Question 1

(a)

item	x_1	x_2	x_3	x_4	y
0	0	0	1	0	0
1	0	1	0	0	0
2	0	0	1	1	1
3	1	0	0	1	1
4	0	1	1	0	0
5	1	1	0	0	0
6	0	1	0	1	0

1) **Level One**
Available attributes = $\{x_1, x_2, x_3, x_4\}$, Data = {0,1,2,3,4,5,6}
$p_+ = 2/7$, $p_- = 5/7$
$H(S) = -\left(\frac{2}{7}\right)\log\frac{2}{7} - \left(\frac{5}{7}\right)\log\frac{5}{7} = 0.86$

x_1:
- 0: 5 of 7 examples, $p_+ = 1/5$, $p_- = 4/5$
  $H(x_1 = 0) = -\left(\frac{1}{5}\right)\log\frac{1}{5} - \left(\frac{4}{5}\right)\log\frac{4}{5} = 0.72$
- 1: 2 of 7 examples, $p_+ = 1/2$, $p_- = 1/2$
  $H(x_1 = 1) = -\left(\frac{1}{2}\right)\log\frac{1}{2} - \left(\frac{1}{2}\right)\log\frac{1}{2} = 1$
- IG: $0.86 - \left(\frac{5}{7} * 0.72 + \frac{2}{7} * 1\right) = \mathbf{0.06}$

x_2:
- 0: 3 of 7 examples, $p_+ = 2/3$, $p_- = 1/3$
  $H(x_2 = 0) = -\left(\frac{2}{3}\right)\log\frac{2}{3} - \left(\frac{1}{3}\right)\log\frac{1}{3} = 0.92$
- 1: 4 of 7 examples, $p_+ = 0/4$, $p_- = 4/4$
  $H(x_2 = 1) = 0$
- IG: $0.86 - \left(\frac{3}{7} * 0.92 + \frac{4}{7} * 0\right) = \mathbf{0.47}$

x_3:
- 0: 4 of 7 examples, $p_+ = 1/4$, $p_- = 3/4$
  $H(x_3 = 0) = -\left(\frac{1}{4}\right)\log\frac{1}{4} - \left(\frac{3}{4}\right)\log\frac{3}{4} = 0.81$
- 1: 3 of 7 examples, $p_+ = 1/3$, $p_- = 2/3$
  $H(x_3 = 1) = -\left(\frac{1}{3}\right)\log\frac{1}{3} - \left(\frac{2}{3}\right)\log\frac{2}{3} = 0.92$
- IG: $0.86 - \left(\frac{4}{7} * 0.81 + \frac{3}{7} * 0.92\right) = \mathbf{0.003}$

x_4:
- 0: 4 of 7 examples, $p_+ = 0/4$, $p_- = 4/4$
  $H(x_4 = 0) = 0$
- 1: 3 of 7 examples, $p_+ = 2/3$, $p_- = 1/3$
  $H(x_4 = 1) = -\left(\frac{2}{3}\right)\log\frac{2}{3} - \left(\frac{1}{3}\right)\log\frac{1}{3} = 0.92$
- IG: $0.86 - \left(\frac{4}{7} * 0 + \frac{3}{7} * 0.92\right) = \mathbf{0.47}$

Highest information gain is from splitting on either x_2 or x_4, so lets choose x_2.

2) **Level Two**, splitting on x_2.

$x_2 = 0$: Available Attributes = $\{x_1, x_3, x_4\}$, Data = {0,2,3}
$H(S) = 0.92$, as previously calculated.

item	x_1	x_3	x_4	y
0	0	1	0	0
2	0	1	1	1
3	1	0	1	1

x_1:

0: 2 of 3 examples, $p_+ = 1/2$, $p_- = 1/2$
$H(x_1 = 0|x_2 = 0) = -\left(\frac{1}{2}\right)\log\frac{1}{2} - \left(\frac{1}{2}\right)\log\frac{1}{2} = 1$

1: 1 of 3 examples, $p_+ = 1$, $p_- = 0$
$H(x_1 = 1|x_2 = 0) = 0$

IG: $0.92 - \left(\frac{2}{3} * 1 + \frac{1}{3} * 0\right) = \mathbf{0.25}$

x_3:

0: 1 of 3 examples, $p_+ = 1$, $p_- = 0$
$H(x_3 = 0|x_2 = 0) = 0$

1: 2 of 3 examples, $p_+ = 1/2$, $p_- = 1/2$
$H(x_3 = 1|x_2 = 0) = -\left(\frac{1}{2}\right)\log\frac{1}{2} - \left(\frac{1}{2}\right)\log\frac{1}{2} = 1$

IG: $0.86 - \left(\frac{1}{3} * 0 + \frac{2}{3} * 1\right) = \mathbf{0.25}$

x_4:

0: 1 of 3 examples, $p_+ = 0/4$, $p_- = 1$
$H(x_4 = 0|x_2 = 0) = 0$

1: 2 of 3 examples, $p_+ = 2/2$, $p_- = 0$
$H(x_4 = 1|x_2 = 0) = 0$

IG: $0.86 - \left(\frac{1}{3} * 0 + \frac{2}{3} * 0\right) = \mathbf{0.86}$

So the highest information gain for this node is x_4, which makes sense since it uniquely identifies instance.

$x_2 = 1$: Available Attributes = $\{x_1, x_3, x_4\}$, Data = {1,4,5,6}
$H(S) = 0$, as previously calculated.

item	x_1	x_3	x_4	y
1	0	0	0	0
4	0	1	0	0
5	1	0	0	0
6	0	0	1	0

This node uniquely identifies each instance and thus becomes a leaf with $y = 0$.

