HW1

Ben Orkild, u1196243

CS 6350 - September 22, 2023

1 Part 1

1.1 Problem 1A

In problem 1A we will utilize the following equations:

$$Entropy(S) = -p_+ log_2(p_+) - p_- log_2(p_-) \tag{1}$$

$$Gain = Entropy(S) - \Sigma \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

Step 1: Attributes: $[x_1, x_2, x_3, x_4]$
Now we use equations 1 and 2:
Overall Entropy:
$H = \frac{-1}{2} * log_2(2/7) - \frac{-5}{7} * log_2(5/7)$
$S_v = x_1$:
$H_s(x_1 = 1) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$H_s(x_1 = 0) = \frac{-1}{5} * log_2(1/5) - \frac{4}{5} * log_2(4/5) = 0.72$
$Gain = 0.86 - (\frac{2}{7} * 1 + \frac{5}{7} * 0.72) = 0.06$
$S_v = x_2$:
$H_s(x_2 = 1) = \frac{0}{4} * log_2(0/4) - \frac{4}{4} * log_2(4/4) = 0$
$H_s(x_2 = 0) = \frac{-2}{3} * log_2(2/3) - \frac{1}{3} * log_2(1/3) = 0.92$
$Gain = 0.86 - (\frac{4}{7} * 0 + \frac{3}{7} * 0.92) = 0.46$
$S_v = x_3$
$H_s(x_3 = 1) = \frac{-1}{3} * log_2(1/3) - \frac{2}{3} * log_2(2/3)$
$H_s(x_3 = 0) = \frac{-1}{4} * log_2(1/4) - \frac{3}{4} * log_2(3/4)$
$Gain = 0.86 - (\frac{4}{7} * 0.81 + \frac{3}{7} * 0.92) = 0.0029$
$S_v = x_4$
$H_s(x_4 = 1) = \frac{-2}{3} * log_2(2/3) - \frac{1}{3} * log_2(1/3)$
$H_s(x_4 = 0) = \frac{0}{4} * log_2(0/4) - \frac{4}{4} * log_2(4/4)$
$Gain = 0.86 - (\frac{4}{7} * 0 + \frac{3}{7} * 0.92) = 0.46$

Now we select x_2 because it had the highest gain. We then start with the subtree $x_2 = 1$. All labels on this subtree are the same, so we return a leaf-node, $y = 0$.

Generate subtree below $x_2 = 0$:

Attributes: $[x_1, x_3, x_4]$
Overall Entropy: $\frac{-2}{3} * log_2(2/3) - \frac{1}{3} * log_2(1/3) = 0.92$
For $S_v = x_1$
$H_s(x_1 = 1) = \frac{-1}{1} * log_2(1/1) - \frac{0}{1} * log_2(0/1) = 0$
$H_s(x_1 = 0) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$Gain = 0.92 - (\frac{1}{3} * 0 + \frac{2}{3} * 1) = 0.25$
For $S_v = x_3$
$H_s(x_3 = 1) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$H_s(x_3 = 0) = \frac{-1}{1} * log_2(1/1) - \frac{0}{1} * log_2(0/1) = 0$
$Gain = 0.92 - (\frac{1}{3} * 0 + \frac{2}{3} * 1) = 0.25$
For $S_v = x_4$
$H_s(x_4 = 1) = \frac{-2}{2} * log_2(2/2) - \frac{0}{2} * log_2(0/2) = 0$
$H_s(x_4 = 0) = \frac{-1}{1} * log_2(1/1) - \frac{0}{1} * log_2(0/1) = 0$
$Gain = 0.92 - (\frac{1}{3} * 0 + \frac{2}{3} * 0) = 0.92$

Select x_4 as next attribute to split. We then look at $x_4 = 1$, and all labels match, so we assign leaf-node $y = 1$. We then look at $x_4 = 0$, and again all labels match, so we assign the leaf-node $y = 0$. Our tree is now complete.

