1 Decision Tree Learning

1.1

Value(Outlook) = sunny, overcast, rain S = [9+, 5-]
S sunny $\leftarrow [2+, 3-]$
$Sovercast \leftarrow [4+, 0-]$
$Srain \leftarrow [3+, 2-]$
$Entropy(S) = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.9403$
$Entropy(Ssunny) = -\frac{2}{5} \cdot \log_2(\frac{2}{5}) - \frac{3}{5} \cdot \log_2(\frac{3}{5}) = 0.9709$
$Entropy(Sovercast) = -1 \cdot \log_2(1) - 0 \cdot \log_2(0) = 0$
$Entropy(Srain) = -\frac{3}{5} \cdot \log_2(\frac{3}{5}) - \frac{2}{5} \cdot \log_2(\frac{2}{5}) = 0.9709$
$Gain(S, Outlook) = Entropy(S) - \frac{5}{14} \cdot Entropy(Ssunny) - \frac{4}{14} \cdot Entropy(Sovercast) - \frac{5}{14} \cdot Entropy(Srain) = 0.2467$

$Value(Humidity) = high(> 75), low(\leq 75)$
$S = [9+, 5-]$
$Shigh \leftarrow [5+, 4-]$
$Slow \leftarrow [4+, 1-]$
$Entropy(S) = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.9403$
$Entropy(Shigh) = -\frac{5}{9} \cdot \log_2(\frac{5}{9}) - \frac{4}{9} \cdot \log_2(\frac{4}{9}) = 0.9911$
$Entropy(Slow) = -\frac{4}{5} \cdot \log_2(\frac{4}{5}) - \frac{1}{5} \cdot \log_2(\frac{1}{5}) = 0.7219$
$Gain(S, Humidity) = Entropy(S) - \frac{9}{14} \cdot Entropy(Shigh) - \frac{5}{14} \cdot Entropy(Slow)$
$= 0.04524$

1.2

SplitInfo = - (5/14) $\cdot \log_2(\frac{5}{14}) - \frac{4}{14} \cdot \log_2(\frac{4}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 1.5774$
$GainRatio(Outlook) = \frac{Gain(Outlook)}{Splitinfo} = 0.156$

$SplitInfo = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.9403$
$GainRatio(Humidity) = \frac{Gain(Humidity)}{Splitinfo} = 0.048$

1.3

Value(Temp) = high(¿70), low($\leq$ 70)
$S = [9+, 5-]$
$Shigh \leftarrow [5+, 4-]$
$Slow \leftarrow [4+, 1-]$
$Entropy(S) = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.9403$
$Entropy(S\,high) = -\frac{5}{9} \cdot \log_2(\frac{5}{9}) - \frac{4}{9} \cdot \log_2(\frac{4}{9}) = 0.9911$
$Entropy(S\,low) = -\frac{4}{5} \cdot \log_2(\frac{4}{5}) - \frac{1}{5} \cdot \log_2(\frac{1}{5}) = 0.7219$
$Gain(S, Temp) = Entropy(S) - \frac{9}{14} \cdot Entropy(Shigh) - \frac{5}{14} \cdot Entropy(Slow)$
$= 0.04524$

$Value(Windy) = true, false$
$S = [9+, 5-]$
$Strue \leftarrow [3+, 3-]$
$Sfalse \leftarrow [6+, 2-]$
$Entropy(S) = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.9403$
$Entropy(Strue) = -\frac{1}{2} \cdot \log_2(\frac{1}{2}) - \frac{1}{2} \cdot \log_2(\frac{1}{2}) = 1$
$Entropy(Sfalse) = -\frac{3}{4} \cdot \log_2(\frac{3}{4}) - \frac{1}{4} \cdot \log_2(\frac{1}{4}) = 0.6226$
$Gain(S, Windy) = Entropy(S) - \frac{6}{14} \cdot Entropy(Strue) - \frac{8}{14} \cdot Entropy(Sfalse)$
$= 0.1559$

Explanation: Outlook is the attribute with the largest information gain, therefore, it is chosen as the decision node.
We can draw a decision tree by mapping through the root node to the leaf node one by one