CS 5525

Biplav Choudhury (906089826)

7 October 2020

1 Solutions to Assignment 2

1.1 Problem 1

Based on Gender The distribution for gender and the classes C0, C1 looks like -

Table 1: Count Matrix

	Male	**Female**
C0	6	4
C1	4	6

At parent, number of C0=10 and number of C1=10. Total $n = 20$.

Gini for parent (G) =$1 - [(\frac{10}{20})^2 + (\frac{10}{20})^2] = 0.5$

Gini at first child based on all male $(G_1) = 1 - [(\frac{6}{10})^2 + (\frac{4}{10})^2] = 0.48$. Number of elements here $n_1 = 10$

Gini at second child based on all female $(G_2) = 1 - [(\frac{4}{10})^2 + (\frac{6}{10})^2] = 0.48$. Number of elements here $n_2 = 10$

Overall Gini at child (G_{split})= $\frac{10}{20} \times G_1 + \frac{10}{20} \times G_2 = 0.48$

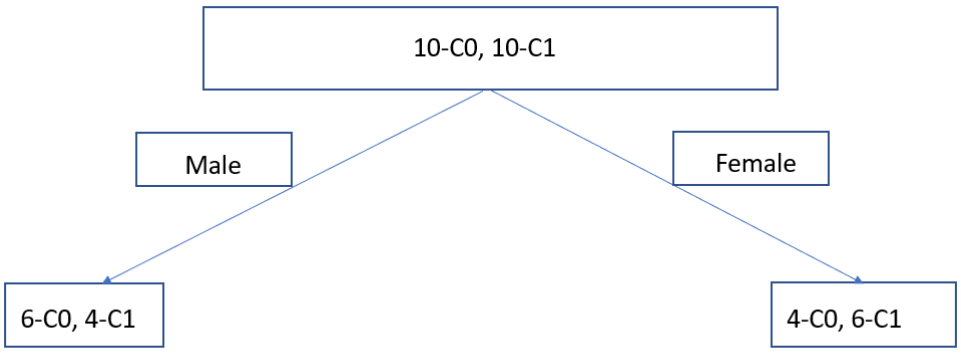

Figure 1: Split based on Gender

Table 2: Count Matrix

	Family	**Sports**	**Luxury**
C0	1	8	1
C1	3	0	7

Based on Car type The distribution for car-type and the classes C0, C1 looks like - At parent, number of C0=10 and number of C1=10. Total $n = 20$.

Gini for parent (G) =$1 - [(\frac{10}{20})^2 + (\frac{10}{20})^2] = 0.5$

Gini at first child based on family car $(G_1) = 1 - [(\frac{1}{4})^2 + (\frac{3}{4})^2] = 0.375$. Number of elements here $n_1 = 4$

Gini at second child based on all sports car $(G_2) = 1 - [(\frac{8}{8})^2 + (\frac{0}{10})^2] = 0$. Number of elements here $n_2 = 8$

Gini at third child based on all luxury car $(G_3) = 1 - [(\frac{1}{8})^2 + (\frac{7}{8})^2] = 0.2187$. Number of elements here $n_3 = 8$

Overall Gini at child (G_{split})= $\frac{4}{20} \times G_1 + \frac{8}{20} \times G_2 + \frac{8}{20} \times G_3 = 0.1624$

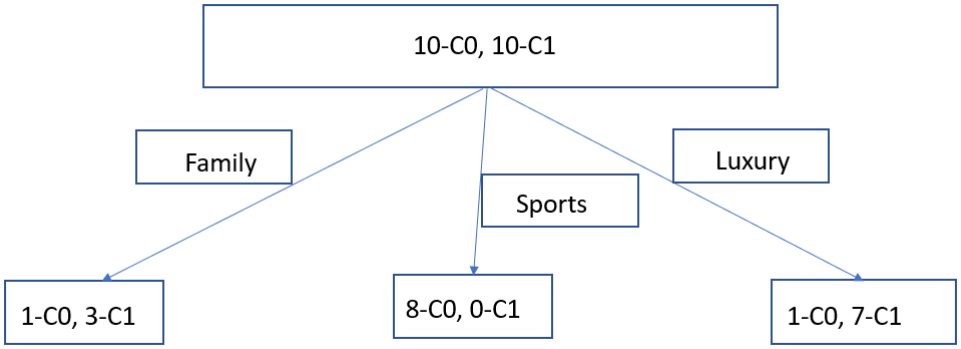

Figure 2: Split based on Gender

1.2 Problem 2 Part 1

At the parent, 50 positive and negative samples. Therefore Gini at parent G_p=0.5

For the first split, 3 options - $X, Y or Z$.

