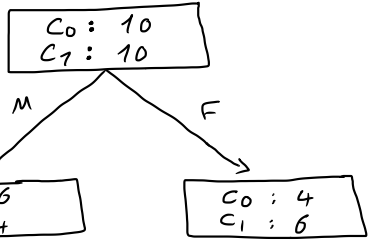
HW2 Yusuf Elnady

Q1: • Gini Index $= 1 - \sum_j [P(j|t)^2]$, Gini Index of Parent node $= 1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = 1/2$

(i) • Contingency table for parent node

C_0	10
C_1	10



• Split based on Gender

	M	F
C_0	6	4
C_1	4	6

• Gini index for male node $= 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$

• Gini index for female node $= 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48$

• Gini Children $= \sum_{i=1}^{k} \frac{n_i}{N} Gini(i) = \frac{10}{20}(0.48) + \frac{10}{20}(0.48) = 0.48$

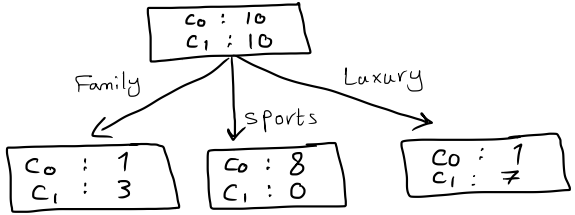
(ii) • Contingency table for parent node

C_0	10
C_1	10

, Gini Index of Parent node $= 1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = 1/2$

• split based on Car Type

	Family	Sports	Luxury
C_0	1	8	1
C_1	3	0	7



• Gini index for Family node $= 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 3/8 = 0.375$

• ~ ~ ~ Sports ~ $= 1 - (\frac{8}{8})^2 - 0 = 0$

• ~ ~ ~ Luxury ~ $= 1 - (\frac{1}{8})^2 - (\frac{7}{8})^2 = 7/32 = 0.21875$

• Gini Children $= \sum_{i-1}^{k} \frac{n_i}{N} Gini(i) = \frac{4}{20}(0.375) + 0 + \frac{8}{20}(\frac{7}{32}) = \frac{13}{80} = 0.1625$

So we should split using the CarType because it has lower impurity #

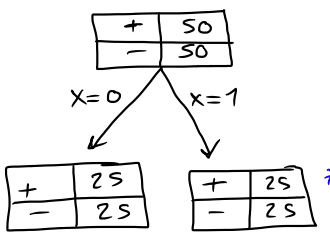
Q2 * Gini index of the parent node without splitting $= 1 - (\frac{50}{100})^2 - (\frac{50}{100})^2 = \frac{1}{2}$

+	50
0	50

① Level ①

Splitting based on X

X	0	1
+	25	25
−	25	25



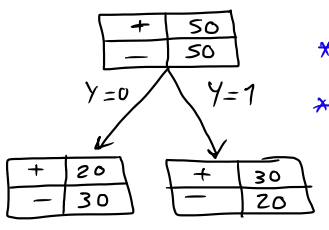
* Gini (0) $= 1 - (\frac{25}{50})^2 - (\frac{25}{50})^2 = \frac{1}{2}$

* Gini(1) $= 1 - (\frac{25}{50})^2 - (\frac{25}{50})^2 = \frac{1}{2}$

* Gini children $= (\frac{50}{100})(\frac{1}{2}) + (\frac{50}{100})(\frac{1}{2}) = \frac{1}{2} = 0.500$

Splitting based on Y

Y	0	1
+	20	30
−	30	20



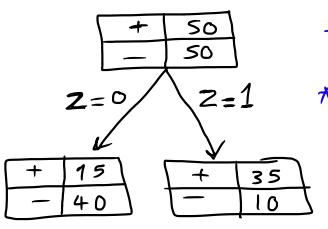
* Gini (0) $= 1 - (\frac{20}{50})^2 - (\frac{30}{50})^2 = \frac{12}{25}$

* Gini(1) $= 1 - (\frac{30}{50})^2 - (\frac{20}{50})^2 = \frac{12}{25}$

* Gini children $= (\frac{50}{100})(\frac{12}{25}) + (\frac{50}{100})(\frac{12}{25}) = \frac{12}{25} = 0.48$