3) **Level 3**, splitting on x_4 given $x_2 = 0$.

$x_4 = 0$: Available Attributes = $\{x_1, x_3\}$, Data = $\{0\}$
$H(S) = 0$, as previously calculated.

item	x_1	x_3	y
0	0	1	0

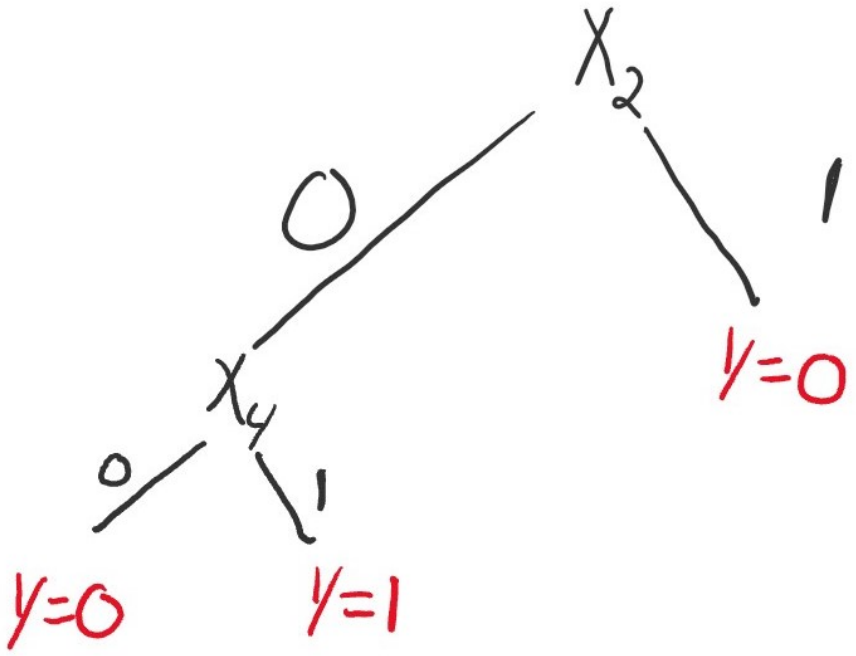
This node uniquely identifies each instance and thus becomes a leaf with $y = 0$.

$x_4 = 1$: Available Attributes = $\{x_1, x_3\}$, Data = $\{2, 3\}$
$H(S) = 0$, as previously calculated.

item	x_1	x_3	y
2	0	1	1
3	1	0	1

This node uniquely identifies each instance and thus becomes a leaf with $y = 1.4$

All training data is labeled at this point so the tree ends here, an image of the tree can be found below.



(b) The function is $\neg x_2 \wedge x_4$. This will result in $y = 1$ only if $x_2 = 0$ and $x_4 = 1$. In all other cases $y = 0$. It seems a bit overkill to include the full table for this but whatever, here it is.

x_1	x_2	x_3	x_4	y
0	0	0	0	0
0	0	0	1	1
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	1
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

Question 2

item	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

(a) **Majority Error**

1) **Level One**

Available Attributes = $\{O, T, H, W\}$, Data = $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$
$p_+ = 9/14$, $p_- = 5/14$
$ME = 5/14$

O:

S: 5 of 14 examples, $p_+ = 2/5$, $p_- = 3/5$
$ME(O = S) = 2/5$

O: 4 of 14 examples, $p_+ = 4/4$, $p_- = 0$
$ME(O = O) = 0$

R: 5 of 14 examples, $p_+ = 3/5$, $p_- = 2/5$
$ME(O = R) = 2/5$

IG: $\frac{5}{14} - \left(\frac{5}{14} * \frac{2}{5} + \frac{4}{14} * 0 + \frac{5}{14} * \frac{2}{5}\right) = \mathbf{\frac{1}{14}}$

T:

H: 4 of 14 examples, $p_+ = 2/4$, $p_- = 2/4$
$ME(T = H) = 2/4 = 1/2$

M: 6 of 14 examples, $p_+ = 4/6$, $p_- = 2/6$
$ME(T = M) = 2/6 = 1/3$

C: 4 of 14 examples, $p_+ = 3/4$, $p_- = 1/4$
$ME(T = C) = 1/4$

IG: $\frac{5}{14} - \left(\frac{4}{14} * \frac{1}{2} + \frac{6}{14} * \frac{1}{3} + \frac{4}{14} * \frac{1}{4}\right) = \mathbf{0}$

H:

H: 7 of 14 examples, $p_+ = 3/7$, $p_- = 4/7$
$ME(H = H) = 3/7$

N: 7 of 14 examples, $p_+ = 6/7$, $p_- = 1/7$
$ME(H = N) = 1/7$

L: 0 of 14 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L) = 0$
IG: $\frac{5}{14} - \left(\frac{7}{14} * \frac{3}{7} + \frac{7}{14} * \frac{1}{7} + \frac{0}{14} * 0\right) = \mathbf{\frac{1}{14}}$

W:

S: 6 of 14 examples, $p_+ = 3/6$, $p_- = 3/6$
$ME(W = S) = 3/6 = 1/2$

W: 8 of 14 examples, $p_+ = 6/8$, $p_- = 2/8$
$ME(W = W) = 2/8 = 1/4$

IG: $\frac{5}{14} - \left(\frac{6}{14} * \frac{1}{2} + \frac{8}{14} * \frac{1}{4}\right) = \mathbf{0}$

So we find that Outlook and Humidity give us the same information gain and can pick either of them. Lets, go with Outlook.