If the first split is based on X:

Table 3: Count Matrix

	X=0	**X=1**
+	25	25
-	25	25

Therefore Gini $G_x = \frac{50}{100}[1-(\frac{25}{50})^2-(\frac{25}{50})^2]+\frac{50}{100}[1-(\frac{25}{50})^2-(\frac{25}{50})^2] = 0.5$. Gain $= 0.5-0.5 = 0$

If the first split is based on Y:

Table 4: Count Matrix

	Y=0	**Y=1**
+	20	30
-	30	20

Therefore Gini $G_y = \frac{50}{100}[1-(\frac{20}{50})^2-(\frac{30}{50})^2]+\frac{50}{100}[1-(\frac{30}{50})^2-(\frac{20}{50})^2] = 0.48$. Gain $= 0.5-0.48 = 0.02$

If the first split is based on z:

Table 5: Count Matrix

	Z=0	**Z=1**
+	15	35
-	40	10

Therefore Gini $G_z = \frac{55}{100}[1-(\frac{15}{55})^2-(\frac{40}{55})^2]+\frac{45}{100}[1-(\frac{35}{45})^2-(\frac{10}{45})^2] = 0.3736$. Gain $= 0.5-0.3736 = 0.1264$

As first split with Z gives the highest gain, the first splitting criteria is based on Z and the tree looks like Fig. 3 -

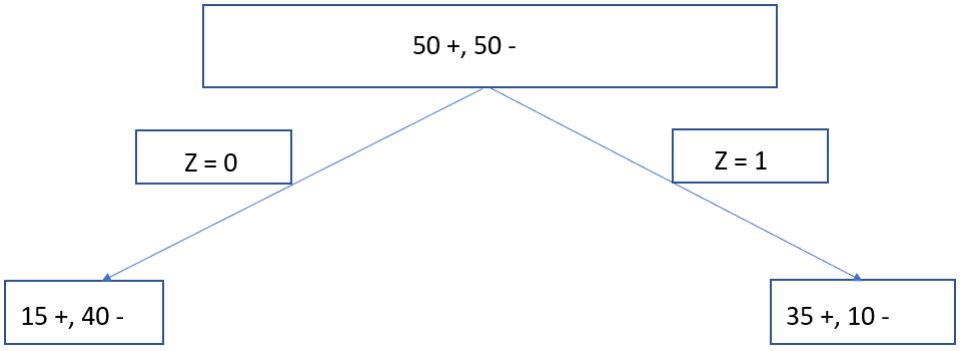

Figure 3: First split based on Z

Now we have two parent nodes at the second level due to the split based on Z. One has 15 positive and 40 negative samples, and the other has 35 positive and 10 negative examples. Now for the second split, we have either X or Y.

If the second split is based on X:

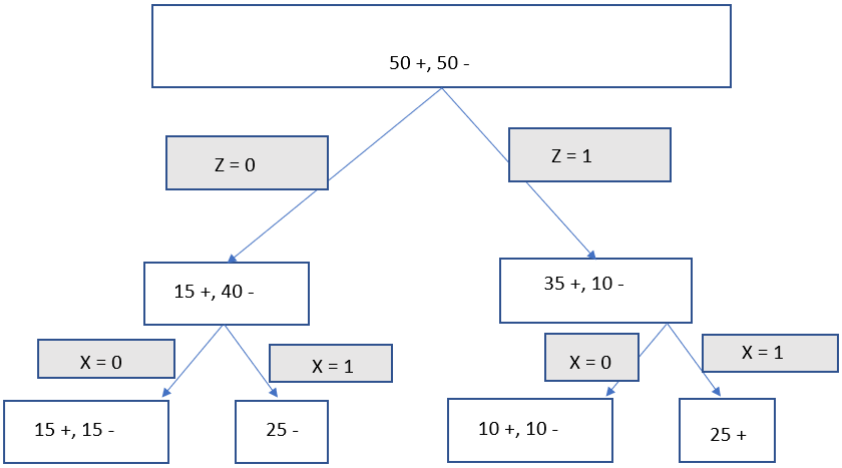

Figure 4: Second split based on X

For the parent node with 15 positive and 40 negative samples, Gini $G_{p1} = 1 - (\frac{15}{55})^2 - (\frac{40}{55})^2 = 0.3966$

