

**CS 5990 (Advanced Data Mining) - Assignment #2**  
**Maximum Points: 100 pts.**

**Bronco ID: 017314921**

**Last Name: Panda**

**First Name: Subham**

**Answers:**

**1) Binary Classification Problem:**

a) The entropy of the collection of training examples with respect to the class attribute

$$Entropy = - \sum_{i=0}^{c-1} p_i(t) \log_2 p_i(t)$$

$$p(+) = \frac{4}{9} = 0.44$$

$$p(-) = \frac{5}{9} = 0.55$$

$$\begin{aligned} \therefore E &= -0.44 \log_2 0.44 - 0.55 \log_2 0.55 \\ &= -0.44(-1.1844) - 0.55(-0.8625) \\ &= 0.52 + 0.47 \\ &= 0.9911 \end{aligned}$$

**b)** The information gains of  $a_1$  and  $a_2$  relative to the training examples is  
Information Gain = P-M

Information Gain:

$$Gain_{split} = Entropy(p) - \sum_{i=1}^k \frac{n_i}{n} Entropy(i)$$

Parent node,  $p$  is split into  $k$  partitions (children)

$n_i$  is number of records in child node  $i$

$$Entropy(p) = 0.9911$$

$$E(a_1) = \frac{4}{9} \left[ -\frac{3}{4} \log_2 \left( \frac{3}{4} \right) - \frac{1}{4} \log_2 \left( \frac{1}{4} \right) \right] + \frac{5}{9} \left[ -\frac{1}{5} \log_2 \left( \frac{1}{5} \right) - \frac{4}{5} \log_2 \left( \frac{4}{5} \right) \right]$$
$$= 0.7616$$

$$\therefore Information\ Gain(a_1) = Entropy(p) - E(a_1)$$

$$= 0.9911 - 0.7616 = 0.2294$$

$$E(a_2) = \frac{4}{9} \left[ -\frac{2}{4} \log_2 \left( \frac{2}{4} \right) - \frac{2}{4} \log_2 \left( \frac{2}{4} \right) \right] + \frac{5}{9} \left[ -\frac{2}{5} \log_2 \left( \frac{2}{5} \right) - \frac{3}{5} \log_2 \left( \frac{3}{5} \right) \right]$$
$$= 0.9839$$

$$\therefore Information\ Gain(a_2) = Entropy(p) - E(a_2)$$

$$= 0.9911 - 0.9839 = 0.0072$$

c) For  $a_3$ , the information gain for every possible split using the efficient computation technique which includes sorting, splitting, filling class distribution

<i>For <math>a_3</math></i>	+	−	+	− −	+	− +	−
<i>Sorted Values</i>	1.0	3.0	4.0	5.0 5.0	6.0	7.0 7.0	8.0
<i>Split Values</i>	0.5	2.0	3.5	4.5	5.5	6.5	7.5
+		1 3	1 3	2 2	2 2	3 1	4 0
−		0 5	1 4	1 4	3 2	3 2	4 1
<i>Entropy(<math>a_3</math>)</i>		0.8484	0.9885	0.9183	0.9838	0.9728	0.8889
<i>Information Gain(<math>a_3</math>)</i>		0.1427	0.0026	0.0728	0.0072	0.0183	0.1022

∴ The best split for  $a_3$  occurs at split value 2.0 with an information gain of 0.1427

**d)** According to the information gain  $a_1$  produces the best split.

e)

$$\text{Gain Ratio} = \frac{\text{Gain}_{\text{split}}}{\text{Split Info}} \quad \text{Split Info} = - \sum_{i=1}^k \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

$$\begin{aligned} SI(a_1) &= -\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} \\ &= -0.4444(-1.17) - 0.5555(-0.848) \\ &= 0.991 \end{aligned}$$

$$SI(a_2) = 0.991$$

$$\begin{aligned} SI(a_3) &= -\frac{1}{9} \log_2 \frac{1}{9} - \frac{8}{9} \log_2 \frac{8}{9} \\ &= -0.1111(-3.17) - 0.8888(-0.16993) \\ &= 0.503 \end{aligned}$$

$$GR(a_1) = \frac{0.2294}{0.991} = 0.231$$

$$GR(a_2) = \frac{0.0072}{0.991} = 0.0072$$

$$GR(a_3) = \frac{0.1427}{0.991} = 0.2836$$

$\therefore a_3$  has best split according to gain ratio as per information gain.