2) **Level Two**, splitting on O
Available Attributes = $\{T, H, W\}$

O = S: Data = $\{1, 2, 8, 9, 11\}$
$ME = 2/5$

T:

H: 2 of 5 examples, $p_+ = 0/2$, $p_- = 2/2$
$ME(T = H|O = S) = 0$

M: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(T = M|O = S) = 1/2$

C: 1 of 5 examples, $p_+ = 1/1$, $p_- = 0/1$
$ME(T = C|O = S) = 0$

IG: $\frac{2}{5} - \left(\frac{2}{5} * 0 + \frac{2}{5} * \frac{1}{2} + \frac{1}{5} * 0\right) = \mathbf{\frac{1}{5}}$

H:

H: 3 of 5 examples, $p_+ = 0/3$, $p_- = 3/3$
$ME(H = H|O = S) = 0$

N: 2 of 5 examples, $p_+ = 2/2$, $p_- = 0/2$
$ME(H = N|O = S) = 0$

L: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L|O = S) = 0$

IG: $\frac{2}{5} - \left(\frac{3}{5} * 0 + \frac{2}{5} * 0 + \frac{0}{5} * 0\right) = \mathbf{\frac{2}{5}}$

W:

S: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(W = S|O = S) = 1/2$

W: 3 of 5 examples, $p_+ = 1/3$, $p_- = 2/3$
$ME(W = W|O = S) = 1/3$

IG: $\frac{2}{5} - \left(\frac{2}{5} * \frac{1}{2} + \frac{3}{5} * \frac{1}{3}\right) = \mathbf{0}$

The most information can be gained by splitting this node on Humidity.

O = O: Data = $\{3, 7, 12, 13\}$
$ME = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

O = R: Data = $\{4, 5, 6, 10, 14\}$
$ME = 2/5$
T:
H: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$ME(T = H|O = R) = 0$
M: 3 of 5 examples, $p_+ = 2/3$, $p_- = 1/3$
$ME(T = M|O = R) = 1/3$
C: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(T = C|O = R) = 1/2$
IG: $\frac{2}{5} - \left(\frac{0}{5} * 0 + \frac{3}{5} * \frac{1}{3} + \frac{2}{5} * \frac{1}{2}\right) = \mathbf{0}$
H:
H: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(H = H|O = R) = 1/2$
N: 3 of 5 examples, $p_+ = 2/3$, $p_- = 1/3$
$ME(H = N|O = R) = 1/3$
L: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L|O = R) = 0$
IG: $\frac{2}{5} - \left(\frac{2}{5} * \frac{1}{2} + \frac{3}{5} * \frac{1}{3} + \frac{0}{5} * 0\right) = \mathbf{0}$
W:
S: 2 of 5 examples, $p_+ = 0/2$, $p_- = 2/2$
$ME(W = S|O = R) = 0$
W: 3 of 5 examples, $p_+ = 3/3$, $p_- = 0/3$
$ME(W = W|O = R) = 0$
IG: $\frac{2}{5} - \left(\frac{2}{5} * 0 + \frac{3}{5} * 0\right) = \mathbf{\frac{2}{5}}$
The most information can be gained by splitting this node on Wind.

3) **Level Three**

Branch One, splitting on H given O = S
Available Attributes = $\{T, W\}$

H = H: Data = $\{1, 2, 8\}$
$ME = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

H = N: Data = $\{9, 11\}$
$ME = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

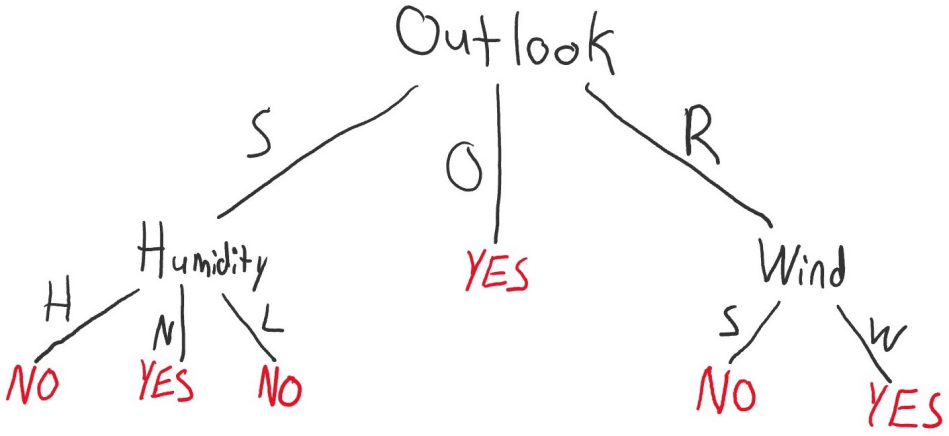
H = L: Data = $\{\}$
We have no training data matching this case so it will become a leaf with *play* equal to the most common occurrence of it's parent node. Thus $play = No$ for this.

Branch Two, splitting on W given O = R
Available Attributes = $\{T, H\}$

W = S: Data = $\{6, 14\}$ $ME = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

W = W: Data = $\{4, 5, 10\}$ $ME = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

All training data is labeled at this point and the tree stops here. An image of the tree can be found below.



