CS 4375 Assignment 2 Due: Oct. 12, 2019 Jaemin Lee

1 a) $H(\text{Classification}) = H(H, I, S, B) = H(1,2,2,2) = -\frac{1}{7}\log_2\frac{1}{7} - \left(\frac{2}{7}\log_2\frac{2}{7}\right)\times 3 = \boxed{1.95}$

(Classification = C)

$P(H) = \frac{1}{7}$, $P(I) = P(S) = P(B) = \frac{2}{7}$

$H(C|\text{Fever}) = P(\text{No})\cdot H(1,0,0,1) + P(\text{avg})\cdot H(0,1,1,1) + P(\text{high})\cdot H(0,1,1,0)$

$= \left(\frac{2}{7}\right)\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \left(\frac{3}{7}\right)\left[\left(-\frac{1}{3}\log_2\frac{1}{3}\right)\times 3\right] + \left(\frac{2}{7}\right)\left[\left(-\frac{1}{2}\log_2\frac{1}{2}\right)\times 2\right]$

$= \frac{2}{7} + 0.6793 + \frac{2}{7} = \boxed{1.25}$

$IG(C|\text{Fever}) = H(C) - H(C|\text{Fever}) = 1.95 - 1.25 = \boxed{0.7}$

$H(C|\text{Vomit}) = P(Y)\cdot H(0,0,1,2) + P(N)\cdot H(1,2,1,0)$

$= \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{4}{7}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) = 0.3936 + \frac{6}{7} = \boxed{1.25}$

$IG(C|\text{Vomit}) = H(C) - H(C|\text{Vomit}) = 1.95 - 1.25 = \boxed{0.7}$

$H(Y|\text{Diarrhea}) = P(Y)\cdot H(0,0,2,2) + P(N)\cdot H(1,2,0,0)$

(Diarrhea = Diar)

$= \frac{4}{7}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{3}{7}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = \frac{4}{7} + 0.39355 = \boxed{0.9649}$

$IG(C|\text{Diar}) = 1.95 - 0.9649 = \boxed{0.985}$

$H(C|\text{Shiver}) = P(Y)\cdot H(0,1,0,0) + P(N)\cdot H(1,1,2,2)$

$= 0 + \frac{6}{7}\left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) = \boxed{1.644}$

$IG(C|\text{Shiver}) = 1.95 - 1.644 = \boxed{0.306}$

$\therefore$ $IG(C|\text{Diarrhea})$ is the highest $\Rightarrow$ set it as a root node (* D = Diarrhea)

$H_{D_{yes}} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = \boxed{1}$

$IG(D_{yes}|\text{Fever}) = 1 - \left[\frac{1}{4}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right) + \frac{2}{4}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{4}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right)\right]$

$= 1 - [0 + \frac{2}{4} + 0] = 1 - \frac{2}{4} = \frac{2}{4} = \boxed{\frac{1}{2}}$

$IG(D_{yes}|\text{Vomit}) = 1 - \left[\frac{3}{4}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{1}{4}\left(-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right)\right] = 1 - \left[\frac{3}{4}(0.92) + \frac{1}{4}(0)\right] = \boxed{0.31}$

$IG(D_{yes}|\text{Shiver}) = 1 - \left[\frac{0}{4} + \frac{4}{4}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)\right] = 1 - (0+1) = \boxed{0}$

$\therefore$ $IG(D_{yes}|\text{Fever})$ is the highest $\Rightarrow$ set it as a left child of the root node

$H_{D_{no}} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = \boxed{0.92}$

$IG(D_{no}|\text{Fever}) = 0.92 - \left[\frac{1}{3}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{3}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right)\right]$

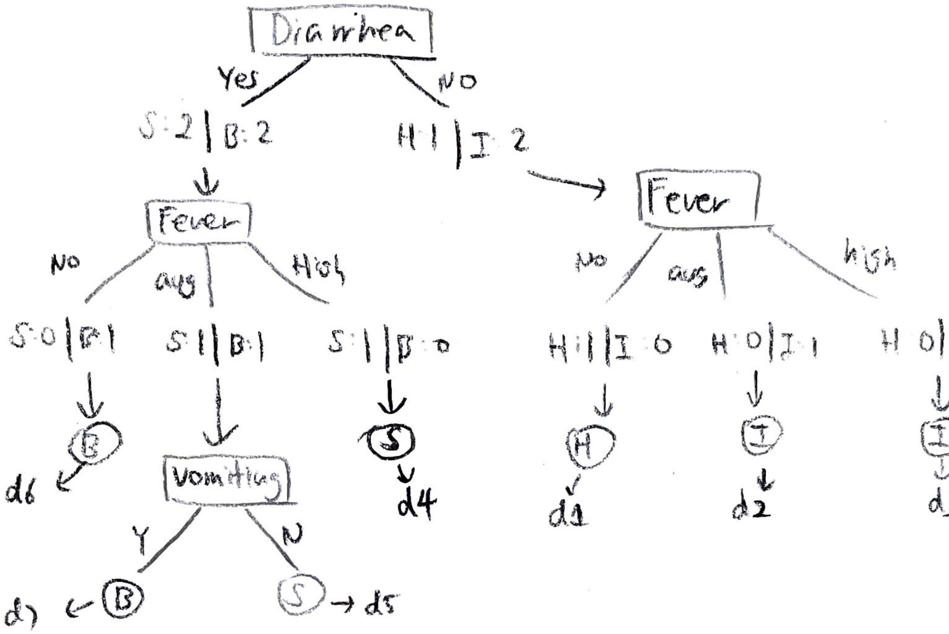
$= 0.92 - [0+0+0] = \boxed{0.92}$

$IG(D_{no}|\text{Vomit}) = 0.92 - \left[\frac{0}{3}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{3}{3}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)\right] = 0.92 - [0 + 0.92] = \boxed{0}$