Gini at the two childs $G_{c1} = \frac{30}{55}[1 - (\frac{15}{30})^2 - (\frac{15}{30})^2] + \frac{25}{55}[1 - (\frac{25}{25})^2] = 0.2727$. Gain $= G_{p1} - G_{c1} = 0.3966 - 0.2727 = 0.1239$

For the parent node with 35 positive and 10 negative samples, Gini $G_{p2} = 1 - (\frac{35}{45})^2 - (\frac{10}{45})^2 = 0.3456$

Gini at the two childs $G_{c2} = \frac{20}{45}[1 - (\frac{10}{20})^2 - (\frac{10}{20})^2] + \frac{25}{55}[1 - (\frac{25}{25})^2] = 0.2222$. Gain $= G_{p2} - G_{c2} = 0.3456 - 0.2222 = 0.1234$

If the second split is based on Y:

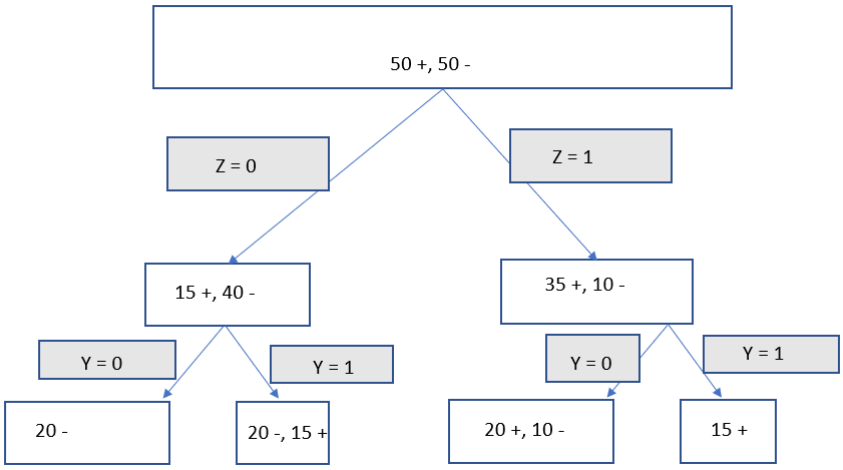

Figure 5: Second split based on Y

For the parent node with 15 positive and 40 negative samples, Gini $G_{p1} = 1 - (\frac{15}{55})^2 - (\frac{40}{55})^2 = 0.3966$

Gini at the two childs $G_{c1} = \frac{20}{55}[1 - (\frac{20}{20})^2] + \frac{35}{55}[1 - (\frac{20}{35})^2 - (\frac{15}{35})^2] = 0.3116$. Gain $= G_{p1} - G_{c1} = 0.3966 - 0.3116 = 0.085$

For the parent node with 35 positive and 10 negative samples, Gini $G_{p2} = 1 - (\frac{35}{45})^2 - (\frac{10}{45})^2 = 0.3456$

Gini at the two childs $G_{c2} = \frac{30}{45}[1 - (\frac{20}{30})^2 - (\frac{10}{30})^2] + \frac{15}{45}[1 - (\frac{15}{15})^2] = 0.2962$. Gain $= G_{p2} - G_{c2} = 0.3456 - 0.2962 = 0.0494$

Therefore suing X as the second split has more gain in both the splits at the second level, and therefore X will be used as the second split. The final tree looks like Fig. 4.

The error is $\frac{25}{100}$.