(b) **Gini Index**

1) **Level One**
Available Attributes = $\{O, T, H, W\}$, Data = $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$
$p_+ = 9/14$, $p_- = 5/14$
$GI = 1 - (\frac{9}{14}^2 + \frac{5}{14}^2 = 0.46$

O:

S: 5 of 14 examples, $p_+ = 2/5$, $p_- = 3/5$
$GI(O = S) = 1 - \left(\frac{2}{5}^2 + \frac{3}{5}^2\right) = 0.48$

O: 4 of 14 examples, $p_+ = 4/4$, $p_- = 0$
$GI(O = O) = 1 - \left(\frac{4}{4}^2 + \frac{0}{4}^2\right) = 0$

R: 5 of 14 examples, $p_+ = 3/5$, $p_- = 2/5$
$GI(O = R) = 1 - \left(\frac{3}{5}^2 + \frac{2}{5}^2\right) = 0.48$

IG: $0.46 - \left(\frac{5}{14} * 0.48 + \frac{4}{14} * 0 + \frac{5}{14} * 0.48\right) = \mathbf{0.12}$

T:

H: 4 of 14 examples, $p_+ = 2/4$, $p_- = 2/4$
$GI(T = H) = 1 - \left(\frac{2}{4}^2 + \frac{2}{4}^2\right) = 0.5$

M: 6 of 14 examples, $p_+ = 4/6$, $p_- = 2/6$
$GI(T = M) = 1 - \left(\frac{4}{6}^2 + \frac{2}{6}^2\right) = 0.44$

C: 4 of 14 examples, $p_+ = 3/4$, $p_- = 1/4$
$GI(T = C) = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = 0.375$
IG: $0.46 - \left(\frac{4}{14} * 0.5 + \frac{6}{14} * 0.44 + \frac{4}{14} * 0.375\right) = \mathbf{0.02}$

H:
H: 7 of 14 examples, $p_+ = 3/7$, $p_- = 4/7$
$GI(H = H) = 1 - \left(\frac{3}{7}^2 + \frac{4}{7}^2\right) = 0.49$
N: 7 of 14 examples, $p_+ = 6/7$, $p_- = 1/7$
$GI(H = N) = 1 - \left(\frac{6}{7}^2 + \frac{1}{7}^2\right) = 0.24$
L: 0 of 14 examples, $p_+ = 0$, $p_- = 0$
$GI(H = L) = 1$
IG: $0.46 - \left(\frac{7}{14} * 0.49 + \frac{7}{14} * 0.24 + \frac{0}{14} * 1\right) = \mathbf{0.095}$

W:
S: 6 of 14 examples, $p_+ = 3/6$, $p_- = 3/6$
$GI(W = S) = 1 - \left(\frac{3}{6}^2 + \frac{3}{6}^2\right) = 0.5$
W: 8 of 14 examples, $p_+ = 6/8$, $p_- = 2/8$
$GI(W = W) = 1 - \left(\frac{6}{8}^2 + \frac{2}{8}^2\right) = 0.375$
IG: $0.46 - \left(\frac{6}{14} * 0.5 + \frac{8}{14} * 0.375\right) = \mathbf{0.031}$

So we find that Outlook gives the most information gain and it becomes our initial split point.

2) **Level Two**, splitting on O
Available Attributes = $\{T, H, W\}$

O = S: Data = $\{1, 2, 8, 9, 11\}$
$GI = 0.48$

T:
H: 2 of 5 examples, $p_+ = 0/2$, $p_- = 2/2$
$GI(T = H|O = S) = 1 - \left(\frac{0}{2}^2 + \frac{2}{2}^2\right) = 0$
M: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$GI(T = M|O = S) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$
C: 1 of 5 examples, $p_+ = 1/1$, $p_- = 0/1$
$GI(T = C|O = S) = 1 - \left(\frac{1}{1}^2 + \frac{0}{1}^2\right) = 0$
IG: $0.48 - \left(\frac{2}{5} * 0 + \frac{2}{5} * 0.5 + \frac{1}{5} * 0\right) = \mathbf{0.28}$

H:
H: 3 of 5 examples, $p_+ = 0/3$, $p_- = 3/3$
$GI(H = H|O = S) = 1 - \left(\frac{0}{3}^2 + \frac{3}{3}^2\right) = 0$
N: 2 of 5 examples, $p_+ = 2/2$, $p_- = 0/2$
$GI(H = N|O = S) = 1 - \left(\frac{2}{2}^2 + \frac{0}{2}^2\right) = 0$
L: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$GI(H = L|O = S) = 1 - \left(0^2 + 0^2\right) = 1$
IG: $0.48 - \left(\frac{3}{5} * 0 + \frac{2}{5} * 0 + \frac{0}{5} * 1\right) = \mathbf{0.48}$

W:

S: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$GI(W = S|O = S) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$
W: 3 of 5 examples, $p_+ = 1/3$, $p_- = 2/3$
$GI(W = W|O = S) = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = 0.44$
IG: $0.48 - \left(\frac{2}{5} * 0.5 + \frac{3}{5} * 0.44\right) = \mathbf{0.016}$
The most information can be gained by splitting this node on Humidity.

O = O: Data = $\{3, 7, 12, 13\}$
$GI = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

O = R: Data = $\{4, 5, 6, 10, 14\}$
$GI = 0.48$
T:
H: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$GI(T = H|O = R) = 1 - \left(0^2 + 0^2\right) = 1$
M: 3 of 5 examples, $p_+ = 2/3$, $p_- = 1/3$
$GI(T = M|O = R) = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = 0.44$
C: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$GI(T = C|O = R) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$
IG: $0.48 - \left(\frac{0}{5} * 1 + \frac{3}{5} * 0.44 + \frac{2}{5} * 0.5\right) = \mathbf{0.016}$
H:
H: 2 of 5 examples, $p_+ = 1/2$, $p_- = 1/2$
$GI(H = H|O = R) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$
N: 3 of 5 examples, $p_+ = 2/3$, $p_- = 1/3$
$GI(H = N|O = R) = 1 - \left(\frac{2}{3}^2 + \frac{1}{3}^2\right) = 0.44$
L: 0 of 5 examples, $p_+ = 0$, $p_- = 0$
$GI(H = L|O = R) = 1 - \left(0^2 + 0^2\right) = 1$
IG: $0.48 - \left(\frac{2}{5} * 0.5 + \frac{3}{5} * 0.44 + \frac{0}{5} * 1\right) = \mathbf{0.016}$
W:
S: 2 of 5 examples, $p_+ = 0/2$, $p_- = 2/2$
$GI(W = S|O = R) = 1 - \left(\frac{0}{2}^2 + \frac{2}{2}^2\right) = 0$
W: 3 of 5 examples, $p_+ = 3/3$, $p_- = 0/3$
$GI(W = W|O = R) = 1 - \left(\frac{0}{2}^2 + \frac{2}{2}^2\right) = 0$
IG: $0.48 - \left(\frac{2}{5} * 0 + \frac{3}{5} * 0\right) = \mathbf{0.48}$
The most information can be gained by splitting this node on Wind.