Splitting based on Z

Z	0	1
+	15	35
−	40	10



* Gini (0) $= 1 - (\frac{15}{55})^2 - (\frac{40}{55})^2 = \frac{48}{121}$

* Gini(1) $= 1 - (\frac{35}{45})^2 - (\frac{10}{45})^2 = \frac{28}{81}$

* Gini Children $= (\frac{55}{100})(\frac{48}{121}) + (\frac{45}{100})(\frac{28}{81}) = \frac{37}{99} = 0.373$

- For Level splitting, the best attribute to use is Z, as it has the lowest Impurity (Gini Index)

Level ②

- when Z = 0

X	0	1
+	15	0
−	15	25

Y	0	1
+	0	15
−	20	20

* Gini index for children of Z = 0 when splitting using X
$= \frac{30}{55}(1 - (\frac{15}{30})^2 - (\frac{15}{30})^2) + \frac{25}{55}(1 - 0 - (\frac{25}{25})^2) = \frac{3}{11} = 0.2727$

* Gini index for children of Z = 0 when splitting using Y
$= \frac{20}{55}(1 - (\frac{20}{20})^2) + \frac{35}{55}(1 - (\frac{15}{35})^2 - (\frac{20}{35})^2) = \frac{24}{77} = 0.3116$

* We use attr X to split from Z = 0 since it has the lowest impurity (Gini Index)

- when Z = 1

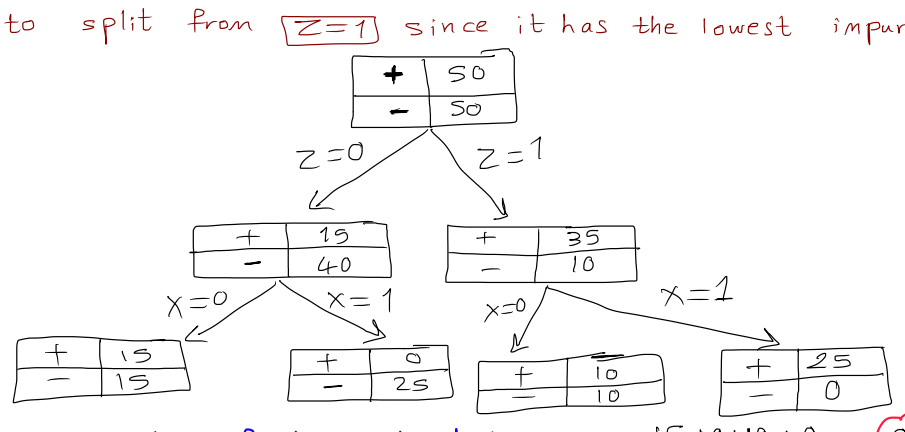
X	0	1
+	10	25
−	10	0

Y	0	1
+	20	15
−	10	0

* Gini index for children of Z = 1 when splitting using X
$= \frac{20}{45}(1 - (\frac{10}{20})^2 - (\frac{10}{20})^2) + \frac{25}{45}(1 - 0 - (\frac{25}{25})^2) = \frac{2}{9} = 0.2222$

* Gini index for children of Z = 1 when splitting using Y
$= \frac{30}{45}(1 - (\frac{20}{30})^2 - (\frac{10}{30})^2) + \frac{15}{45}(1 - (\frac{15}{15})^2) = \frac{8}{27} = 0.2962$

* we use attr X to split from Z = 1 since it has the lowest impurity (Gini Index)

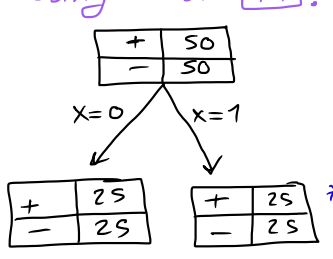


* The overall Training error rate of the induced tree is $\frac{15+0+10+0}{100} = 0.25$ # ①

② Level ① is already splitted using attr X.

Splitting based on X

X	0	1
+	25	25
−	25	25



* Gini (0) $= 1 - (\frac{25}{50})^2 - (\frac{25}{50})^2 = \frac{1}{2}$

* Gini(1) $= 1 - (\frac{25}{50})^2 - (\frac{25}{50})^2 = \frac{1}{2}$

* Gini children $= (\frac{50}{100})(\frac{1}{2}) + (\frac{50}{100})(\frac{1}{2}) = \frac{1}{2} = 0.5$

Level ②

when X = 0

Y	0	1
+	0	25
−	25	0

Z	0	1
+	15	10
−	15	10

* Gini index for children of X = 0 when splitting on Y
$= (\frac{25}{50})(1 - (\frac{25}{25})^2) + (\frac{25}{50})(1 - (\frac{25}{25})^2) = 0$

