0480$$

Question 1 – First Feature Selection Gains

Gain(S, Outlook) = 0.2470
Gain(S, Temperature) = 0.0292
Gain(S, Humidity) = 0.1515
Gain(S, Wind) = 0.0480

Split based on Outlook

Question 1 – Sunny Outlook Entropy

$$\mathrm{E}\left(\mathrm{S}_{\mathrm{Sunny}}\right) = -\left(\left(\frac{\mathrm{yes}}{\mathrm{Sunny}}\right)\log_2\left(\frac{\mathrm{yes}}{\mathrm{Sunny}}\right) + \left(\frac{\mathrm{no}}{\mathrm{Sunny}}\right)\log_2\left(\frac{\mathrm{no}}{\mathrm{Sunny}}\right)\right)$$

$$= -\left(\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)\right)$$

$$= -(-0.5287 + -0.4422)$$

$$= 0.9710$$

Question 1 – Temperature Entropy given Sunny

$$\mathrm{E(Sunny, Hot)} = -\left(\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + -\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right)\right) = 0.000$$

$$\mathrm{E(Sunny, Mild)} = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right) = 1.000$$

$$\mathrm{E(Sunny, Cool)} = -\left(\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) + \left(\frac{0}{1}\right)\log_2\left(\frac{0}{1}\right)\right) = 0.000$$

Question 1 – Temperature Entropy given Sunny

$$\text{Gain}\left(S_{\text{Sunny}}, \text{Temperature}\right) = E\left(S_{\text{Sunny}}\right) - \left(\left(\frac{2}{5}\right)E(\text{Sunny, Hot}) + \left(\frac{2}{5}\right)E(\text{Sunny,Mild}) + \left(\frac{1}{5}\right)E(\text{Sunny,Cool})\right)$$

$$= 0.971 - (0 + 0.4000 + 0) = 0.5710$$

Question 1 – Humidity Entropy given Sunny

$$\mathrm{E(Sunny,High)} = -\left(\left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right) + -\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right)\right) = 0.000$$

$$\mathrm{E(Sunny,Normal)} = -\left(\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) + \left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right)\right) = 0.000$$

$$\text{Gain}\left(\mathrm{S_{Sunny}, Humidity}\right)$$

$$= \mathrm{E}\left(\mathrm{S_{Sunny}}\right) - \left(\left(\frac{3}{5}\right)\mathrm{E(Sunny,High)} + \left(\frac{2}{5}\right)\mathrm{E(Sunny,Normal)}\right)$$

$$= 0.971 - (0 + 0) = 0.9710$$

Question 1 – Wind Entropy given Sunny

$$\mathrm{E(Sunny,Weak)} = -\left(\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) + -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right)\right) = 0.918$$

$$\mathrm{E(Sunny,Strong)} = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right) = 1.000$$

$$\text{Gain}\left(\mathrm{S_{Sunny}}, \mathrm{Wind}\right)$$

$$= \mathrm{E}\left(\mathrm{S_{Sunny}}\right) - \left(\left(\frac{3}{5}\right)\mathrm{E(Sunny,Weak)} + \left(\frac{2}{5}\right)\mathrm{E(Sunny,Strong)}\right)$$

$$= 0.971 - (0.5508 + 0.4000) = 0.0202$$

Question 1 – Rain Outlook Entropy

$$E(S_{\text{Rain}}) = -\left(\left(\frac{\text{yes}}{\text{Rain}}\right)\log_2\left(\frac{\text{yes}}{\text{Rain}}\right) + \left(\frac{\text{no}}{\text{Rain}}\right)\log_2\left(\frac{\text{no}}{\text{Rain}}\right)\right)$$

$$= -\left(\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)\right)$$

$$= -(-0.4422 + -0.5287)$$

$$= 0.9710$$

Question 1 – Temperature Entropy given Rain

$$\text{E(Rain, Hot)} = -\left(\left(\frac{0}{0}\right)\log_2\left(\frac{0}{0}\right) + -\left(\frac{0}{0}\right)\log_2\left(\frac{0}{0}\right)\right) = 0.000$$

$$\text{E(Rain, Mild)} = -\left(\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)\right) = 0.918$$

$$\text{E(Rain, Cool)} = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right) = 1.000$$