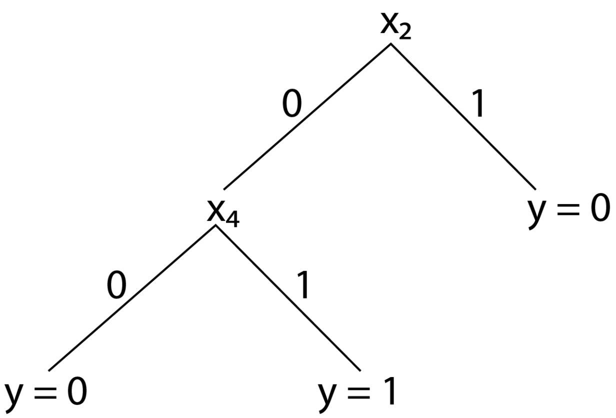

Figure 1: Decision Tree for problem 1A of Part 1.

1.2 Problem 1B

	$x_2 = 0$	$x_2 = 1$
$x_4 = 0$	y = 0	y = 0
$x_4 = 1$	y = 1	y = 0

Table 1: Table describing the function from part A.

1.3 Problem 2A

Attributes: $[Outlook, Temperature, Humidity, Wind]$

Majority Error of whole dataset: $ME(S) = \frac{5}{14}$

$S_v = Outlook$

$ME_s(Outlook = S) = 2/5$

$ME_s(Outlook = O) = 0/4$

$ME_s(Outlook = R) = 2/5$

$Gain = \frac{5}{14} - (\frac{5}{14}\frac{2}{5} + \frac{4}{14}\frac{0}{4} + \frac{5}{14}\frac{2}{5}) = 0.0714$

$S_v = Temp$

$ME_s(Temp = H) = 2/4$

$ME_s(Temp = M) = 2/6$

$ME_s(Temp = C) = 1/4$

$Gain = \frac{5}{14} - (\frac{4}{14}\frac{2}{4} + \frac{6}{14}\frac{2}{6} + \frac{4}{14}\frac{1}{4}) = 0$

$S_v = Hum$

$ME_s(Hum = H) = 3/7$

$ME_s(Hum = N) = 1/7$

$ME_s(Hum = L) = 0$

$Gain = \frac{5}{14} - (\frac{7}{14}\frac{3}{7} + \frac{7}{14}\frac{1}{7}) = 0.0714$

$S_v = Wind$

$ME_s(Wind = S) = 3/6$

$ME_s(Wind = W) = 2/8$

$Gain = \frac{5}{14} - (\frac{6}{14}\frac{3}{6} + \frac{8}{14}\frac{2}{8}) = 0$

We select Humidity as our first attribute. The Low value of Humidity doesn't have any values, so we assign the most common label as a leaf node, which is $play = +$. Now we start the subtree for $Humidity = high$.

Attributes $= [Outlook, Temperature, Wind]$

Overall ME: 3/7

$S_v = Outlook$

$ME_s(Outlook = S) = 0/3$

$ME_s(Outlook = O) = 0/2$

$ME_s(Outlook = R) = 1/2$

$Gain = \frac{3}{7} - (\frac{3}{7}\frac{0}{3} + \frac{2}{7}\frac{0}{2} + \frac{1}{2}\frac{2}{7}) = 0.2857$

$S_v = Temp$

$ME_s(Temp = H) = 1/3$

$ME_s(Temp = M) = 2/4$

$ME_s(Temp = C) = 0$

$Gain = \frac{3}{7} - (\frac{3}{7}\frac{1}{3} + \frac{4}{7}\frac{2}{4}) = 0$

$S_v = Wind$

$ME_s(Wind = S) = 1/3$

$ME_s(Wind = W) = 2/4$

$Gain = \frac{3}{7} - (\frac{3}{7}\frac{1}{3} + \frac{4}{7}\frac{2}{4}) = 0$

We select Outlook as the attribute to split on for the subtree below humidity = high. Now we handle the subtree for $humidity = Norm$:

Overall ME: 1/7

$S_v = Outlook$

$ME_s(Outlook = S) = 0/2$
$ME_s(Outlook = O) = 0/2$
$ME_s(Outlook = R) = 1/3$
$Gain = \frac{1}{7} - (\frac{2}{7}\frac{0}{2} + \frac{2}{7}\frac{0}{2} + \frac{1}{3}\frac{3}{7}) = 0$
$S_v = Temp$
$ME_s(Temp = H) = 0/1$
$ME_s(Temp = M) = 0/2$
$ME_s(Temp = C) = 1/4$
$Gain = \frac{1}{7} - (\frac{1}{7}\frac{0}{1} + \frac{2}{7}\frac{0}{2} + \frac{4}{7}\frac{1}{4}) = 0$
$S_v = Wind$
$ME_s(Wind = S) = 1/3$
$ME_s(Wind = W) = 0/4$
$Gain = \frac{3}{7} - (\frac{3}{7}\frac{1}{3} + \frac{4}{7}\frac{0}{4}) = 0$