1.3 Problem 2 Part 2

Given, first split based on X

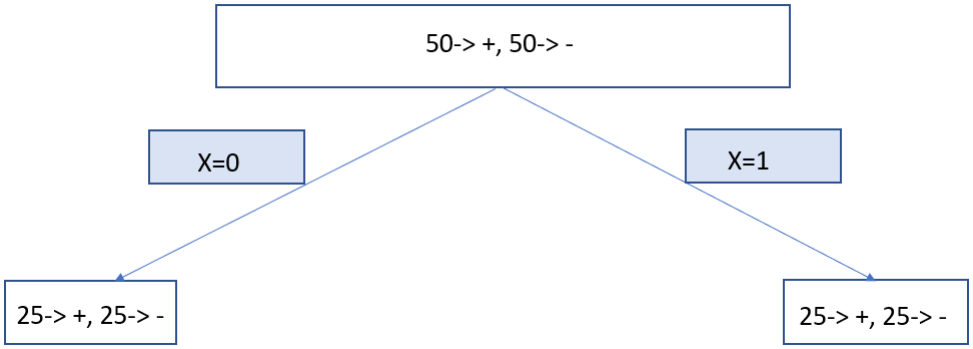

Figure 6: First split based on X

Node N1 will be split on Y and N2 will be split based on Z as shown in Fig. **??**.

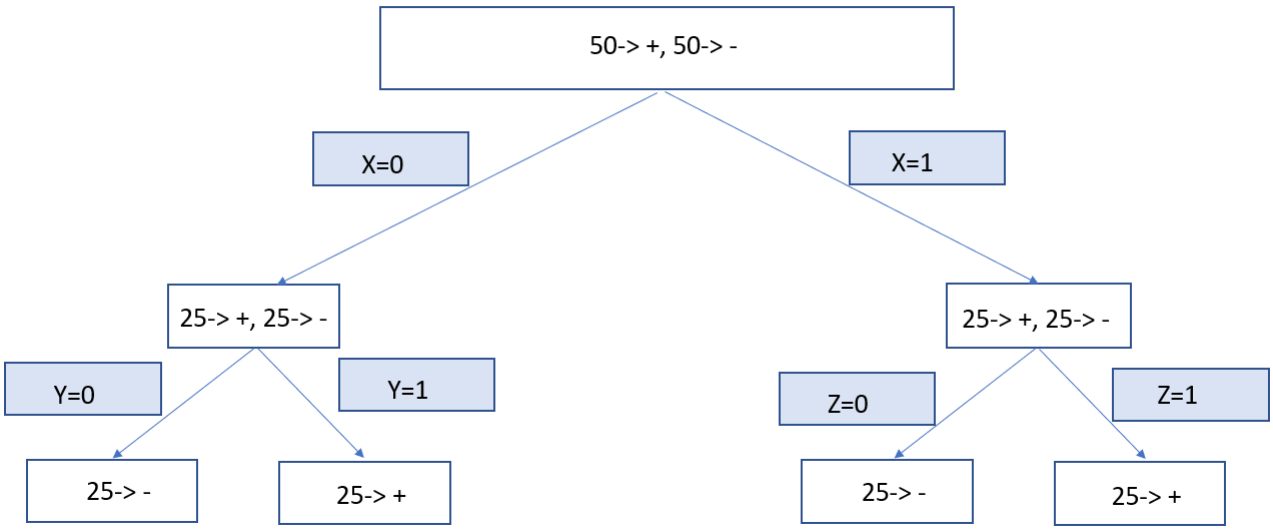

Figure 7: Second level split

1.4 Problem 2 Part 3

Compared to $\frac{25}{100}$ previously, error has reduced to 0. Therefore it shows that greedily classifying based on Gini may not always be the best approach.

1.5 Problem 3 Part 1

Entropy is given by

$$E = -p(x)\log p(x) = p(x)[-\log p(x)]$$

Using $\log a^b = b\log a$, $-\log p(x)$ can be written as $\log \frac{1}{p(x)}$

$$\text{Therefore} \qquad E = p(x)\log \frac{1}{p(x)}$$

Now $p(x) \geq 0$, and $\log \frac{1}{p(x)} \geq 0$ as $0 \leq p(x) \leq 1 \implies \frac{1}{p(x)} \geq 1$. Therefore entropy is the product of two non-negative terms and is itself a non-negative term too.