* Gini index for children of X = 0 when splitting on Z
$= \frac{30}{50}(1 - (\frac{15}{30})^2 - (\frac{15}{30})^2) + \frac{20}{50}(1 - (\frac{10}{20})^2 - (\frac{10}{20})^2)$
$= \frac{1}{2}$

* we use attr Y to split from X=0 since it has the lowest impurity (Gini Index)

when X = 1

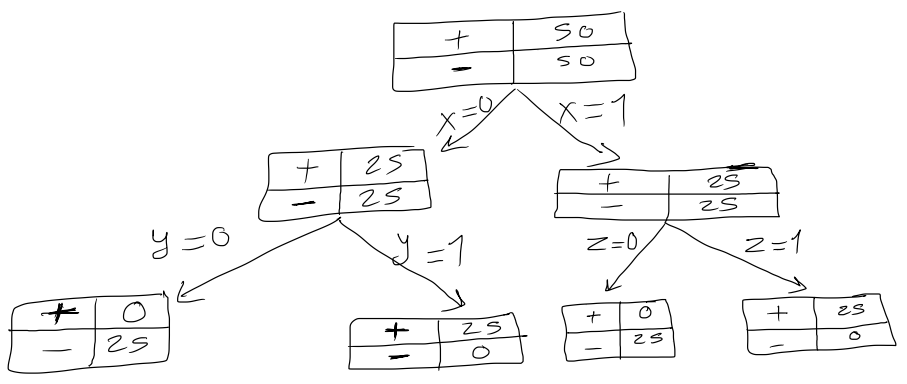
Y	0	1
+	20	5
−	5	20

Z	0	1
+	0	25
−	25	0

* Gini index for children of X = 1 when splitting on Y
$= (\frac{25}{50})(1 - (\frac{20}{25})^2 - (\frac{5}{25})^2) + (\frac{25}{50})(1 - (\frac{20}{25})^2 - (\frac{5}{25})^2) = \frac{8}{25} = 0.32$

* Gini index for children of X = 1 when splitting on Z
$= (\frac{25}{50})(1 - (\frac{25}{25})^2) + (\frac{25}{50})(1 - (\frac{25}{25})^2) = 0$

* we use attr Z to split from X = 1 since it has the lowest impurity (Gini Index)



* The overall Training error rate of the induced tree is $\frac{0+0+0+0}{100} = 0$ # ②

③ From ①, ②, we see that using the greedy heuristic approach doesn't always result in the best optimal solution ? because the training error of the induced tree in ② was zero.

Q3

① • Entropy measure $= -P(x)\log_2 P(x)$ ∵ as $P(x)$ is the relative frequency, so $0 < P(x) \leq 1$

• For any base $(b) > 1$, the logarithmis of numbers between 0 and 1 are $\log_b(P(x)) \leq 0$

• So, $-P(x)\log_2 P(x) \geq 0$ (a non-negative value)

② [a] $H(Y) = -\sum_j P(j|y)\log_2 P(j|t) = -\frac{4}{7}\log_2(\frac{4}{7}) - \frac{3}{7}\log_2(\frac{3}{7}) =$ **0.9852 bits**

[b]

	A=0	A=1
C_0	3	1
C_1	0	3

($y=0 \sim C_0$, $y=1 \sim C_1$)

* Entropy $(A=0) = -\frac{3}{3}\log\frac{3}{3} = 0$
* Entropy $(A=1) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.8112$
* Information Gain $I(Y;A) = 0.9852 - (\frac{3}{7})(0) - (\frac{4}{7})(0.8112) =$ **0.5217 bits**

[c]

	B=0	B=1
C_0	2	2
C_1	1	2

($y=0 \sim C_0$, $y=1 \sim C_1$)

* Entropy $(B=0) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9183$
* Entropy $(B=1) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$
* $I(Y;B) = 0.9852 - (\frac{3}{7})(0.9183) - (\frac{4}{7})(1) =$ **0.0202 bits**

[d]

	C=0	C=1	C=2
C_0	2	1	1
C_1	1	0	2

($y=0 \sim C_0$, $y=1 \sim C_1$)

* Entropy $(C=0) = -\frac{2}{3}\log(\frac{2}{3}) - \frac{1}{3}\log(\frac{1}{3}) = 0.9183$
* Entropy $(C=1) = -1\log 1 = 0$
* Entropy $(C=2) = -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3}) = 0.9183$
* $I(Y;C) = 0.9852 - (\frac{3}{7})\,0.9183 - 0 - (\frac{3}{7})(0.9183) =$ **0.1981 bits**