3) **Level Three**

Branch One, splitting on H given O = S
Available Attributes = $\{T, W\}$

H = H: Data = $\{1, 2, 8\}$
$GI = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

H = N: Data = $\{9, 11\}$
$GI = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

H = L: Data = $\{\}$
We have no training data matching this case so it will become a leaf with *play* equal to the most common occurrence of it's parent node. Thus $play = No$ for this.

Branch Two, splitting on W given O = R
Available Attributes = $\{T, H\}$

W = S: Data = $\{6, 14\}$ $GI = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

W = W: Data = $\{4, 5, 10\}$ $GI = 0$
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

All training data is labeled at this point and the tree stops here. This tree is identical to the one formed by using ME so the image can be found a few pages back.

(c) **Comparison**

The trees generated by all three methods are the same. This makes sense because all three methods are using a measure of purity to determine the best attribute to split on so we would expect them to give similar results in terms of which attribute gives the most purity.

Question 3

(a) You don't say which method to use so I'm going to use Majority Error since it is the quickest to do by hand. Additionally, for this part I am going to assume that Outlook is Sunny since there is a tie between it and rainy.
$ME(S) = \frac{5}{15}$

O:

S: 6 of 15 examples, $p_+ = 3/6$, $p_- = 3/6$
$ME(O = S) = 1/2$

O: 4 of 15 examples, $p_+ = 4/4$, $p_- = 0$
$ME(O = O) = 0$

R: 5 of 15 examples, $p_+ = 3/5$, $p_- = 2/5$
$ME(O = R) = 2/5$

IG: $\frac{5}{15} - \left(\frac{5}{15} * \frac{1}{2} + \frac{4}{15} * 0 + \frac{5}{15} * \frac{2}{5}\right) = \mathbf{\frac{1}{30}}$

T:

H: 4 of 15 examples, $p_+ = 2/4$, $p_- = 2/4$
$ME(T = H) = 2/4 = 1/2$

M: 7 of 15 examples, $p_+ = 5/7$, $p_- = 2/7$
$ME(T = M) = 2/7$

C: 4 of 15 examples, $p_+ = 3/4$, $p_- = 1/4$
$ME(T = C) = 1/4$

IG: $\frac{5}{15} - \left(\frac{4}{15} * \frac{1}{2} + \frac{7}{15} * \frac{2}{7} + \frac{4}{15} * \frac{1}{4}\right) = \mathbf{0}$

H:

H: 7 of 15 examples, $p_+ = 3/7$, $p_- = 4/7$
$ME(H = H) = 3/7$

N: 8 of 15 examples, $p_+ = 7/8$, $p_- = 1/8$
$ME(H = N) = 1/8$

L: 0 of 15 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L) = 0$

IG: $\frac{5}{15} - \left(\frac{7}{15} * \frac{3}{7} + \frac{8}{15} * \frac{1}{8} + \frac{0}{15} * 0\right) = \mathbf{\frac{1}{15}}$

W:

S: 6 of 15 examples, $p_+ = 3/6$, $p_- = 3/6$
$ME(W = S) = 3/6 = 1/2$

W: 9 of 15 examples, $p_+ = 7/9$, $p_- = 2/9$
$ME(W = W) = 2/9$

IG: $\frac{5}{15} - \left(\frac{6}{15} * \frac{1}{2} + \frac{9}{15} * \frac{2}{9}\right) = \mathbf{0}$

With this extra data the new best splitting point becomes Humidity.

(b) The most common value among the Yes labeled data is Overcast so that is what we will assign to this new point.
$ME(S) = \frac{5}{15}$

O:

S: 5 of 15 examples, $p_+ = 2/5$, $p_- = 3/5$
$ME(O = S) = 2/5$

O: 5 of 15 examples, $p_+ = 5/5$, $p_- = 0$
$ME(O = O) = 0$

R: 5 of 15 examples, $p_+ = 3/5$, $p_- = 2/5$
$ME(O = R) = 2/5$

IG: $\frac{5}{15} - \left(\frac{5}{15} * \frac{2}{5} + \frac{5}{15} * 0 + \frac{5}{15} * \frac{2}{5}\right) = \mathbf{\frac{1}{15}}$

T:

H: 4 of 15 examples, $p_+ = 2/4$, $p_- = 2/4$
$ME(T = H) = 2/4 = 1/2$

M: 7 of 15 examples, $p_+ = 5/7$, $p_- = 2/7$
$ME(T = M) = 2/7$

C: 4 of 15 examples, $p_+ = 3/4$, $p_- = 1/4$
$ME(T = C) = 1/4$

IG: $\frac{5}{15} - \left(\frac{4}{15} * \frac{1}{2} + \frac{7}{15} * \frac{2}{7} + \frac{4}{15} * \frac{1}{4}\right) = \mathbf{0}$

H:

H: 7 of 15 examples, $p_+ = 3/7$, $p_- = 4/7$
$ME(H = H) = 3/7$

N: 8 of 15 examples, $p_+ = 7/8$, $p_- = 1/8$
$ME(H = N) = 1/8$

L: 0 of 15 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L) = 0$

IG: $\frac{5}{15} - \left(\frac{7}{15} * \frac{3}{7} + \frac{8}{15} * \frac{1}{8} + \frac{0}{15} * 0\right) = \mathbf{\frac{1}{15}}$

W:

S: 6 of 15 examples, $p_+ = 3/6$, $p_- = 3/6$
$ME(W = S) = 3/6 = 1/2$

W: 9 of 15 examples, $p_+ = 7/9$, $p_- = 2/9$
$ME(W = W) = 2/9$

IG: $\frac{5}{15} - \left(\frac{6}{15} * \frac{1}{2} + \frac{9}{15} * \frac{2}{9}\right) = \mathbf{0}$

In this case the tree doesn't change from what it originally was, that is that our best splitting attribute for the first step is a tie between Outlook and Humidity.

(c) In order to add this new point we need to know the proportion of each possible outlook. Looking at the data we find the following $Outlook = \{5/14 Sunny, 4/14 Overcast, 5/14 Rain\}$. So we will use those fractions to fill in the missing data.
$ME = 5/15$

O:

S: $5 + \frac{5}{14}$ of 15 examples, $p_+ = (2 + \frac{5}{14})/(5 + \frac{5}{14})$, $p_- = 3/(5 + \frac{5}{14})$
$ME(O = S) = (2 + \frac{5}{14})/(5 + \frac{5}{14})$

O: $4 + \frac{4}{14}$ of 15 examples, $p_+ = (4 + \frac{4}{14}/(4 + \frac{4}{14})$, $p_- = 0$
$ME(O = O) = 0$

R: $5 + \frac{5}{14}$ of 15 examples, $p_+ = (3 + \frac{5}{14})/(5 + \frac{5}{14})$, $p_- = 2/(5 + \frac{5}{14})$
$ME(O = R) = 2/(5 + \frac{5}{14})$

IG: $\frac{5}{15} - \left(\left(\frac{5\frac{5}{14}}{15}\right) * \frac{2\frac{5}{14}}{5\frac{5}{14}} + \left(\frac{4\frac{4}{14}}{15}\right) * 0 + \left(\frac{5\frac{5}{14}}{15}\right) * \frac{2}{5\frac{5}{14}}\right) = \mathbf{0.043}$

T:

H: 4 of 15 examples, $p_+ = 2/4$, $p_- = 2/4$
$ME(T = H) = 2/4 = 1/2$

M: 7 of 15 examples, $p_+ = 5/7$, $p_- = 2/7$
$ME(T = M) = 2/7$

C: 4 of 15 examples, $p_+ = 3/4$, $p_- = 1/4$
$ME(T = C) = 1/4$

IG: $\frac{5}{15} - \left(\frac{4}{15} * \frac{1}{2} + \frac{7}{15} * \frac{2}{7} + \frac{4}{15} * \frac{1}{4}\right) = \mathbf{0}$

H:

H: 7 of 15 examples, $p_+ = 3/7$, $p_- = 4/7$
$ME(H = H) = 3/7$

N: 8 of 15 examples, $p_+ = 7/8$, $p_- = 1/8$
$ME(H = N) = 1/8$

L: 0 of 15 examples, $p_+ = 0$, $p_- = 0$
$ME(H = L) = 0$

IG: $\frac{5}{15} - \left(\frac{7}{15} * \frac{3}{7} + \frac{8}{15} * \frac{1}{8} + \frac{0}{15} * 0\right) = \mathbf{0.067}$

W:

S: 6 of 15 examples, $p_+ = 3/6$, $p_- = 3/6$
$ME(W = S) = 3/6 = 1/2$

W: 9 of 15 examples, $p_+ = 7/9$, $p_- = 2/9$
$ME(W = W) = 2/9$

IG: $\frac{5}{15} - \left(\frac{6}{15} * \frac{1}{2} + \frac{9}{15} * \frac{2}{9}\right) = \mathbf{0}$

This means that humidity is now our best attribute to split on meaning that I'm going to have to continue dealing with that nightmare of fractions...

(d) Using Humidity to split as identified in the previous problem we can continue building the tree as follows.
Available Attributes = $\{O, T, W\}$

H = H: Data = $\{1, 2, 3, 4, 8, 12, 14\}$
ME = $\frac{3}{7}$

O:

S: 3 of 7 examples, $p_+ = 0/3$, $p_- = 3/3$
$ME(O = S|H = H) = 0$

O: 2 of 7 examples, $p_+ = 2/2$, $p_- = 0/2$
$ME(O = O|H = H) = 0$

R: 2 of 7 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(O = R|H = H) = 1/2$

IG: $\frac{3}{7} - \left(\frac{3}{7} * 0 + \frac{2}{7} * 0 + \frac{2}{7} * \frac{1}{2}\right) = \mathbf{0.286}$

T:

H: 3 of 7 examples, $p_+ = 1/3$, $p_- = 2/3$
$ME(T = H|H = H) = 1/3$

M: 4 of 7 examples, $p_+ = 2/4$, $p_- = 2/42$
$ME(T = M|H = H) = 2/4$

C: 0 of 7 examples, $p_+ = 0$, $p_- = 0$
$ME(T = C|H = H) = 0$

IG: $\frac{3}{7} - \left(\frac{3}{7} * \frac{1}{3} + \frac{4}{7} * \frac{2}{4} + \frac{0}{7} * 0\right) = \mathbf{0}$

W:

S: 3 of 7 examples, $p_+ = 1/3$, $p_- = 2/3$
$ME(W = S|H = H) = 1/3$

W: 4 of 7 examples, $p_+ = 2/4$, $p_- = 2/43$
$ME(W = W|H = H) = 2/4$

IG: $\frac{3}{7} - \left(\frac{3}{7} * \frac{1}{3} + \frac{4}{7} * \frac{2}{4}\right) = \mathbf{0}$

Thus this node is best split using Outlook.