Question 1 – Temperature Entropy given Rain

Gain$\left(S_{\mathrm{Rain}}, \mathrm{Temperature}\right)$

$$= \mathrm{E}\left(S_{\mathrm{Rain}}\right) - \left((\frac{0}{5})\mathrm{E}(\mathrm{Rain, Hot}) + (\frac{3}{5})\mathrm{E}(\mathrm{Rain{,}Mild}) + (\frac{2}{5})\mathrm{E}(\mathrm{Rain{,}Cool})\right)$$

$$= 0.971 - (0 + 0.5508 + 0.4000) = 0.0202$$

Question 1 – Humidity Entropy given Rain

$$\mathrm{E(Rain,High)} = -\left(\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) + -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)\right) = 1.000$$

$$\mathrm{E(Rain,Normal)} = -\left(\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)\right) = 0.551$$

$$\text{Gain}\left(S_{\mathrm{Rain}}, \mathrm{Humidity}\right)$$

$$= \mathrm{E}\left(S_{\mathrm{Rain}}\right) - \left(\left(\frac{2}{5}\right)\mathrm{E(Rain,High)} + \left(\frac{3}{5}\right)\mathrm{E(Rain,Normal)}\right)$$

$$= 0.971 - (0.4000 + 0.3306) = 0.2404$$

Question 1 – Wind Entropy given Rain

$$\text{E(Rain,Weak)} = -\left(\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) + -\left(\frac{0}{3}\right)\log_2\left(\frac{0}{3}\right)\right) = 0.000$$

$$\text{E(Rain,Strong)} = -\left(\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) + \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right)\right) = 0.000$$

$$\text{Gain}\left(S_{\text{Rain}}, \text{Wind}\right)$$

$$= \text{E}\left(S_{\text{Rain}}\right) - \left(\left(\frac{3}{5}\right)\text{E(Rain,Weak)} + \left(\frac{2}{5}\right)\text{E(Rain,Strong)}\right)$$

$$= 0.971 - (0 + 0) = 0.971$$

Question 1 – Second Feature Selection Gains

Gain(S_{Sunny}, Temperature) = 0.5710
Gain(S_{Sunny}, Humidity) = 0.9710
Gain(S_{Sunny}, Wind) = 0.0202

Gain(S_{Rain}, Temperature) = 0.0202
Gain(S_{Rain}, Humidity) = 0.2404
Gain(S_{Rain}, Wind) = 0.971

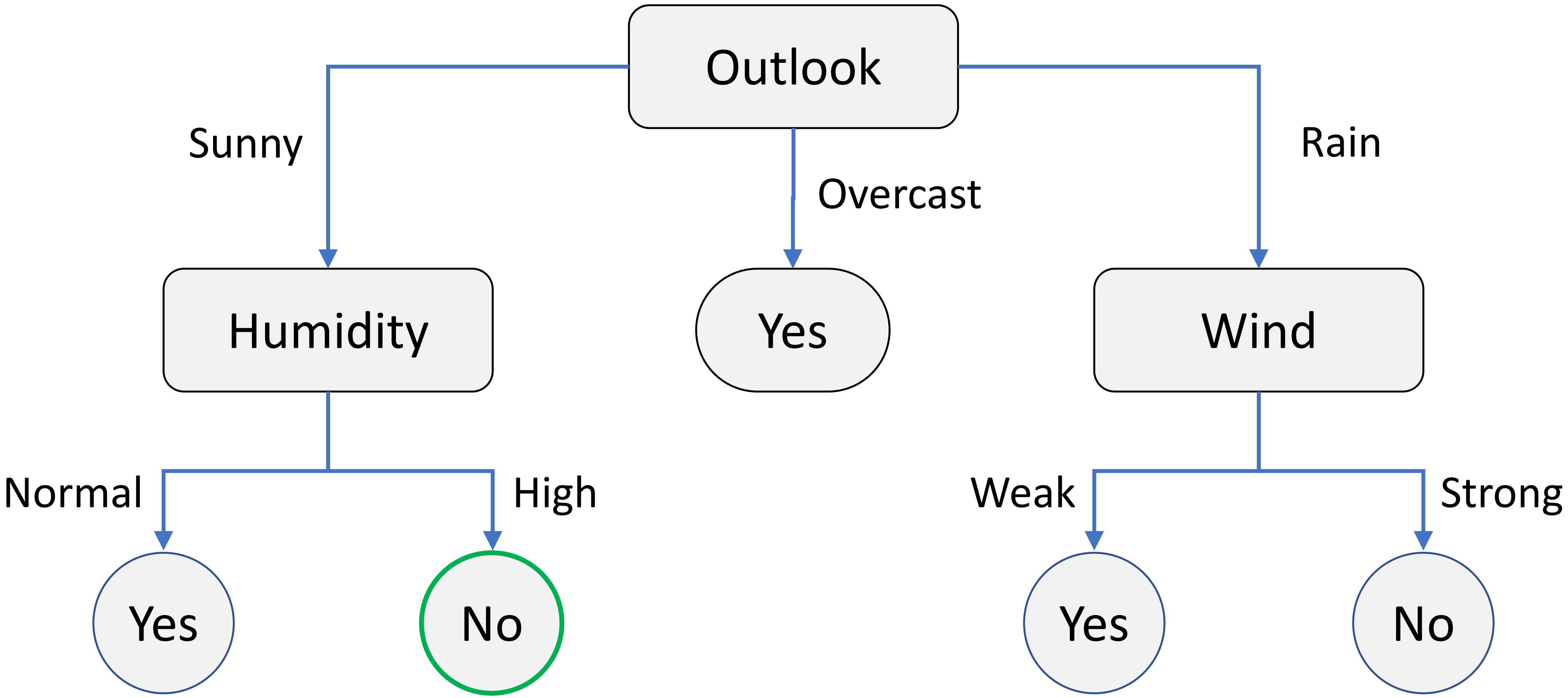
Split based on Humidity for Sunny
Split based on Wind for Rain