[e] Based on the count matrix above, it's clear it will be perfectly classified after having **depth = 3**, which means we have splitted using all 3 attributes

-Also from the remaining parts of the questions (f, g, h) ⇒ the final tree will have **depth = 3**

[f] -the mutual information is another name for Information Gain

- The highest information gain was in splitting using **A**

[g] When $A=0$ ⇒ No need to further split, because all rows belong to C_0 $(y=0)$

when $A=1$ ⇒

* Entropy $(A=1) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.8112$ //Parent node

If to split using **B**

	B=0	B=1
C_0	0	1
C_1	1	2

($y=0 \sim C_0$, $y=1 \sim C_1$)

* Entropy $(B=0) = -1\log(1) = 0$
* Entropy $(B=1) = -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3}) = 0.9183$
* $I(B;A) = 0.8112 - 0 - \frac{3}{4}(0.9183) = 0.1225$

If to split using **C**

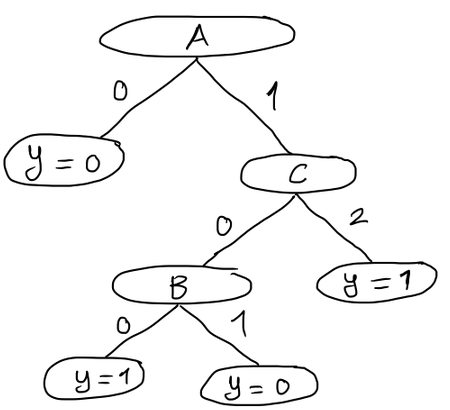
	C=0	C=2
C_0	1	0
C_1	1	2

($y=0 \sim C_0$, $y=1 \sim C_1$)

* Entropy $(C=0) = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(\frac{1}{2}) = 1$
* Entropy $(C=1) = -\frac{2}{2}\log(\frac{2}{2}) = 0$
* $I(C;A) = 0.8112 - \frac{2}{4}(1) = 0.3112$

- The highest information gain was in splitting using **C**

[h]



Q 4

a $Entropy = -P(x)\log_2 P(x)$? $Entropy\ of\ parent = -\frac{10}{20}\log(\frac{10}{20}) - \frac{10}{20}\log(\frac{10}{20}) = 1$

using ID as splitting attr ⇒ $Information\ Gain = 1 - (20)\left[(\frac{1}{20})(-1)\log(1)\right] = 1$

b $Entropy\ of\ parent = -\frac{10}{20}\log(\frac{10}{20}) - \frac{10}{20}\log(\frac{10}{20}) = 1$

using Handedness as splitting attr ⇒

	L	R
+	9	1
-	1	9

* $Entropy\ (Handedness = Left) = -\frac{9}{10}\log(\frac{9}{10}) - \frac{1}{10}\log(\frac{1}{10}) = 0.4690$
* $Entropy\ (Handedness = Right) = -\frac{1}{10}\log(\frac{1}{10}) - \frac{9}{10}\log(\frac{1}{10}) = 0.4690$
* $Information\ Gain = 1 - \frac{10}{20}(0.4690) - \frac{10}{20}(0.4690) = 0.531$

c *From @,ⓑ using the Information Gain ⇒ It suggests splitting using ID (has the highest gain)

* Many small nodes (not efficient solution)

d $Gain\ Ratio = \frac{Info\ Gain}{Split\ Info}$? $Split\ Info = -\sum_{i=1}^{K}\frac{n_i}{N}\log\frac{n_i}{N}$

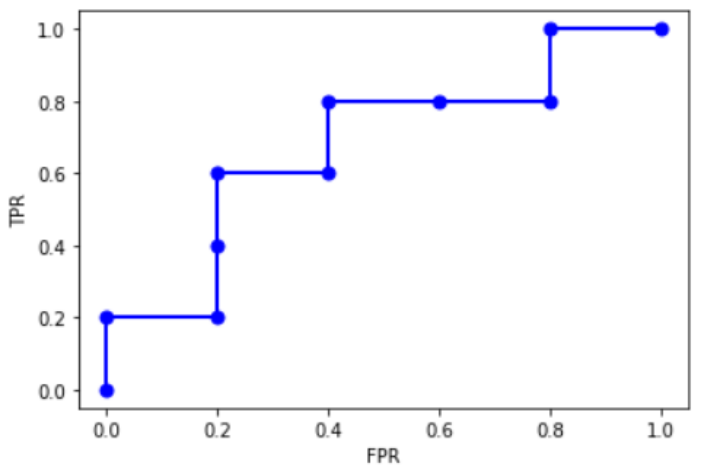
using ID to split ⇒ $Gain\ Ratio = \frac{1}{(20)(\frac{-1}{20})\log\frac{1}{20}} = 0.2314$