$IG(D_{no}|\text{Shiver}) = 0.92 - \left[\frac{1}{3}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)\right] = 0.92 - [0 + \frac{2}{3}] = \boxed{0.25}$

$\therefore$ $IG(D_{no}|\text{Fever})$ is the highest $\Rightarrow$ set it as a right child of the root node

For the final branch for when fever = "avg", **Vomiting** is a perfect classifier that splits 2 classes (S and B), whereas Shivering doesn't. Thus use vomiting as the child node for fever = "avg"



1. b) The above decision tree represents that there is no overlapping leaf (they all are disjoint). Therefore, the resulting decision provides a disjoint definition of the classes

2. a) $H(Y) = -\frac{6}{11}\log_2\frac{6}{11} - \frac{5}{11}\log_2\frac{5}{11} = \boxed{0.99}$

$IG(Y|X1) = 0.99 - \left[\left(\frac{5}{11}\right)\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{6}{11}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right)\right]$

$= 0.99 - \left[0.4413 + \frac{6}{11}\right] = \boxed{0.0032}$

$IG(Y|X2) = 0.99 - \left[\frac{5}{11}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{11}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)\right]$

$= 0.99 - \left(0.4413 + \frac{4}{11}\right) = \boxed{0.185}$

$IG(Y|X3) = 0.99 - \left[\frac{5}{11}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)\right] = 0.99 - 0.4413 = \boxed{0.5487}$

$\therefore$ $IG(Y|X3)$ is the highest $\Rightarrow$ Set X3 as the root node

$H_{X3_v} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = \boxed{0.97}$

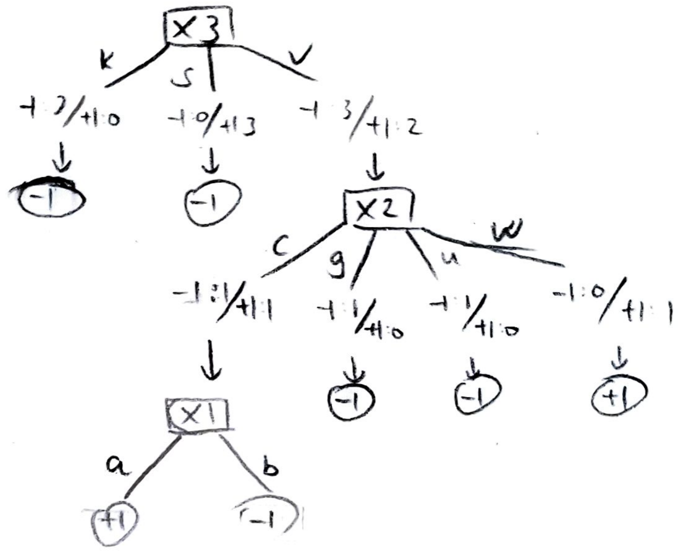
$IG(X1|X3_v) = 0.97 - \left[\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)\right] = 0.97 - \left[\frac{2}{5} + \frac{3}{5}(0.551)\right]$

$= \boxed{0.239}$

$IG(X2|X3_v) = 0.97 - \left[\frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{5}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right) + \frac{1}{5}\left(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right)\right.$
$\left. + \frac{1}{5}\left(\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right)\right] = 0.97 - \left[\frac{2}{5}(1) + 0 + 0 + 0\right] = \boxed{0.57}$

$IG(X2|X3_v)$ is the highest $\Rightarrow$ Use X2 as the child node for X3 = "v"

X1 is the perfect classifier for when X2 = "C". Thus, X1 is the child node for X2 = "c"

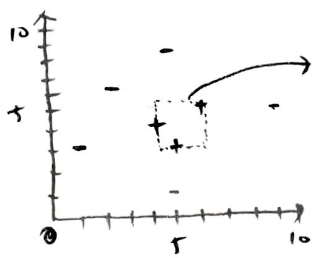


2(b) k => -1 , S => -1 , S => +1 , V => W => +1

Correct, wrong, Correct, correct

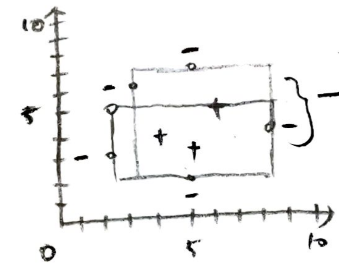
Accuracy = $\frac{3}{4}$ = 75%

3. a)



$\rightarrow S = \{4 \le x \le 6,\ 3 \le y \le 5\}$ or $G = \{(4,5), (6,3)\}$

b)



$\rightarrow G = \{(3 \le x \le 8,\ 2 \le y \le 7),\ (2 \le x \le 8,\ 2 \le y \le 5)\}$ or

$G = \{[(3,7), (8,2)], [(2,5), (8,2)]\}$

c) Pick any spot on border of S. Then it'll shrink if it's a "-" example. (i.e, (6, 3))

d) The smallest number you can provide is 2, because you only need 2 points for the candidate elimination algorithm to learn the target concept $(3 \le x \le 5,\ 2 \le y \le 9)$