Question 1 – Second Feature Selection Gains

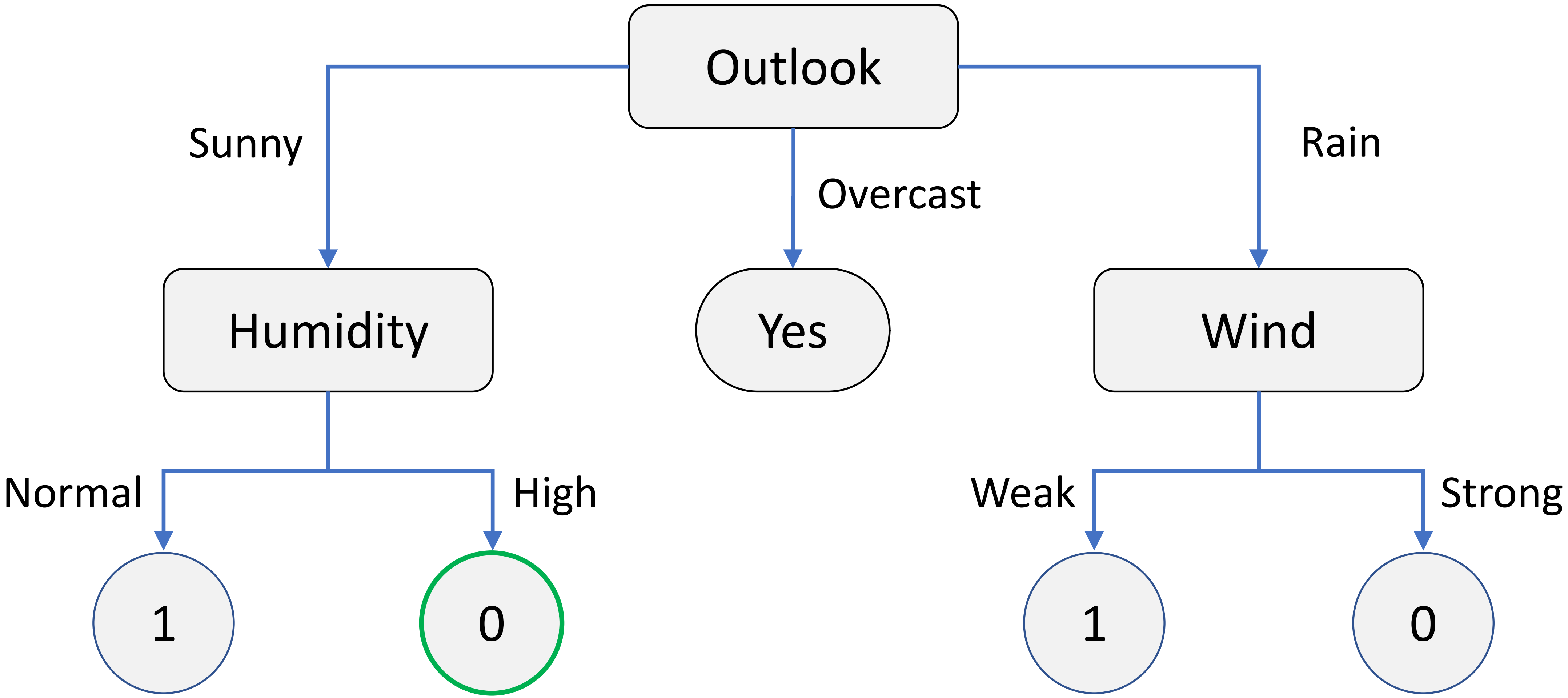
No other splits are required since the entropy at the current splits is 0.

This means that no bits are required to represent the information since there is only one answer.