f)

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

$$GI(a_1) = \frac{4}{9} \left[ 1 - \left( \frac{3}{4} \right)^2 - \left( \frac{1}{4} \right)^2 \right] + \frac{5}{9} \left[ 1 - \left( \frac{1}{5} \right)^2 - \left( \frac{4}{5} \right)^2 \right] = 0.3444$$

$$GI(a_2) = \frac{5}{9} \left[ 1 - \left( \frac{2}{5} \right)^2 - \left( \frac{3}{5} \right)^2 \right] + \frac{4}{9} \left[ 1 - \left( \frac{2}{4} \right)^2 - \left( \frac{2}{4} \right)^2 \right] = 0.4889$$

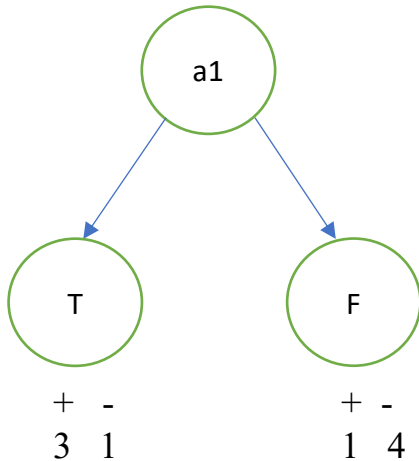
$\therefore$  Gini Index of  $a_1$  is smaller & it produces better split.

g)

Classification Error Rate

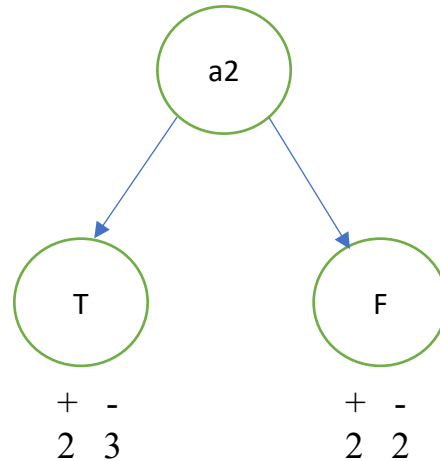
$$\text{Classification error} = 1 - \max[p_i(t)]$$

But using shortcut method:



$$\text{Minority} = 1 + 1 = 2$$

$$E(t)a_1 = \frac{2}{9} = 0.222$$



$$\text{Minority} = 1 + 1 = 2$$

$$E(t)a_2 = \frac{4}{9} = 0.444$$

Since  $E(t)$  of  $a_1$  is less than  $a_2$

$\therefore a_1$  will provide best split according to the classification error rate.

## 2) Two-level decision tree:

a)

$X$	$C_1$	$C_2$
<b>0</b>	60	60
<b>1</b>	40	40

$$Class_{err}(X) = \frac{40 + 40}{200} = 0.4$$

$Y$	$C_1$	$C_2$
<b>0</b>	40	60
<b>1</b>	60	40

$$Class_{err}(Y) = \frac{40 + 40}{200} = 0.4$$

$Z$	$C_1$	$C_2$
<b>0</b>	30	70
<b>1</b>	70	30

$$Class_{err}(Z) = \frac{30 + 30}{200} = 0.3$$

$Z$  is the lowest in terms of classification error and is chosen as the splitting attribute at level 1.