All the gains are equal, so we select Wind as the attribute to split on. Now we move to the Outlook node. We see that the Sunny branch has all the same labels, so it is assigned the $play = -$ leaf node. We see that the Overcast branch has all the same label, so we assign the $play = +$ leaf node. We now repeat the algorithm for the Outlook = Rain branch:

Attributes: $[Temperature, Wind]$
Overall ME: 1/2
$S_v = Temp$
$ME_s(Temp = H) = 0$
$ME_s(Temp = M) = 1/2$
$ME_s(Temp = C) = 0$
$Gain = \frac{1}{2} - (\frac{1}{2}\frac{2}{2}) = 0$
$S_v = Wind$
$ME_s(Wind = S) = 0/1$
$ME_s(Wind = W) = 0/1$
$Gain = \frac{1}{2} - (\frac{0}{1}\frac{1}{2} + \frac{0}{1}\frac{1}{2}) = 1/2$

We select Wind as the next attribute. The labels are the same for Wind = Strong, so we assign the leaf node $Play = -$. The labels are the same for Wind = Weak, so we assign the leaf node $Play = +$. We then move to the Wind = Strong node:

Attributes = $[Outlook, Temperature]$
Overall ME: 1/3
$S_v = Outlook$
$ME_s(Outlook = S) = 0/1$
$ME_s(Outlook = O) = 0/1$
$ME_s(Outlook = R) = 0/1$
$Gain = \frac{1}{3} - (\frac{0}{1}\frac{1}{3} + \frac{0}{1}\frac{1}{3} + \frac{0}{1}\frac{1}{3}) = 1/3$
$S_v = Temp$
$ME_s(Temp = H) = 0$
$ME_s(Temp = M) = 0/1$
$ME_s(Temp = C) = 1/2$
$Gain = \frac{1}{3} - (0 + 0 + \frac{2}{3}\frac{1}{2}) = 0$

We select Outlook as our attribute. Each value of Outlook has the same label, so our tree is complete. The completed tree is shown in 2