1.6 Problem 3 Part 2 a

Entropy of Y $= E = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} = 0.9852$ bits

1.7 Problem 3 Part 2 b

Based on split w.r.t A,

Table 6: Count Matrix

	A=0	**A=1**
0	3	1
1	0	3

Entropy at child after split on A $= \frac{3}{7} \times [-\frac{3}{3}\log\frac{3}{3}] + \frac{4}{7} \times [-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}] = 0.4635$ bits
Gain $= 0.9852 - 0.4635 = 0.5217$ bits

1.8 Problem 3 Part 2 c

Table 7: Count Matrix

	B=0	**B=1**
0	2	2
1	1	2

Entropy at child after split on B $= \frac{3}{7} \times [-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}] + \frac{4}{7} \times [-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}] = 0.9649$ bits

Gain $= 0.9852 - 0.9649 = 0.0203$ bits

1.9 Problem 3 Part 2 d

Table 8: Count Matrix

	C=0	**C=1**	**C=2**
0	2	1	1
1	1	0	2

Entropy at child after split on $C = \frac{3}{7} \times [-\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3}] + \frac{1}{7} \times [-\frac{1}{1} \log \frac{1}{1}] - \frac{3}{7} \times [-\frac{2}{3} \log \frac{2}{3} = 0.7871$ bits

Gain $= 0.9852 - 0.7871 = 0.1981$ bits

1.10 Problem 3 Part 2 e

Based on information gain, the full tree will look like Fig. 8.

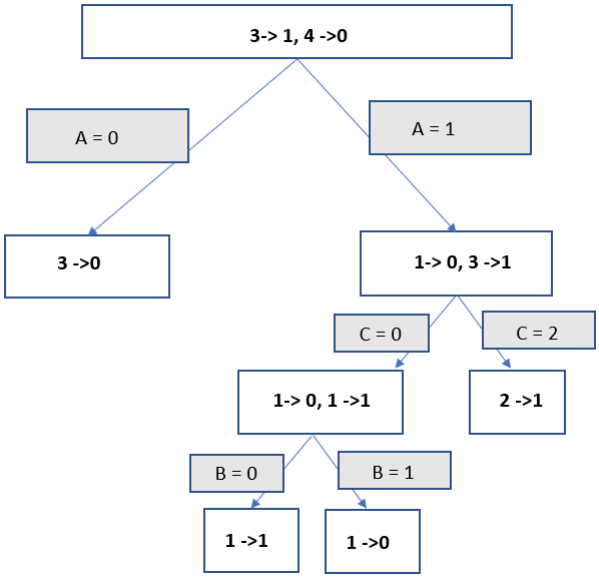

Figure 8: Tree for full classification

The depth is 3.

1.11 Problem 3 Part 2 f

First split is based on A.

1.12 Problem 3 Part 2 g

Second split is based on C.

1.13 Problem 3 Part 2 h

Whole tree with labels is shown in Fig. 8.

1.14 Problem 4 Part 1

At parent, 10 samples for each positive and negative. $p_1 = p_2 = 0.5$. Entropy at parent $E_p = -p_1 \log p_1 - p_2 \log p_2 = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$. This is because both classes is in equal proportion and entropy/randomness is maximum.

When using ID as the splitting attribute, it will lead to 20 pure child nodes as ID is unique to

each player. Entropy at one such child node $= -1 \times \log 1 = 0$. So average entropy across 20 child nodes $= E_{c,ID} = 0$. This is because only one class is in 1 node and therefore there is no entropy/randomness.

Information Gain $(IG_1) = E_p - E_{c,ID} = 1 - 0 = 1$.

As we get pure nodes, the information gain is very high.

1.15 Problem 4 Part 2

Using handedness as the splitting attribute,

Table 9: Count Matrix

	Left	**Right**
+	9	1
-	1	9

We have entropy at parent $E_p = 1$.

From Table. 9, we have at first child of left handed players, $p_1 = \frac{9}{10}, p_2 = \frac{1}{10}$. Therefore entropy $E_{c1} = -p_1 \log p_1 - p_2 \log p_2 = 0.4689$

Similarly for second child of right handed players, $p_1 = \frac{1}{10}, p_2 = \frac{9}{10}$. Therefore entropy $E_{c1} = -p_1 \log p_1 - p_2 \log p_2 = 0.4689$

Average entropy at child $E_{c,hand} = \frac{10}{20} E_{c1} + \frac{10}{20} E_{c2} = 0.4689$

Information Gain $IG_2 = E_p - E_{c,hand} = 1 - 0.4689 = 0.5311$

1.16 Problem 4 Part 3

As $IG_1 > IG_2$, splitting based on ID is better.