H = N: Data = $\{5, 6, 7, 9, 10, 11, 13, 15\}$
ME = $\frac{1}{8}$

O:

S: $2 + \frac{5}{14}$ of 8 examples, $p_+ = (2 + \frac{5}{14})/(2 + \frac{5}{14})$, $p_- = 0$
$ME(O = S|H = N) = 0$

O: $2 + \frac{4}{14}$ of 8 examples, $p_+ = (2 + \frac{4}{14})/(2 + \frac{4}{14})$, $p_- = 0$
$ME(O = O|H = N) = 0$

R: $3 + \frac{5}{14}$ of 8 examples, $p_+ = (2 + \frac{5}{14})/(3 + \frac{5}{14})$, $p_- = 1/(3 + \frac{5}{14})$
$ME(O = R|H = N) = 1/(3 + \frac{5}{14})$

IG: $\frac{1}{8} - \left(\frac{2\frac{5}{14}}{8} * 0 + \frac{2\frac{4}{14}}{8} * 0 + \frac{3\frac{5}{14}}{8} * \frac{1}{3\frac{5}{14}}\right) = \mathbf{0}$

T:

H: 1 of 8 examples, $p_+ = 1/1$, $p_- = 0/1$
$ME(T = H|H = N) = 0$

M: 3 of 8 examples, $p_+ = 3/3$, $p_- = 0/3$
$ME(T = M|H = N) = 0$

C: 4 of 8 examples, $p_+ = 3/4$, $p_- = 1/4$
$ME(T = C|H = N) = 1/4$

IG: $\frac{1}{8} - \left(\frac{1}{8} * 0 + \frac{3}{8} * 0 + \frac{4}{8} * \frac{1}{4}\right) = \mathbf{0}$

W:

S: 3 of 8 examples, $p_+ = 2/3$, $p_- = 1/3$
$ME(W = S|H = N) = 1/3$

W: 5 of 8 examples, $p_+ = 5/5$, $p_- = 0/5$
$ME(W = W|H = N) = 0$
IG: $\frac{1}{8} - \left(\frac{3}{8} * \frac{1}{3} + \frac{5}{8} * 0\right) = \mathbf{0}$

Either I did something wrong, which I would not rule out given how tedious this is to do by hand, or they're all equally bad choices and it doesn't matter what we chose. Hoping it's the second case I'm going to choose to split on Outlook so I don't have to deal with the fractions anymore.

H = L: Data = {}
There are no training examples with this value so we will make this a leaf node with the label of the majority of it's parent (which is the full dataset for this tree). In this case $play = Yes$.

Level Three Branch One Splitting on Outlook given Humidity = High
Available Attributes = $\{T, W\}$

O = S: Data = $\{1, 2, 8\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

O = O: Data = $\{3, 12\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

O = R: Data = $\{4, 14\}$
ME = $\frac{1}{2}$

T:

H: 0 of 2 examples, $p_+ = 0$, $p_- = 0$
$ME(T = H|O = R, H = H) = 0$

M: 2 of 2 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(T = M|O = R, H = H) = 1/2$

C: 0 of 2 examples, $p_+ = 0$, $p_- = 0$
$ME(T = C|O = R, H = H) = 0$

IG: $\frac{1}{2} - \left(\frac{0}{2} * 0 + \frac{2}{2} * \frac{1}{2} + \frac{0}{2} * 0\right) = \mathbf{0}$

W:

S: 1 of 2 examples, $p_+ = 0/1$, $p_- = 1/1$
$ME(W = S|O = R, H = H) = 0$

W: 1 of 2 examples, $p_+ = 1/1$, $p_- = 0/1$
$ME(W = W|O = R, H = H) = 0$

IG: $\frac{1}{2} - \left(\frac{1}{2} * 0 + \frac{1}{2} * 0\right) = \frac{1}{2}$

So this node should further be split on Wind.

Level Three Branch Two Splitting on Outlook given Humidity = Normal
Available Attributes = $\{T, W\}$

O = S: Data = $\{9, 11, 15\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

O = O: Data = $\{7, 13, 15\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

O = R: Data = $\{5, 6, 10, 15\}$
ME = $\frac{1}{3\frac{5}{14}}$

T:

H: 0 of 4 examples, $p_+ = 0$, $p_- = 0$
$ME(T = H|O = R, H = N) = 0$

M: $1 + \frac{5}{14}$ of 4 examples, $p_+ = (1 + \frac{5}{14})/2$, $p_- = 0$
$ME(T = M|O = R, H = N) = 0$

C: 2 of 4 examples, $p_+ = 1/2$, $p_- = 1/2$
$ME(T = C|O = R, H = N) = 1/2$

IG: $\frac{1}{3\frac{5}{14}} - \left(\frac{0}{4} * 0 + \frac{1\frac{5}{14}}{4} * 0 + \frac{2}{4} * \frac{1}{2}\right) = \mathbf{0.048}$

W:

S: 1 of 4 examples, $p_+ = 0$, $p_- = 1/1$
$ME(W = S|O = R, H = N) = 0$

W: 3 of 4 examples, $p_+ = 3/3$, $p_- = 0$
$ME(W = W|O = R, H = N) = 0$

IG: $\frac{1}{3\frac{5}{14}} - \left(\frac{1}{4} * 0 + \frac{3}{4} * 0\right) = \frac{1}{3\frac{5}{14}}$

So this node should also be further split on Wind.