Now,  $Z = 0$

$X$	$C_1$	$C_2$
<b>0</b>	15	45
<b>1</b>	25	25

$Y$	$C_1$	$C_2$
<b>0</b>	15	45
<b>1</b>	15	25

$$Class_{err}Z_0(X, Y) = \frac{15 + 15}{100} = 0.3$$

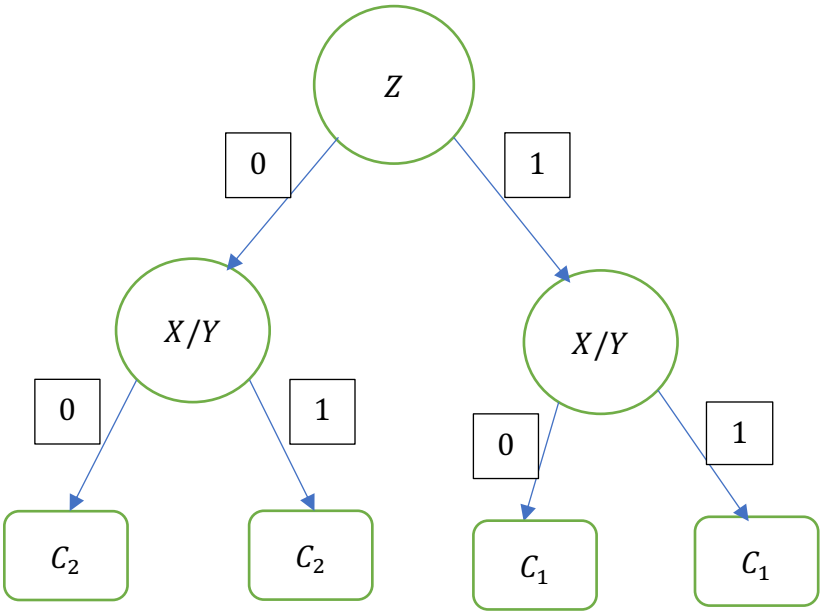
Now,  $Z = 1$

$X$	$C_1$	$C_2$
<b>0</b>	45	15
<b>1</b>	25	15

$Y$	$C_1$	$C_2$
<b>0</b>	25	15
<b>1</b>	45	15

$$Class_{err}Z_1(X, Y) = \frac{15 + 15}{100} = 0.3$$





**b)** Overall error rate of the induced tree is  $\frac{(15+15+15+15)}{200} = 0.3$

### 3) Model decision tree:

Ground Truth	$B +$		$B -$		$C +$		$C -$	
Predicted	+	−	+	−	+	−	+	−
Training - ID	1, 2	NA	3	4	5, 6	7, 9, 10	8	NA
Validation - ID	11	NA	12	NA	13, 15	NA	NA	14
Test - ID	23	21	16	NA	17, 22	NA	19	18, 20

a) The generalization error rate of the tree using the optimistic approach

$$err_{tr} = \frac{5}{10} = 0.5$$

b)

Ground Truth	$B +$		$B -$		$C +$		$C -$	
Predicted	+	−	+	−	+	−	+	−
Training - ID	1, 2	NA	3	4	5, 6	7, 9, 10	8	NA
Validation - ID	11	NA	12	NA	13, 15	NA	NA	14
Test - ID	23	21	16	NA	17, 22	NA	19	18, 20

The generalization error rate of the tree using the pessimistic approach with factor of 0.5

$$err_{gen}(T) = err(T) + \Omega \times \frac{k}{N_{train}}$$

$$err_{gen}(T) = \frac{5}{10} + 0.5 \left( \frac{4}{10} \right) = 0.7$$

c)

Ground Truth	<i>B</i> +		<i>B</i> −		<i>C</i> +		<i>C</i> −	
Predicted	+	−	+	−	+	−	+	−
Training - ID	1, 2	NA	3	4	5, 6	7, 9, 10	8	NA
Validation - ID	11	NA	12	NA	13, 15	NA	NA	14
Test - ID	23	21	16	NA	17, 22	NA	19	18, 20

The generalization error rate of the tree using the validation set

$$err_{gen}(V) = \frac{1}{5} = 0.2$$

d) Accuracy of the model on the test set

	<i>ID</i>	<i>Total</i>
<i>True Positive (TP)</i>	17,22,23	3
<i>True Negative (TN)</i>	18,20	2
<i>False Positive (FP)</i>	16,19	2
<i>False Negative (FN)</i>	21	1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3 + 2}{3 + 2 + 2 + 1} = 0.625$$

e)

