ANS 1

(a)

Entropy of the dataset:

$$S = -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right) \approx 0.991$$

(b)

For a1:

$$Entropy(T) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$
$$Entropy(F) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.7219$$
$$Gain(a_1) = 0.9907 - \left(\frac{4}{9}(0.8113) + \frac{5}{9}(0.7219)\right) = 0.229$$

For a2:

$$Entropy(T) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$
$$Entropy(F) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183$$
$$Gain(a_2) = 0.9907 - \left(\frac{6}{9}(1) + \frac{3}{9}(0.9183)\right) = 0.018$$

(c)

Information Gain for each possible split on a3:
Sorted a3: 1.0, 3.0, 4.0, 5.0, 5.0, 6.0, 7.0, 7.0, 8.0
Possible midpoints (where class label changes):
Candidates: {2.0,3.5,4.5,5.5,6.5,7.0,7.5}

Compute Gain for each:

- Split at 2.0: Gain = 0.143
- Split at 3.5: Gain = 0.003
- Split at 4.5: Gain = 0.073
- Split at 5.5: Gain = 0.007
- Split at 6.5: Gain = 0.018
- Split at 7.0: Gain = 0.102
- Split at 7.5: Gain = 0.102

Best split: a3 ≤ 2.0

(d)

Best split according to Information Gain:

$$Gain(a_1) = 0.229, \quad Gain(a_2) = 0.018, \quad \max Gain(a_3) = 0.143$$

Therefore, Best split = a1.

(e)

Gain Ratio:

$$SplitInfo(a_1) = -\left(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right) = 0.991$$

$$SplitInfo(a_2) = -\left(\frac{6}{9}\log_2\frac{6}{9} + \frac{3}{9}\log_2\frac{3}{9}\right) = 0.918$$

$$GainRatio(a_1) = \frac{0.229}{0.991} = 0.231, \quad GainRatio(a_2) = \frac{0.018}{0.918} = 0.020$$

Best split by Gain Ratio = a1

(f)

Gini Index:

$$Gini(T) = 1 - (3/4)^2 - (1/4)^2 = 0.375, \quad Gini(F) = 1 - (1/5)^2 - (4/5)^2 = 0.32$$

$$Gini(a_1) = \frac{4}{9}(0.375) + \frac{5}{9}(0.32) = 0.344$$

$$Gini(T) = 1 - (3/6)^2 - (3/6)^2 = 0.5, \quad Gini(F) = 1 - (1/3)^2 - (2/3)^2 = 0.444$$

$$Gini(a_2) = \frac{6}{9}(0.5) + \frac{3}{9}(0.444) = 0.481$$

Best split by Gini = a1

(g)

Classification Error:

$$Error(a_1) = \frac{4}{9}\left(\frac{1}{4}\right) + \frac{5}{9}\left(\frac{1}{5}\right) = 0.222$$

$$Error(a_2) = \frac{6}{9}\left(\frac{3}{6}\right) + \frac{3}{9}\left(\frac{1}{3}\right) = 0.444$$

Best split by Classification Error = a1

ANS 2

(a)

Root Node:

$$\text{Total C1} = 100, \quad \text{Total C2} = 100, \quad \text{Error} = \frac{\min(100,100)}{200} = 0.5$$

Calculating Info Gain for X, Y, Z:

X	C1	C2	Error
0	60	60	0.5
1	40	40	0.5

$$Gain(X) = 0$$

Y	C1	C2	Error
0	40	60	0.4
1	60	40	0.4

$$Gain = 0.5 - (0.5 \cdot 0.4 + 0.5 \cdot 0.4) = 0.5 - 0.4 = 0.1$$

Z	C1	C2	Error
0	95	65	0.406
1	5	35	0.125

$$Gain = 0.5 - \left(\frac{160}{200}(0.406) + \frac{40}{200}(0.125)\right) = 0.5 - 0.36 = 0.14$$

Choosing the best split:

Best Root: Z

Expand left child of Z (Z=0)
Subset where Z=0:

Y	X	C1	C2
0	0	5	40
0	1	10	5
1	0	10	5
1	1	25	0
1	1	45	0

Total = 95C1, 65C2

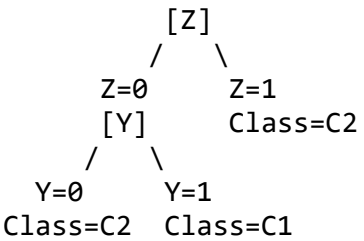
Try splitting on Y (Z=0 subset):

Y=0: (5+10 C1, 40+5 C2) → 15 C1, 45 C2 → error = 0.25

Y=1: (10+25+45 C1, 5+0+0 C2) → 80 C1, 5 C2 → error = 0.0588

$$Weightederror = \frac{60}{160}(0.25) + \frac{100}{160}(0.0588) = 0.093$$
$$\Rightarrow Gain = 0.406 - 0.093 = 0.313$$

Decision Tree:



(b)

Leaf nodes:

Z=1 → 5 C1, 35 C2 → misclassified = 5

Z=0, Y=0 → 15 C1, 45 C2 → misclassified = 15

Z=0, Y=1 → 80 C1, 5 C2 → misclassified = 5

$$\text{Total Error} = 5 + 15 + 5 = 25$$
$$\Rightarrow \quad \frac{25}{200} = 0.125$$

Error rate = 12.5%

(c)

- Root: Z → Gain = 0.14
- Second level (Z=0): Y → Gain = 0.313
- X was not used

Importance(Z)=0.14,Importance(Y)=0.313,Importance(X)=0

Ranking:

1. Y (0.313)
2. Z (0.14)
3. X (0)

ANS 3

(a)

Generalization Error (Optimistic)
Use training data. Count misclassified instances:
Use the tree:
If A = 0 → check B
If A = 1 → check C

Instance	A	B	C	True	Predicted	Match
1	0	0	0	+	+	✓
2	0	0	1	+	+	✓
3	0	1	0	+	−	✗
4	0	1	1	−	−	✓
5	1	0	0	+	+	✓
6	1	0	0	+	+	✓

7	1	0	1	−	−	✓
8	1	1	0	+	+	✓
9	1	1	0	−	+	✗
10	1	1	0	−	+	✗

$$\text{Error}_{opt} = \frac{3}{10} = 0.3$$

(b)

Generalization Error (Pessimistic)

There are 4 leaf nodes. Add 0.5 error to each leaf.

$$E_{pess} = \frac{3 + 0.5 \times 4}{10} = \frac{5}{10} = 0.5$$

(c)

Reduced Error Pruning (Validation Set)

Use validation set to test tree:

Instance	A	B	C	True	Predicted	Match
11	0	0	0	+	+	✓
12	0	1	1	+	−	✗
13	1	1	0	+	+	✓
14	1	0	1	−	−	✓
15	1	0	0	+	+	✓

$$\text{Error}_{val} = \frac{1}{5} = 0.2$$

(d)

Test Set Classification and Accuracy

Inst	A	B	C	True	Pred	TP	TN	FP	FN
16	0	1	0	+	−				1
17	1	0	0	+	+	1			
18	1	1	1	−	−		1		
19	1	0	1	+	−				1
20	1	1	1	−	−		1		
21	0	0	1	−	+			1	
22	1	0	0	+	+	1			
23	0	0	1	+	+	1			

$$\text{Accuracy} = \frac{TP + TN}{Total} = \frac{5}{8} = 0.625$$

(e)