Level Four Branch One Splitting on Wind given Outlook = Rain and Humidity = High
Available Attributes = $\{T\}$

W = S: Data = $\{14\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

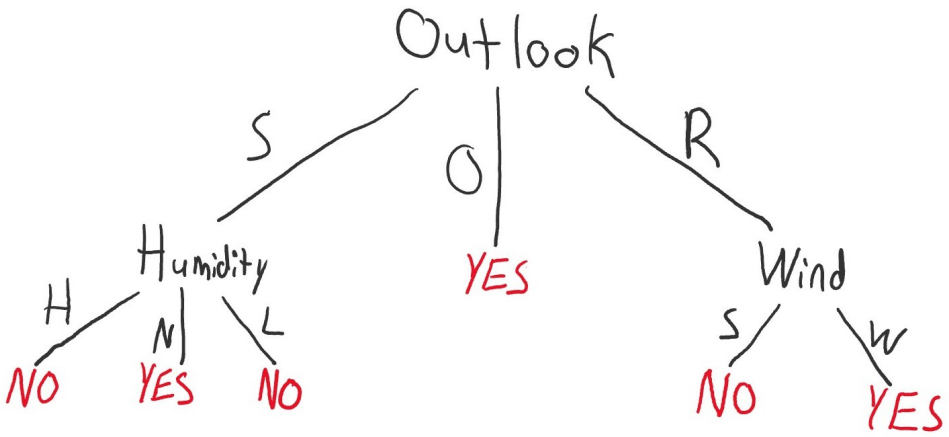
W = W: Data = $\{4\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

Level Four Branch Two Splitting on Wind given Outlook = Rain and Humidity = Normal
Available Attributes = $\{T\}$

W = S: Data = $\{6\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = No$.

W = W: Data = $\{5, 10, 15\}$
ME = 0
This node uniquely identifies each instance and as such becomes a leaf with $play = Yes$.

All training data is labeled at this point and the tree stops here. A image of the tree can be found below.



Part Two

1. Link to my github is https://github.com/Tweal/CS6350.

2. (c) As can be seen in the table below the accuracy of each method increases as we increase the max depth of the tree. This makes sense since the tree gets more options to classify on. Additionally we can see that the Max Error method preforms slightly worse than the other two. Again this kind of makes sense because it takes into account the least amount of information and is the easiest to implement.

Training accuracy:

	1	2	3	4	5	6
Entropy	0.698	0.778	0.819	0.918	0.973	1.0
Max Error	0.698	0.708	0.807	0.889	0.964	1.0
Gini Index	0.698	0.778	0.824	0.911	0.973	1.0

Test accuracy:

	1	2	3	4	5	6
Entropy	0.703	0.777	0.804	0.849	0.916	0.916
Max Error	0.703	0.687	0.788	0.821	0.887	0.887
Gini Index	0.703	0.777	0.816	0.863	0.916	0.916

3. (c) Comparing the training errors and testing errors across the various max depths is kind of interesting in this case. We see that with all methods the test data actually starts getting slightly worse with deeper trees. I'm assuming this has to do with over fitting the training data since the training data gets better with more levels, which makes sense. Additionally, I find it interesting that there isn't much of a difference when replacing unknowns vs using them as a label. If I had to venture a guess I'd say it's because of the sheer amount of data that we have, 5000 is quite a lot in my opinion.

Figure 1: Using 'unknown' as a attribute value

Training accuracy:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Entropy	0.881	0.894	0.899	0.921	0.939	0.953	0.965	0.971	0.977	0.983	0.986	0.986	0.986	0.986	0.986	0.986
Max Error	0.891	0.896	0.904	0.918	0.928	0.933	0.936	0.941	0.949	0.957	0.962	0.968	0.972	0.980	0.985	0.986
Gini Index	0.891	0.896	0.907	0.925	0.940	0.953	0.965	0.973	0.979	0.983	0.985	0.986	0.986	0.986	0.986	0.986

Test accuracy:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Entropy	0.875	0.889	0.893	0.886	0.880	0.870	0.860	0.856	0.852	0.847	0.846	0.847	0.847	0.847	0.847	0.847
Max Error	0.883	0.891	0.888	0.884	0.885	0.882	0.881	0.879	0.874	0.864	0.860	0.855	0.850	0.845	0.844	0.843
Gini Index	0.883	0.891	0.888	0.880	0.874	0.863	0.851	0.844	0.842	0.838	0.835	0.835	0.835	0.835	0.835	0.835

Figure 2: Using 'unknown' as missing and replacing with majority

Training accuracy:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Entropy	0.881	0.894	0.898	0.913	0.929	0.943	0.955	0.961	0.968	0.974	0.977	0.978	0.978	0.978	0.978	0.978
Max Error	0.891	0.895	0.902	0.914	0.922	0.927	0.931	0.934	0.940	0.944	0.951	0.957	0.964	0.969	0.975	0.978
Gini Index	0.891	0.895	0.899	0.912	0.926	0.943	0.955	0.963	0.971	0.975	0.978	0.978	0.978	0.978	0.978	0.978

Test accuracy:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Entropy	0.875	0.889	0.891	0.885	0.879	0.867	0.860	0.854	0.850	0.846	0.844	0.846	0.846	0.846	0.846	0.846
Max Error	0.883	0.890	0.886	0.883	0.884	0.879	0.877	0.875	0.868	0.865	0.860	0.857	0.852	0.849	0.846	0.845
Gini Index	0.883	0.890	0.892	0.886	0.880	0.869	0.859	0.854	0.848	0.845	0.844	0.845	0.845	0.845	0.845	0.845