	<i>ID</i>	<i>Total</i>
<i>True Positive (TP)</i>	17,22,23	3
<i>True Negative (TN)</i>	18,20	2
<i>False Positive (FP)</i>	16,19	2
<i>False Negative (FN)</i>	21	1

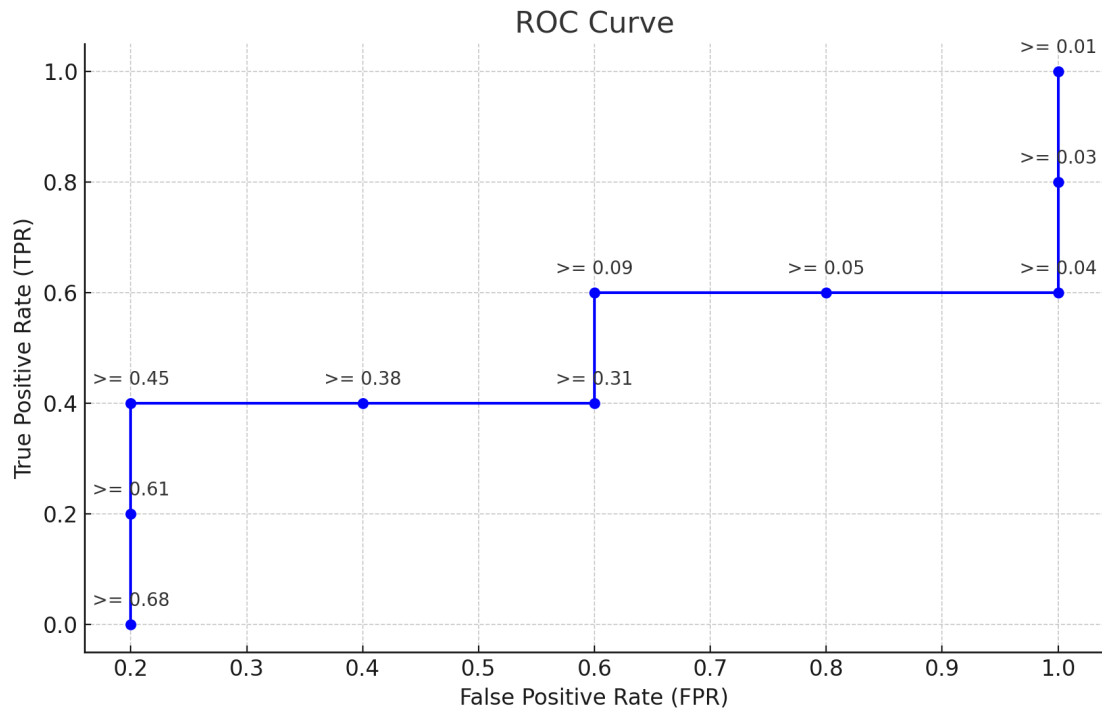
$$\text{Precision (P)} = \frac{TP}{TP + FN} = \frac{3}{3 + 2} = 0.6$$

$$\text{Recall (R)} = \frac{TP}{TP + FN} = \frac{3}{3 + 1} = \frac{3}{4} = 0.75$$

$$F1 \text{ Score} = 2 * \frac{PR}{P + R} = \frac{2(0.6)(0.75)}{0.6 + 0.75} = \frac{0.9}{1.35} = 0.667$$

#### 4) ROC Curve:

Class	+	+	-	-	+	-	-	+	+	-
Threshold ≥	0.01	0.03	0.04	0.05	0.09	0.31	0.38	0.45	0.61	0.68
<b>TP</b>	5	4	3	3	3	2	2	2	1	0
<b>FP</b>	5	5	5	4	3	3	2	1	1	1
<b>TN</b>	0	0	0	1	2	2	3	4	4	4
<b>FN</b>	0	1	2	2	2	3	3	3	4	5
<b>TPR</b>	1.0	0.8	0.6	0.6	0.6	0.4	0.4	0.4	0.2	0.0
<b>FPR</b>	1.0	1.0	1.0	0.8	0.6	0.6	0.4	0.2	0.2	0.2





## 5) Python program (decision tree.py)

6) **Python program (roc\_curve.py)**