e using Handedness to split ⇒ $Gain\ Ratio = \frac{0.531}{-\frac{10}{20}\log\frac{10}{20} - \frac{10}{20}\log\frac{10}{20}} = \frac{0.531}{1} = 0.531$

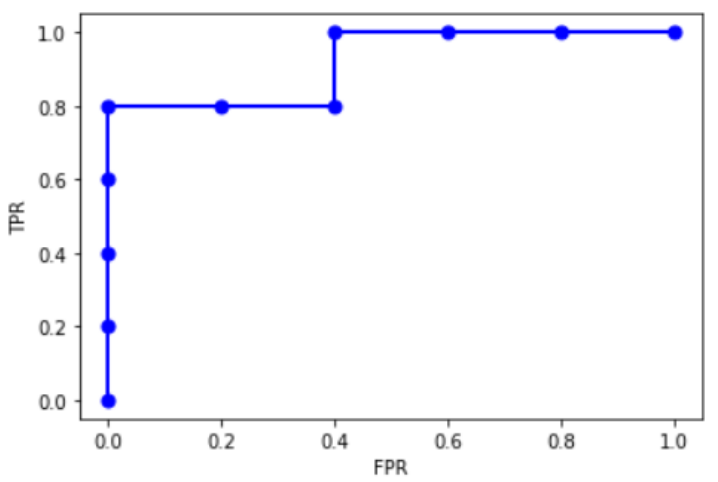
f From ⓓ, ⓔ, using the Gain Ratio ⇒ The highest Gain Ratio is using Handedness, so we should split according to this attr.

Q5 1 (a)

* the calculation of (TPR) and (FPR) for both of the models is found on the Jupyter Notebook
* ROC curve for the First Model:



* ROC curve for the Second Model:



(b) * The Area under ROC curve for the first classifier is $(0.2)(0.2)+(0.2)(0.6)+(0.4)(0.8)+(0.2)(1) = \boxed{0.68}$ #

* The Area under ROC curve for the Second classifier is $(0.4)(0.8)+(0.6)(1) = \boxed{0.92}$ #

* The second classifier C_2 has a larger area under the curve.

(c) * For the First classifier: $(m=n=5)$ $\Rightarrow WMW = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} I(x_i, y_i)}{mn}$

- we will only take the following x values from $c_1 \Rightarrow \{0.15, 0.31, 0.62, 0.77, 0.95\}$
  y values from $c_1 \Rightarrow \{0.1, 0.2, 0.3, 0.4, 0.8\}$

$$WMW = \frac{5+4+4+3+1}{(5)(5)} = \frac{17}{25} = \boxed{0.68}$$

* For the second classifier: $(m=n=5)$

- we will take the following x values from $C_2 \Rightarrow \{0.49, 0.65, 0.66, 0.70, 0.99\}$
  y values from $C_2 \Rightarrow \{0.05, 0.25, 0.35, 0.55, 0.6\}$

$$WMW = \frac{5+5+5+5+3}{(5)(5)} = \frac{23}{25} = \boxed{0.92}$$

* From WMW for both classifiers, we get that $\boxed{C_2}$ has a larger WMW value.

* It's clear that calculating (WMW) is the same as (the area under RUC curve).

Q5 b

Actual class \ Predicted class	Yes	No
Yes	TP 345	FN 225
No	FP 195	TN 235

① Precision $= \frac{TP}{TP+FP} = \frac{345}{345+195} = \frac{23}{36} = 0.6388$

② Recall $= \frac{TP}{TP+FN} = \frac{345}{345+225} = \frac{23}{38} = 0.6053$

③ FPR $= \frac{FP}{FP+TN} = \frac{195}{195+235} = \frac{39}{86} = 0.4535$

④ F-measure $= \frac{2rP}{r+P} = \frac{2TP}{2TP+FP+FN} = 0.6216 = \frac{23}{37}$

⑤ Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{345+235}{345+235+195+225} = \frac{29}{50} = 0.58$