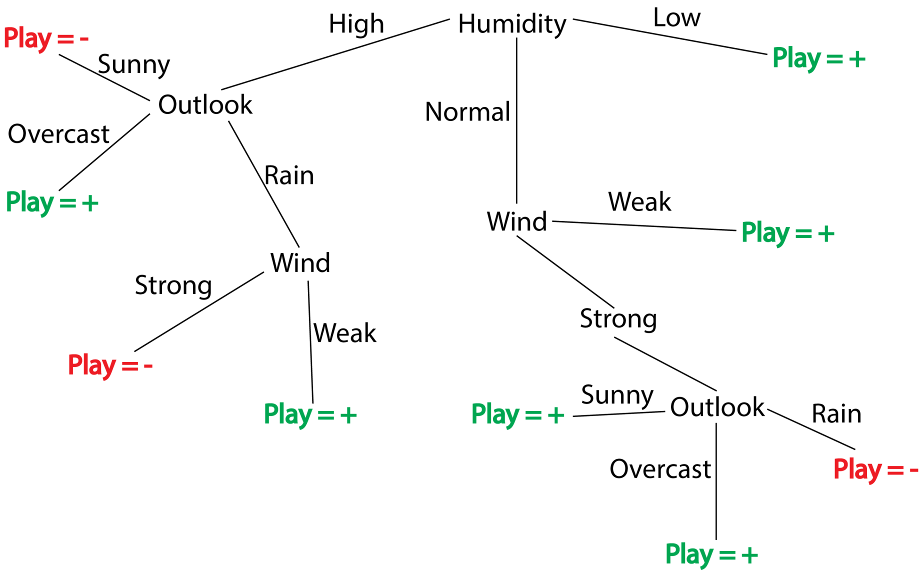

Figure 2: Completed tree for problem 2A

1.4 Problem 2B

The overall gini index is: $GE(S) = 1 - ((5/14)^2 + (9/14)^2) = 0.4592$

Attributes: $[Outlook, Temperature, Humidity, Wind]$
$S_v = Outlook$
$GI_s(Outlook = S) = 1 - ((3/5)^2 + (2/5)^2) = 0.48$
$GI_s(Outlook = O) = 1 - ((0/4)^2 + (4/4)^2) = 0$
$GI_s(Outlook = R) = 1 - ((3/5)^2 + (2/5)^2) = 0.48$
$Gain = 0.4592 - (\frac{5}{14}0.48 + \frac{4}{14}0 + \frac{5}{14}0.48) = 0.1163$
$S_v = Temp$
$GI_s(Temp = H) = 1 - ((2/4)^2 + (2/4)^2) = 0.50$
$GI_s(Temp = M) = 1 - ((4/6)^2 + (2/6)^2) = 0.44$
$GI_s(Temp = C) = 1 - ((1/4)^2 + (3/4)^2) = 0.375$
$Gain = 0.4592 - (\frac{4}{14}0.50 + \frac{6}{14}0.44 + \frac{4}{14}0.375) = 0.0187$
$S_v = Hum$
$GI_s(Hum = H) = 1 - ((4/7)^2 + (3/7)^2) = 0.4898$
$GI_s(Hum = N) = 1 - ((1/7)^2 + (6/7)^2) = 0.2449$
$GI_s(Hum = L) = 0$
$Gain = 0.4592 - (\frac{7}{14}0.4898 + \frac{7}{14}0.2449 + \frac{0}{14}0) = 0.018$
$S_v = Wind$
$GI_s(Wind = S) = 1 - ((3/6)^2 + (3/6)^2) = 0.5$

$GI_s(Wind = W) = 1 - ((2/8)^2 + (6/8)^2) = 0.375$

$Gain = 0.4592 - (\frac{6}{14}0.5 + \frac{8}{14}0.375) = 0.0306$

We select Outlook as the attribute for the first split. We see that overcast all have the same label, so it is assigned a leaf node of $play = +$. Now we start the outlook = sunny subtree:

Attributes: $[Temperature, Humidity, Wind]$

Overall GI: $1 - ((3/5)^2 + (2/5)^2) = 0.48$

$S_v = Temp$

$GI_s(Temp = H) = 1 - ((2/2)^2 + (0/2)^2) = 0$

$GI_s(Temp = M) = 1 - ((1/2)^2 + (1/2)^2) = 0.50$

$GI_s(Temp = C) = 1 - ((0/1)^2 + (1/1)^2) = 0$

$Gain = 0.48 - (\frac{2}{5}0 + \frac{1}{5}0.5 + \frac{2}{5}0.) = 0.28$

$S_v = Hum$

$GI_s(Hum = H) = 1 - ((3/3)^2 + (0/3)^2) = 0$

$GI_s(Hum = N) = 1 - ((0/2)^2 + (2/2)^2) = 0$

$GI_s(Hum = L) = 0$

$Gain = 0.48 - (\frac{3}{5}0 + \frac{2}{5}0 + \frac{0}{5}0) = 0.48$

$S_v = Wind$

$GI_s(Wind = S) = 1 - ((1/2)^2 + (1/2)^2) = 0.5$

$GI_s(Wind = W) = 1 - ((2/3)^2 + (1/3)^2) = 0.444$

$Gain = 0.48 - (\frac{2}{5}0.5 + \frac{3}{5}0.444) = 0.013$

Humidity is selected as the next attribute to split. Each of the branches of Humidity have the same label, leaf nodes are assigned to them. We now move to the Outlook = Rain subtree:

Attributes: $[Temperature, Humidity, Wind]$

Overall GI: $1 - ((3/5)^2 + (2/5)^2) = 0.48$

$S_v = Temp$

$GI_s(Temp = H) = 0$

$GI_s(Temp = M) = 1 - ((1/3)^2 + (2/3)^2) = 0.444$

$GI_s(Temp = C) = 1 - ((1/2)^2 + (1/2)^2) = 0.50$

$Gain = 0.48 - (\frac{0}{5}0 + \frac{3}{5}0.444 + \frac{2}{5}0.5) = 0.0134$

$S_v = Hum$

$GI_s(Hum = H) = 1 - ((1/2)^2 + (1/2)^2) = 0.50$

$GI_s(Hum = N) = 1 - ((1/3)^2 + (2/3)^2) = 0.444$

$GI_s(Hum = L) = 0$

$Gain = 0.48 - (\frac{2}{5}0.5 + \frac{3}{5}0.444 + \frac{0}{5}0) = 0.0134$

$S_v = Wind$

$GI_s(Wind = S) = 1 - ((2/2)^2 + (0/2)^2) = 0$

$GI_s(Wind = W) = 1 - ((0/3)^2 + (3/3)^2) = 0$

$Gain = 0.48 - (\frac{2}{5}0 + \frac{3}{5}0) = 0.48$

We select Wind as the next attribute to split. The set Wind = Strong all have the same label, so a leaf node is created. The set Wind = Weak also all have the same label. The tree is now constructed, as shown in 3.

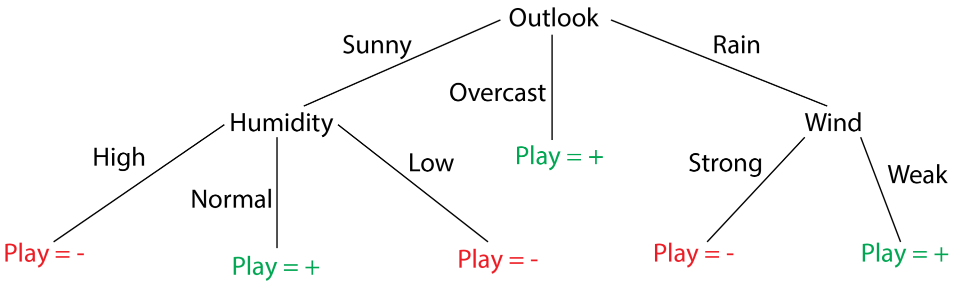

Figure 3: Final tree for problem 2B.

1.5 Problem 2C

The tree in problem 2B matches the one in the slides. The tree in part 2A does not match the one in the slides. The difference is due to the change in the gain function. Majority error gives multiple possible attributes to split on the first node, namely outlook or humidity. We selected humidity for our first split, which likely caused the difference.

1.6 Problem 3A

$H(S) = \frac{-10}{15} * log_2(10/15) - \frac{5}{15} * log_2(5/15) = 0.9183$

$S_v = Outlook$

$H(Out = S) = \frac{-3}{6} * log_2(3/6) - \frac{3}{6} * log_2(3/6) = 1$

$H(Out = O) = \frac{-4}{4} * log_2(4/4) - \frac{0}{4} * log_2(0/4) = 0$

$H(Out = R) = \frac{-2}{5} * log_2(2/5) - \frac{3}{5} * log_2(3/5) = 0.9710$

$Gain = 0.9183 - (\frac{6}{15}1 + \frac{4}{15}0 + \frac{5}{14}0.9710) = 0.1946$

$S_v = Temp$

$H(Temp = H) = \frac{-2}{4} * log_2(2/4) - \frac{2}{4} * log_2(2/4) = 1$

$H(Temp = M) = \frac{-5}{7} * log_2(5/7) - \frac{2}{7} * log_2(2/7) = 0.8631$

$H(Temp = C) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$

$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$

$S_v = Humidity$

$H(Hum = H) = \frac{-4}{7} * log_2(4/7) - \frac{3}{7} * log_2(3/7) = 0.9852$

$H(Hum = N) = \frac{-7}{8} * log_2(7/8) - \frac{1}{8} * log_2(1/8) = 0.5436$

$H(Hum = L) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$

$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$

$S_v = Wind$

$H(Wind = S) = \frac{-3}{6} * log_2(3/6) - \frac{3}{6} * log_2(3/6) = 1$

$H(Wind = W) = \frac{-7}{9} * log_2(7/9) - \frac{2}{9} * log_2(2/9) = 0.7642$

$Gain = 0.9183 - (\frac{6}{15}1 + \frac{9}{15}0.7642) = 0.0598$

Select Outlook as the best feature.