Question 1 – Final Decision Tree



Question 1 – Final Decision Tree



Bayes Theorem - Recap

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

Probability of a given event occurring based on our prior knowledge or evidence

$$P(Hypothesis \mid Evidence) = \frac{P(Evidence \mid Hypothesis) \times P(Hypothesis)}{P(Evidence)}$$

Question 2 – Bayes Classifier

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

$$P(Play \mid Outlook, Temperature, Humidity, Wind)$$

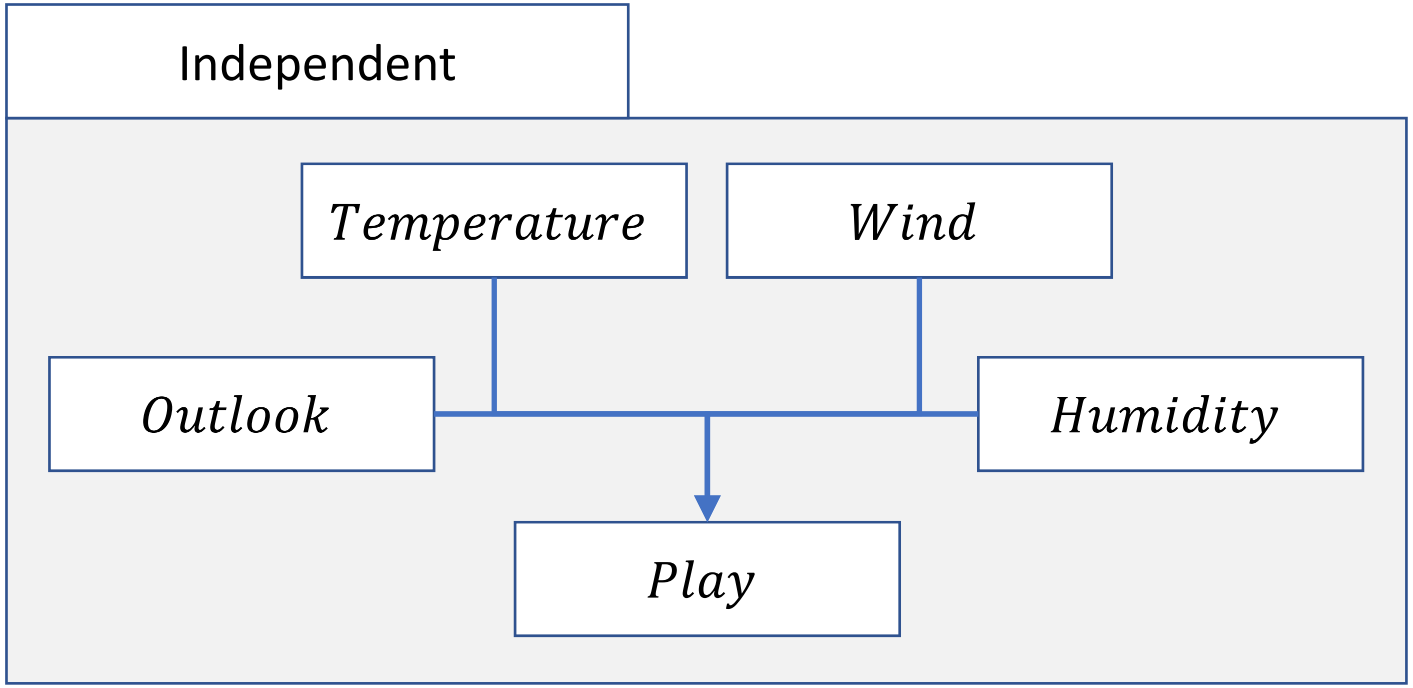
$$= \frac{P(Outlook, Temperature, Humidity, Wind \mid Play) \times P(Play)}{P(Outlook, Temperature, Humidity, Wind)}$$

$$= \frac{P(Sunny, Hot, High, Weak \mid Play) \times P(Play)}{P(Sunny, Hot, High, Weak)}$$

$$= \frac{P(Sunny, Hot, High, Weak \mid \neg Play) \times P(\neg Play)}{P(Sunny, Hot, High, Weak)}$$

Question 2 – Bayes Classifier

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$



Question 2 – Bayes Classifier

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

$$P(Sunny, Hot, High, Weak \mid Play) \times P(Play)$$

$$P(Sunny \mid Play) \times P(Hot \mid Play) \times P(High \mid Play) \times P(Weak \mid Play) \times P(Play)$$

$$= \frac{2}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{9}{14} = 0.00705$$

Question 2 – Bayes Classifier

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

$$P(Sunny, Hot, High, Weak \mid \neg Play) \times P(\neg Play)$$

$$P(Sunny \mid \neg Play) \times P(Hot \mid \neg Play) \times P(High \mid \neg Play) \times P(Weak \mid \neg Play) \times P(\neg Play)$$

$$= \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{2}{5} \times \frac{5}{14} = 0.02743$$

Question 2 – Bayes Classifier

$P(H \mid E) =$	$\dfrac{P(E \mid H) \times P(H)}{P(E)}$

$\dfrac{0.00705}{1} + \dfrac{0.02743}{1} = 0.03448$	No

Notes

- Assignment 2 packages are as follows:
  - Preprocessing toolkit:
    - https://www.nltk.org/
  - Modelling toolkit
    - https://scikit-learn.org/stable/