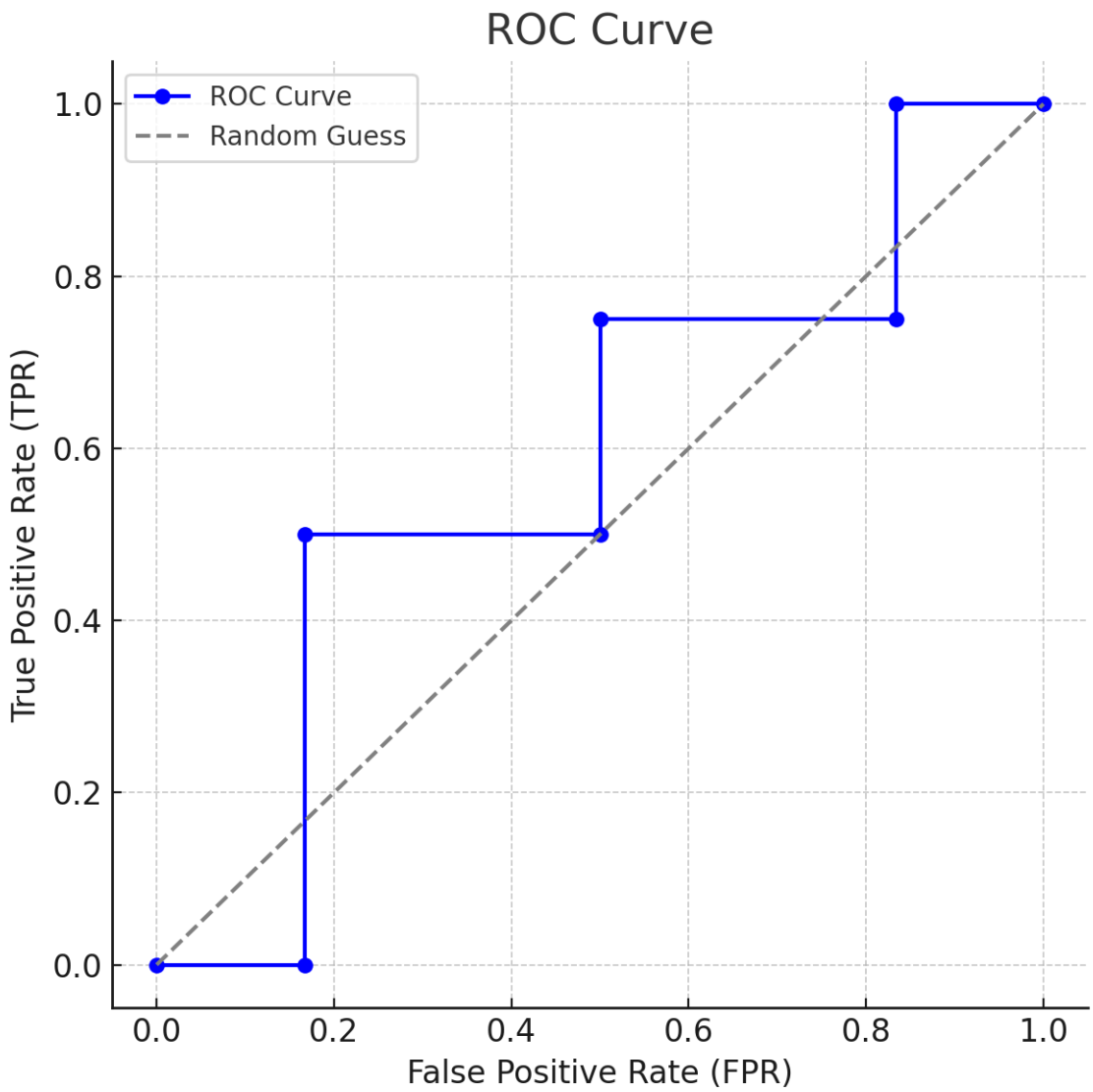
Precision, Recall, F1:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 1} = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{3}{3 + 2} = 0.6$$
$$F1 = \frac{2 \cdot 0.75 \cdot 0.6}{0.75 + 0.6} = \frac{0.9}{1.35} \approx 0.667$$

ANS 4

ROC Curve:



ROC Table Used:

	0.01	0.03	0.04	0.05	0.09	0.31	0.38	0.45	0.61	0.68
TP	5	5	5	5	4	3	3	3	2	1
FP	5	4	4	3	3	3	2	1	0	0
TN	0	1	1	2	2	2	3	4	5	5
FN	0	0	0	0	1	2	2	2	3	4
TPR	1	1	1	1	0.8	0.6	0.6	0.6	0.4	0.2
FPR	1	0.8	0.8	0.6	0.6	0.6	0.4	0.2	0	0

ANS 5 & 6

Link to Github