1.7 Problem 3B

$H(S) = \frac{-10}{15} * log_2(10/15) - \frac{5}{15} * log_2(5/15) = 0.9183$

$S_v = Outlook$

$H(Out = S) = \frac{-3}{5} * log_2(3/5) - \frac{2}{5} * log_2(2/5) = 0.9710$

$H(Out = O) = \frac{-5}{5} * log_2(5/5) - \frac{0}{5} * log_2(0/5) = 0$

$H(Out = R) = \frac{-2}{5} * log_2(2/5) - \frac{3}{5} * log_2(3/5) = 0.9710$

$Gain = 0.9183 - (\frac{5}{15}0.9710 + \frac{5}{15}0 + \frac{5}{14}0.9710) = 0.2710$

$S_v = Temp$

$H(Temp = H) = \frac{-2}{4} * log_2(2/4) - \frac{2}{4} * log_2(2/4) = 1$

$H(Temp = M) = \frac{-5}{7} * log_2(5/7) - \frac{2}{7} * log_2(2/7) = 0.8631$

$H(Temp = C) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$

$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$

$S_v = Humidity$

$H(Hum = H) = \frac{-4}{7} * log_2(4/7) - \frac{3}{7} * log_2(3/7) = 0.9852$

$H(Hum = N) = \frac{-7}{8} * log_2(7/8) - \frac{1}{8} * log_2(1/8) = 0.5436$

$H(Hum = L) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$

$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$

$S_v = Wind$

$H(Wind = S) = \frac{-3}{6} * log_2(3/6) - \frac{3}{6} * log_2(3/6) = 1$

$H(Wind = W) = \frac{-7}{9} * log_2(7/9) - \frac{2}{9} * log_2(2/9) = 0.7642$

$Gain = 0.9183 - (\frac{6}{15}1 + \frac{9}{15}0.7642) = 0.0598$

Select Outlook as the best feature.

1.8 Problem 3C

$H(S) = \frac{-10}{15} * log_2(10/15) - \frac{5}{15} * log_2(5/15) = 0.9183$

$S_v = Outlook$

Outlook = Sunny

$p_+ = (2 + \frac{5}{14})/(5 + \frac{5}{14}) = 0.44$

$p_- = (3)/(5 + \frac{5}{14}) = 0.56$

$H(Out = S) = 0.44 * log_2(0.44) - 0.56 * log_2(0.56) = 0.9896$

Outlook = Overcast

$p_+ = (4 + \frac{4}{14})/(4 + \frac{4}{14}) = 1$

$p_- = (0)/(4 + \frac{4}{14}) = 0$

$H(Out = O) = -1 * log_2(1) - 0 * log_2(0) = 0$

Outlook = Rain

$p_+ = (3 + \frac{5}{14})/(5 + \frac{5}{14}) = 0.6267$

$p_- = (2)/(5 + \frac{5}{14}) = 0.3733$

$H(Out = R) = -0.6267 * log_2(0.6267) - 0.3733 * log_2(0.3733) = 0.9532$

$Gain = 0.9183 - (\frac{5}{15}0.9896 + \frac{5}{15}0 + \frac{5}{14}0.9532) = 0.2244$

$S_v = Temp$

$H(Temp = H) = \frac{-2}{4} * log_2(2/4) - \frac{2}{4} * log_2(2/4) = 1$

$H(Temp = M) = \frac{-5}{7} * log_2(5/7) - \frac{2}{7} * log_2(2/7) = 0.8631$

$H(Temp = C) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$

$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$

$S_v = Humidity$
$H(Hum = H) = \frac{-4}{7} * log_2(4/7) - \frac{3}{7} * log_2(3/7) = 0.9852$
$H(Hum = N) = \frac{-7}{8} * log_2(7/8) - \frac{1}{8} * log_2(1/8) = 0.5436$
$H(Hum = L) = \frac{-1}{4} * log_2(2/5) - \frac{3}{4} * log_2(3/4) = 0.81130$
$Gain = 0.9183 - (\frac{4}{15}1 + \frac{7}{15}0.8631 + \frac{4}{14}0.8113) = 0.0325$
$S_v = Wind$
$H(Wind = S) = \frac{-3}{6} * log_2(3/6) - \frac{3}{6} * log_2(3/6) = 1$
$H(Wind = W) = \frac{-7}{9} * log_2(7/9) - \frac{2}{9} * log_2(2/9) = 0.7642$
$Gain = 0.9183 - (\frac{6}{15}1 + \frac{9}{15}0.7642) = 0.0598$
Select Outlook as the best feature.

