HW 3

Q

Problem 1

i) $S = -\frac{4}{9} \log_2 \frac{4}{9} + \left(-\frac{5}{9} \log_2 \frac{5}{9}\right)$

$= 0.519 + 0.471$

$= 0.99$

For Attribute 1 (a_1):

$S_T = -\frac{3}{4} \log_2 3/4 + \left(-\frac{1}{4} \log_2 \frac{1}{4}\right)$

$= 0.311 + 0.5$

$= 0.811$

$S_F = -\frac{1}{5} \log_2 \frac{1}{5} + \left(-\frac{4}{5} \log_2 \frac{4}{5}\right)$

$= 0.464 + 0.256$

$= 0.72$

$\text{Gain}(S, a_1) = S - \Sigma \frac{|S_i|}{|S|} S_i$

$= 0.99 - \frac{4}{9} S_T - \frac{5}{9} S_F$

$= 0.99 - \frac{4}{9}(0.811) - \frac{5}{9}(0.72)$

$= 0.23$

For Attribute 2 (a_2)

i) Split point $a_2 > 1$

$S_{a_2 \leq 1} = 0$

$S_{a_2 > 1} = -\frac{3}{8} \log_2 3/8 + \left(-\frac{5}{8} \log_2 \frac{5}{8}\right)$

$= 0.53 + 0.423$

$= 0.953$

$\therefore S_{a_2} = 0(1/9) + 0.953(8/9) = 0.847$

ii) Split point $a_2 > 3$:-

$S_{a_2 \le 3} = -\frac{1}{2} \log_2 \left(\frac{1}{2}\right) + \left(-\frac{1}{2} \log_2 \frac{1}{2}\right)$

$= 1$

$S_{a_2 > 3} = -\frac{3}{7} \log_2 3/7 + \left(-4/7 \log_2 4/7\right)$

$= 0.524 + 0.461$

$= 0.985$

$S_{a_2} = \frac{2}{9}(1) + 7/9 (0.985)$

$= 0.222 + 0.766$

$= 0.9888$

iii) Split point $a_2 > 4$:-

$S_{a_2} \le 4 = -2/3 \log_2 (2/3) + (-1/3 \log_2 1/3)$

$= 0.39 + 0.528$

$= 0.918$

$S_{a_2 > 4} = -2/6 \log_2 2/6 + (-4/6 \log_2 4/6)$

$= 0.918$

$S_{a_2} = 3/9 (0.918)$
$+ 6/9 (0.918)$

$= 0.918$

iv) Split point $a_2 > 5$:-

$S_{a_2} \le 5 = -\frac{2}{5} \log_2 2/5 + (-3/5 \log_2 3/5)$

$= 0.969$

$S_{a_2 > 5} = -2/4 \log_2 2/4 + \left(-\frac{2}{4} \log_2 2/4\right)$

$= 1$

$S_{a_2} = 5/9 (0.969) +$
$4/9 (1)$

$= 0.538 + 0.444$

$= 0.982$

v) Split point $a_2 > 6$:-

$S_{a_2 \le 6} = \frac{-3}{6} \log_2 (3/6) + (-3/6 \log_2 3/6)$

$= 1$

$S_{a_2} > 6 = -\frac{1}{3} \log_2 \frac{1}{3} + \left(\frac{-2}{3} \log_2 2/3\right)$

$= 0.918$

$S_{a_2} = 6/9 (1) +$
$3/9 (0.918)$

$= 0.972$

vi) Split point $a_2 > 7$:-

$S_{a_2 \le 7} = -4/8 \log_2 4/8 + (-4/8 \log_2 4/8)$ | $S_{a_2} = 8/9 (1) + 0 (1/9)$

$= 1$ | $= 0.888$

$S_{a_2 > 7} = 0$

Entropy for attribute 2 is min for split $a_2 > 1$

$\therefore$ Gain $(S, a_2) = S - S_{a_2}$

$= 0.99 - 0.847$

$= 0.143$

$\therefore$ Gain is maximized for attribute I

$\therefore$ ~~Att~~ Attribute 1 (a_1) will be chosen as the first splitting for decisia tree.

2) Instance shouldn't be used as another attribute for the decision tree, as instances are unique for target values. So, It will give maximum entropy as compared to other attributes. Information gain will be least here.

Answer = No

Problem 2

1) GINI impurity of dataset $gini(D) = 1 - \Sigma p_i^2$

$$= 1 - P_+^2 - P_-^2$$

$$= 1 - \left(\frac{35}{100}\right)^2 - \left(\frac{65}{100}\right)^2$$

$$= 0.455$$

For attribute A :-

	A=T	A=F
+	20	15
-	30	35

$$gini(A) = \frac{|T|}{100} gini(T) + \frac{|F|}{100} gini(F)$$

$$= \frac{50}{100}(1 - P_+^2 - P_-^2) + \frac{50}{100}(1 - P_+^2 - P_-^2)$$

$$= 0.5\left(1 - \left(\frac{20}{50}\right)^2 - \left(\frac{30}{50}\right)^2\right) + 0.5\left(1 - \left(\frac{15}{50}\right)^2 - \left(\frac{35}{50}\right)^2\right)$$

$$= 0.5(1 - 0.16 - 0.36) + 0.5(1 - 0.09 - 0.49)$$

$$= 0.45$$

$$\Delta gini_A = gini(D) - gini(A) = 0.455 - 0.450 = 0.005$$

For attribute B :-

	B=T	B=F
+	15	20
-	20	45

$$gini(B) = \frac{|T|}{100} gini(T) + \left(\frac{F}{100}\right) gini(F)$$

$$= \frac{15+0+20+0}{100}(1 - P_+^2 - P_-^2) +$$

$$\frac{20+10+0+35}{100}(1 - P_+^2 - P_-^2)$$

$$= 0.35\left(1 - \left(\frac{15}{35}\right)^2 - \left(\frac{20}{35}\right)^2\right) + 0.65\left(1 - \left(\frac{20}{65}\right)^2 - \left(\frac{45}{65}\right)^2\right)$$

