CSE 5243

Homework 3

Name: ***Piyush Chawla***

Q1.

		Predicted Class		
		+	-	Total
Actual	+	350	122	472
Class	-	344	670	1014
	Total	694	792	1486

Accuracy Rate = (True Positive + True Negative)/Total Data
= (350 + 670)/1486
= 0.6864

Error Rate = 1 - Accuracy Rate
= 1 - 0.6864
= 0.3136

True Positive Rate = (True Positive)/(True Positive + False Negative)
= 350/(350+122)
= 0.7415

False Positive Rate = (False Positive)/(False Positive + True Negative)
= 344/(344+670)
= 0.3392

Precision = (True Positive)/(True Positive + False Positive)
= 350/(350+344)
= 0.5043

Recall = (True Positive)/(True Positive + False Negative)
= 350/(350+122)
= 0.7415

F-Measure = (2*Precision*Recall)/(Precision + Recall)
= (2*0.5043*0.7415)/(0.7415 + 0.5043)
= 0.6003

Q2.

	a_1	a_2	a_3	Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a) **Entropy** = $-\sum p_i log(p_i)$

= - (4/9)*log(4/9) - (5/9)*log(5/9)
= 0.9911

b)

i) Analysis for a1

The table below shows distribution of class labels in the leaf nodes after branching on a1

a1 values	**+**	**-**
T	3	1
F	1	4

Entropy = $-\sum \frac{|S_j|}{|S|} p_i log(p_i)$

= (5/9)*(-(⅓)log(⅓) - (⅔)log(⅔)) + (4/9)*(-(¾)log(¾) - (¼)log(¼))
= 0.7616

Gain = 0.9911 - 0.7616
= 0.2295

Intrinsic Info = -(5/9)*log(5/9) - (4/9)*log(4/9)
= 0.9911

Gain ratio = 0.2295/0.9911
= 0.2315

ii) Analysis for a2

The table below shows class label distributions in two leaf nodes (+ and -) after branching on a2

a2 values	**+**	**-**
T	2	3
F	2	2

Entropy = $-\sum \frac{|S_j|}{|S|} p_i log(p_i)$

= (5/9)*(-(2/5)*log(2/5) - (3/5)*log(3/5)) + (4/9)*(-(1/2)*log(1/2) - (1/2)*log(1/2))
= 0.9839

Gain = 0.9911 - 0.9839 = 0.0072

Intrinsic Info = -(5/9)*log2(5/9) - (4/9)*log2(4/9)
= 0.9911

Gain ratio = 0.0072/0.9911
= 0.0073

c)

The table below gives Entropy, Gain, Intrinsic-info and Gain-ratio values for different thresholds.

For instance, for threshold 1 we compute the values as follows:

Entropy = $-\sum \frac{|S_j|}{|S|} p_i log(p_i)$

= (1/9)*(-(1/1)*log(1/1) - (0/1)*log(0/1)) + (8/9)*(-(3/8)*log(3/8) - (5/8)*log(5/8))
= 0.8484

Gain = 0.9911 - 0.8484 = 0.1427

Intrinsic Info = -(1/9)*log2(1/9) - (8/9)*log2(8/9)

= 0.5033

Gain ratio = 0.1427/0.5033
= 0.2835

	<=Threshold		>Threshold					
Threshold	**+**	**-**	**+**	**-**	**Entropy**	**Gain**	**Intrinsic Info**	**Gain Ratio**
1	1	0	3	5	0.8484	0.1427	0.5033	0.2835
3	1	1	3	4	0.9885	0.0026	0.7642	0.0034
4	2	1	2	4	0.9183	0.0728	00.9183	0.0793
5	2	3	2	2	0.9839	0.0072	0.9911	0.0073
6	3	3	1	2	0.9728	0.0183	0.9183	0.0199
7	4	4	0	1	0.8889	0.1022	0.5033	0.2031
8	4	5	0	0	0.6869	0.9911	0	NaN

d)

We see that branching on a1 gives the highest information gain (0.2295). So we use a1.

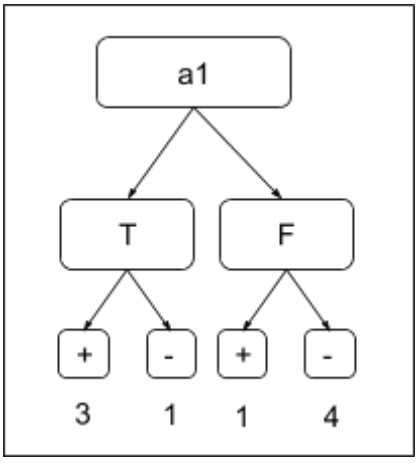

Figure: Decision tree after branching on a1

e)

We see that a3 with threshold value 1 has the largest gain ratio (0.2835). So using the gain ratio criterion, we would choose that instead of a1.

Q3.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Step1: Choosing the attribute for the first split

Table below shows leaf nodes after the first branching on Gender.