1.9 Problem 3D

Outlook was selected as best feature. Outlook = Overcast all have the same label, so it is assigned leaf node play = +. Now we process the Sunny Subtree:
$H(S) = 0.9896$
$S_v = Temp$
$H(Temp = H) = \frac{-0}{2} * log_2(0/2) - \frac{2}{2} * log_2(2/2) = 0$
$H(Temp = M) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$H(Temp = C) = 0$
$Gain = 0.9896 - (\frac{2}{5}1) = 0.5896$
$S_v = Humidity$
$H(Hum = H) = \frac{0}{3} * log_2(0/3) - \frac{3}{3} * log_2(3/3) = 0$
$H(Hum = N) = \frac{0}{2} * log_2(0/2) - \frac{2}{2} * log_2(2/2) = 0$
$H(Hum = L) = 0$
$Gain = 0.9896 - 0 = 0.9896$

We select Humidity without calculating the entropy for wind as humidity has the maximum possible value. All branches of humidity have the same label, so all are given leaf nodes. Now we handle Outlook = Rain:
$H(S) = 0.9532$
$S_v = Temp$
$H(Temp = H) = 0$
$H(Temp = M) = \frac{-2}{3} * log_2(2/3) - \frac{1}{3} * log_2(1/3) = 0.9183$
$H(Temp = C) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$Gain = 0.9532 - (\frac{3}{5}0.9183 + \frac{2}{5}1) = 0.0022$
$S_v = Humidity$
$H(Hum = H) = \frac{-1}{2} * log_2(1/2) - \frac{1}{2} * log_2(1/2) = 1$
$H(Hum = N) = \frac{-2}{3} * log_2(2/3) - \frac{1}{3} * log_2(1/3) = 0.9183$
$H(Hum = L) = 0$
$Gain = 0.9532 - (\frac{3}{5}0.9183 + \frac{2}{5}1) = 0.0022$
$S_v = Wind$
$H(Wind = S) = \frac{0}{3} * log_2(0/3) - \frac{3}{3} * log_2(3/3) = 0$
$H(Wind = W) = \frac{0}{2} * log_2(0/2) - \frac{2}{2} * log_2(2/2) = 0$
$Gain = 0.9532 - 0 = 0.9532$

We select wind as the attribute, and all labels are the same for each branch, so we return the tree shown in figure 4.

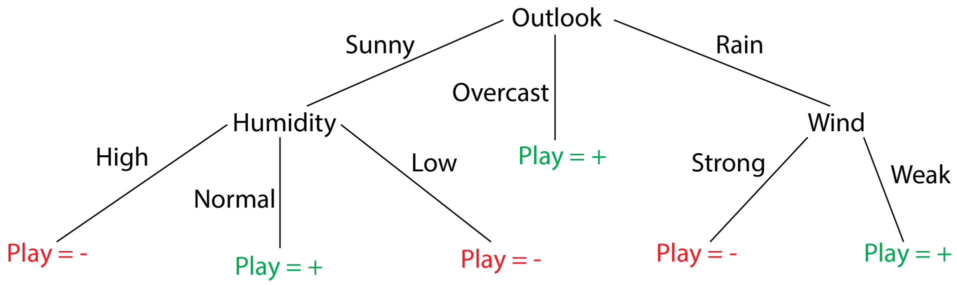

Figure 4: Final tree for problem 3D.

2 Part 2

2.1 Problem 1

Link to github repo: https://github.com/borkild/CS6350.git

2.2 Part 2A

My implementation of the ID3 algorithm can be found in the script $DT_practice.py$. It is a function called ID3, which takes () as inputs and outputs the resulting decision tree.

2.3 Part 2B

depth	train	test
1	69.7%	70.33 %
2	69.7%	70.33%
3	77.7%	77.74%
4	81.8%	80.36%
5	90.9%	84.20%
6	93.9%	86.40%

Table 2: Training and Testing accuracy using Information Gain.

depth	train	test
1	69.7%	70.33%
2	69.7%	70.33%
3	70.7%	68.68%
4	81.9%	80.77%
5	91.3%	84.75%
6	95.9%	89.01%

Table 3: Training and Testing accuracy using Majority Error Gain.

depth	train	test
1	69.7%	70.33%
2	69.7%	70.33%
3	77.7%	77.75%
4	82.3%	81.59%
5	91.0%	86.13%
6	95.8%	89.15%

Table 4: Training and Testing accuracy using Gini Index Gain.