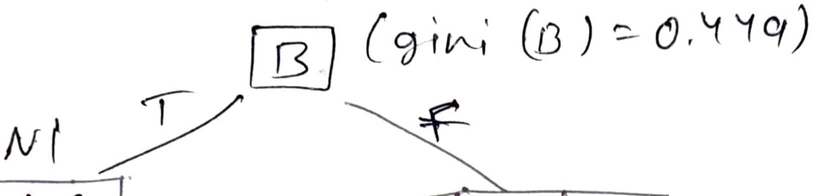
$$= 0.171 + 0.278$$

$$= 0.449$$

$\Delta gini_B = gini(A) - gini(B) = 0.45 - 0.449 = 0.006$

$\therefore$ $gini_B > gini_A$, So, Attribute B will be selected for first split

After splitting at B:-



(gini (B) = 0.449)

	A=T	A=f
+	0	15
−	20	0

	A=T	A=f
+	20	0
−	10	35

$$gini(N1) = \frac{|T|}{35}(1 - P_+^2 - P_-^2) + \frac{|F|}{35}(1 - P_+^2 - P_-^2)$$

$$= \frac{20}{35}(1-0-1) + \frac{15}{35}(1-1-0)$$

$$= 0$$

$$gini(N_2) = \frac{|T|}{65}(1 - P_+^2 - P_-^2) + \frac{|F|}{65}(1 - P_+^2 - P_-^2)$$

$$= \frac{30}{65}\left(1 - \left(\frac{20}{30}\right)^2 - \left(\frac{10}{30}\right)^2\right) + \frac{35}{65}(1-0-1)$$

$$= 0.205$$

$$\Rightarrow gini(\text{chidren}) = \frac{|N1|}{|N1|+|N2|} gini(N1) + \frac{|N2|}{|N2|+|N1|} gini(N2)$$

$$= \frac{35}{100}(0) + \frac{65}{100}(0.205)$$

gini (chidren) = 0.133

gini (parent) = gini (B) = 0.449

Problem 3

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1

For classifier H1, weights = 1/n = 1/10

H1	1	2	3	4	5	6	7	8	9	10
Y ⇒	1	1	1	-1	-1	-1	-1	-1	-1	-1
Weights	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10
Update Weight	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.2	0.2

Total Error = Σ Weights of misclassified features

$$= 1/10 + 1/10 = 2/10$$

∴ Performace $= \frac{1}{2} \log_e \left(\frac{1 - T.E}{T.E} \right)$

$$= 1/2 \log_e \left(\frac{1 - 2/10}{2/10} \right)$$

$$= \frac{1}{2} \log_e 4 \Rightarrow 0.693$$

∴ Upated Weights for misclassified

$\Rightarrow$ weight $\times e^{\text{performance}}$

$\Rightarrow 1/10 \times e^{0.693}$

$\Rightarrow 0.19997 \approx 0.2$

for correctly ~~call~~ classfied

$\Rightarrow$ Weight $\times e^{-\text{perfomace}}$

$\Rightarrow 1/10 \times e^{-0.693}$

$\Rightarrow 0.05$

For classifier H2, weights = 1/n = 1/10

H2	1	2	3	4	5	6	7	8	9	10
Y =>	-1	-1	-1	-1	-1	-1	-1	1	1	1
Weights =>	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10
Updated Weights =>	0.122	0.122	0.122	0.081	0.081	0.081	0.081	0.122	0.081	0.081

Total Error = 1/10 + 1/10 + 1/10 + 1/10 = 4/10

∴ Performace = $\frac{1}{2} \log_e \left(\frac{1 - 4/10}{4/10} \right)$

= $\frac{1}{2} \log_e 1.5 \Rightarrow 0.2$

∴ Updated Weight
miss classified

=> $\frac{1}{10} e^{0.2} \Rightarrow 0.122$

Corectly classified

=> $\frac{1}{10} e^{-0.2} \Rightarrow 0.081$

For classifier H3, weights = 1/m = 1/10

H3	1	2	3	4	5	6	7	8	9	10
Y	1	1	1	-1	-1	-1	-1	-1	-1	1
Updated weights	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.3	0.033

Total Error = 1/10

Performace = $\frac{1}{2} \log_e \left(\frac{1 - 1/10}{1/10} \right)$

= $\frac{1}{2} \log_e 9 \Rightarrow 1.1$

Updated weights
Miss classified

=> $\frac{1}{10} e^{1.1} \Rightarrow 0.3$

Corectly classified

=> $\frac{1}{10} e^{-1.1} \Rightarrow 0.033$

2) All of them

=) All the data instances will be reweighted after the first iteration. After the normalization of weights of all the data points, a new data set will be selected ~~so~~ from the old ~~da~~ data points according to their ~~to~~ new weight.