1.17 Problem 4 Part 4

$GainRatio = \frac{Gain}{SplitInfo}$

For ID as splitting attribute, $Gain = G_1 = 1$

$SplitInfo = -\frac{1}{20} \times \log \frac{1}{20} \times 20 = 4.3219$

Thus $GainRatio = GR_1 = \frac{1}{4.3219} = 0.2313$

1.18 Problem 4 Part 5

For handedness as splitting attribute, $Gain = G_2 = 0.5311$

$SplitInfo = -\frac{10}{20} \times \log \frac{10}{20} \times 2 = 1$

Thus $GainRatio = GR_2 = \frac{0.5311}{1} = 0.5311$

1.19 Problem 4 Part 6

As $GR_1 < GR_2$, handedness is a better splitting attribute now.

1.20 Problem 5 Part 1a

Code is included in the python notebook.

The ROC curve is shown below :

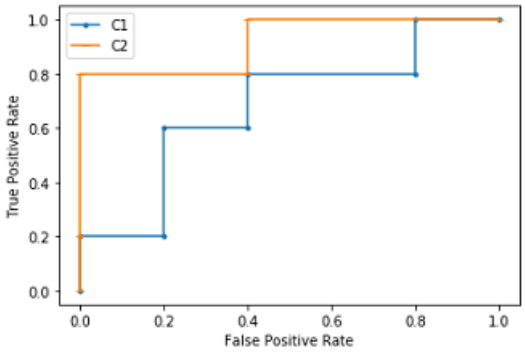

Figure 9: ROC curve

Manual calculation for C1 is shown below in Fig. 10 and Fig. 11-

Thresh	0.1	0.15	0.2	0.3	0.31	0.4	0.62	0.77	0.81	0.95	1
TP	5	5	4	4	4	3	3	2	1	1	0
FP	5	4	4	3	2	2	1	1	1	0	0
TN	0	1	1	2	3	3	4	4	4	5	5
FN	0	0	1	1	1	2	2	3	4	4	5
TPR	1	1	0.8	0.8	0.8	0.6	0.6	0.4	0.2	0.2	0
FPR	1	0.8	0.8	0.6	0.4	0.4	0.2	0.2	0.2	0	0

Figure 10: ROC curve manual calculation C1

Thresh	0.05	0.25	0.35	0.49	0.55	0.6	0.65	0.66	0.7	0.99	1
TP	5	5	5	5	4	4	4	3	2	1	0
FP	5	4	3	2	1	1	0	0	0	0	0
TN	0	1	2	3	4	4	5	5	5	5	5
FN	0	0	0	0	1	1	1	2	3	4	5
TPR	1	1	1	1	0.8	0.8	0.8	0.6	0.4	0.2	0
FPR	1	0.8	0.6	0.4	0.2	0.2	0	0	0	0	0

Figure 11: ROC curve manual calculation C2

1.21 Problem 5 Part 1b

The areas are given as -
C1: ROC AUC=0.680
C2: ROC AUC=0.920

Thus the classifier C2 is better.

1.22 Problem 5 Part 1c

WMW for C1 is 0.68 and for C2 is 0.92. And it is the same value as the AUC.

1.23 Problem 5 Part 2

	Predicted Class = Yes	Predicted Class = No
Actual Class = Yes	345	225
Actual Class = No	195	235

Figure 12: Confusion Matrix

1.23.1 Precision

$= \frac{TP}{TP+FP} = \frac{345}{345+195} = 0.6388$

1.23.2 Recall / True Positive Rate

$= \frac{TP}{TP+FN} = \frac{345}{345+225} = 0.6052$

1.23.3 False Positive Rate

$= \frac{FP}{TN+FP} = \frac{195}{235+195} = 0.4534$

1.23.4 F-score

$= \frac{2 \times recall \times precision}{recall+precision} = 0.6215$

1.23.5 Accuracy

$= \frac{TP+TN}{TP+FP+FN+TN} = \frac{345+235}{1000} = 0.58$