2.4 Part 2C

Our model begins to show signs of overfitting when we hit a depth of 6. Going from a depth of 5 to 6 we see a bigger increase in the training accuracy than in the testing accuracy. If we were to increase the depth beyond 6, we would expect to see the training accuracy increase, but the testing accuracy start to level off, or even decrease.

2.5 Part 3A

depth	train	test
1	88.06%	87.5%
2	88.06%	87.5%
3	89.38%	88.84%
4	90.02%	89.46%
5	91.54%	89.16%
6	93.2%	88.36%
7	94.72%	87.58%
8	95.88%	87.06%
9	97.0%	85.86%
10	97.66%	85.78%
11	98.04%	85.44%
12	98.42%	85.02%
13	98.6%	84.9%
14	98.66%	84.82%
15	98.66%	84.82%
16	98.66%	84.82%

Table 5: Training and Testing accuracy using Information Gain and leaving unknowns in the dataset.

depth	train	test
1	88.06%	87.05%
2	89.1%	88.32%
3	89.56%	89.1%
4	90.34%	88.66%
5	92.12%	88.16%
6	93.3%	87.52%
7	94.22%	87.2%
8	95.16%	86.22%
9	96.02%	85.46%
10	96.7%	84.78%
11	97.36%	84.26%
12	97.88%	83.94%
13	98.16%	83.66%
14	98.44%	83.48%
15	98.66%	83.46%
16	98.66%	83.46%

Table 6: Training and Testing accuracy using Majority Error Gain and leaving unknowns in the dataset.

depth	train	test
1	88.06%	87.5%
2	89.1%	88.32%
3	89.56%	89.1%
4	90.34%	88.66%
5	92.12%	88.16%
6	93.3%	87.52%
7	95.14%	86.34%
8	96.34%	85.46%
9	97.3%	84.92%
10	97.82%	84.78%
11	98.24%	84.5%
12	98.58%	84.12%
13	98.62%	84.14%
14	98.66%	84.04%
15	98.66%	84.04%
16	98.66%	84.04%

Table 7: Training and Testing accuracy using Gini Index Gain and leaving unknowns in the dataset.

2.6 Part 3B

depth	train	test
1	88.06%	87.5%
2	88.06%	87.5%
3	89.38%	88.84%
4	89.84%	89.28%
5	90.98%	88.74%
6	92.74%	87.76%
7	94.06%	87.24%
8	95.3%	86.6%
9	96.24%	85.96%
10	96.88%	85.6%
11	97.52%	85.28%
12	97.86%	85.16%
13	98.16%	85.02%
14	98.18%	85.02%
15	98.18%	85.02%
16	98.18%	85.02%

Table 8: Training and Testing accuracy using Information Gain and removing unknowns from the dataset.

depth	train	test
1	88.06%	87.5%
2	89.1%	88.32%
3	89.48%	88.96%
4	90.22%	88.62%
5	91.28%	88.44%
6	92.58%	87.68%
7	93.7%	86.96%
8	94.82%	86.06%
9	95.04%	85.64%
10	95.88%	85.44%
11	96.52%	84.88%
12	96.98%	84.7%
13	97.28%	84.48%
14	97.72%	84.1%
15	98.1%	83.84%
16	98.18%	83.76%

Table 9: Training and Testing accuracy using Majority Error Gain and removing unknowns from the dataset.

depth	train	test
1	88.06%	87.5%
2	89.1%	88.32%
3	89.46%	88.94%
4	89.88%	89.16%
5	91.16%	88.44%
6	92.78%	87.86%
7	94.36%	87.02%
8	95.54%	86.46%
9	96.44%	86.26%
10	97.02%	85.82%
11	97.62%	85.5%
12	97.98%	85.56%
13	98.14%	85.4%
14	98.18%	85.4%
15	98.18%	85.4%
16	98.18%	85.4%

Table 10: Training and Testing accuracy using Gini Index Gain and removing unknowns from the dataset.

2.7 Part 3C

Removing the unknowns from the dataset did not seem to help increase our decision tree accuracy. This means that the unknowns likely belong to a different category than the most common one. As we increased the tree depth, we saw the training and testing accuracy initially increase, then after a depth of 5, the testing accuracy decreased while the training accuracy increased. This indicates that our model was overfitting to the training data at depths greater than 5.