Gender	**C0**	**C1**
M	6	4
F	4	6

Gini Index = $\frac{|S_j|}{|S|}(1 - \sum p_i^2)$

= (10/20)*(1 - .6*.6 - .4*.4) + (10/20)*(1 - .6*.6 - .4*.4)
= 0.48

The table below shows leaf nodes after first branching on Car Type

Car Type	C0	C1
Family	1	3
Sports	8	0
Luxury	1	7

Gini Index = $\frac{|S_j|}{|S|}(1 - \sum {p_i}^2)$

= (4/20)*(1 - .25*.25 - .75*.75) + (8/20)*(1 - 1*1 - 0*0) + (8/20)*(1 - 1/8*1/8 - 7/8*7/8)

= 0.1625

Table below shows leaf nodes after first branching on Shirt Size

Shirt Size	C0	C1
Small	3	2
Medium	3	4
Large	2	2
Extra Large	2	2

Gini Index = $\frac{|S_j|}{|S|}(1 - \sum {p_i}^2)$

= (5/20)*(1 - .6*.6 - .4*.4) + (7/20)*(1 - (3/7)*(3/7) - (4/7)*(4/7)) + (4/20)*(1 - .5*.5 - .5*.5) +(4/20)*(1 - 0.5*0.5-0.5*0.5)

= 0.4914

Therefore, we choose Car Type for the first split.

Step 2: Second Split

Table below shows leaf nodes obtained after second branching on Gender

Car Type/Gender	C0	C1
Family/M	1	3
Family/F	0	0

Luxury/M	0	1
Luxury/F	1	6
Sports/M	5	0
Sports/F	3	0

Gini Index = $\frac{|S_j|}{|S|}(1 - \sum {p_i}^2)$

= (4/20)*(1 - .25*.25 - .75*.75) + (1/20)*(1 - 0 - 1) +(7/20)*(1 - (1/7)*(1/7)-(6/7)*(6/7)) + (5/20)*(1 - 1) + (3/20)*(1 - 1)

= 0.1607

Table below shows the second branching using Shirt Size.

Car Type/Shirt Size	**C0**	**C1**
Family/Small	1	0
Family/Medium	0	1
Family/Large	0	1
Family/Extra Large	0	1
Luxury/Small	0	2
Luxury/Medium	0	3
Luxury/Large	1	1
Luxury/Extra Large	0	1
Sports/Small	2	0
Sports/Medium	3	0
Sports/Large	1	0
Sports/Extra Large	2	0

Gini Index = $\frac{|S_j|}{|S|}(1 - \sum {p_i}^2)$

= 4*(1/20)*(1 - 1) + (2/20)*(1 - 1) +(3/20)*(1 - 1) + (2/20)*(1 - .5*.5-.5*.5) + (1/20)*(1-1) + (2/20)*(1 -1) + (3/20)*(1 - 1) + (1/20)(1-1) + (2/20)*(1 - 1)

= 0.05

Shirt size gives the minimum gini index (0.05) so we choose Shirt Size as second attribute for decision tree.

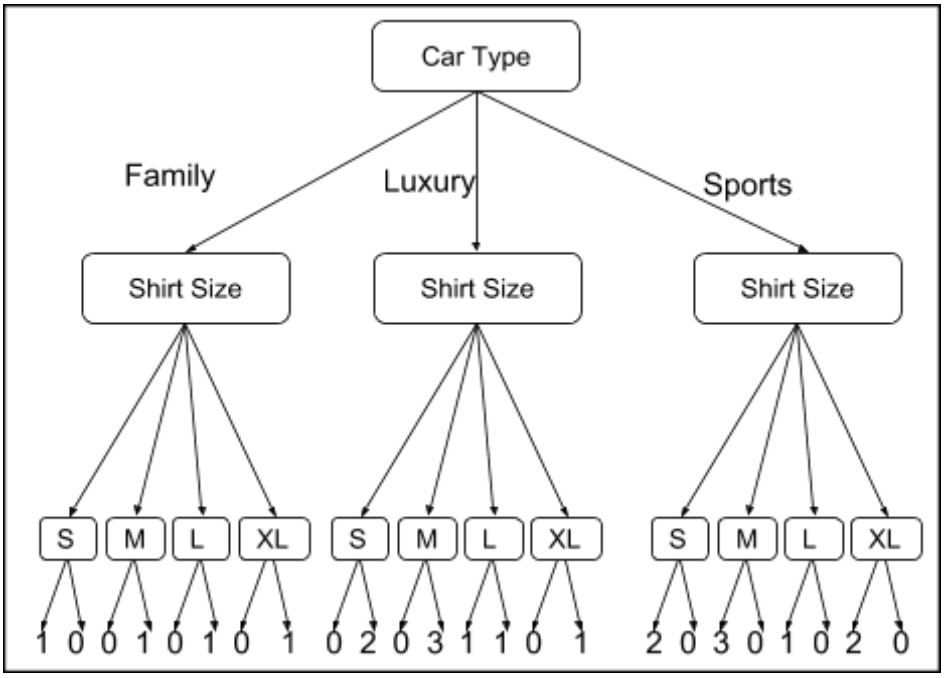

Figure: Final decision tree obtained after the two steps

Q4.

Score Range P(+ I x)	# + Class Examples in the score range	# - Class Examples in the score range	Total Examples in each score range
[0.9 - 1]	118	13	131
[0.8 - 0.9)	94	24	118
[0.7 - 0.8)	81	35	116
[0.6 - 0.7)	74	42	116
[0.5 - 0.6)	51	48	99
[0.4 - 0.5)	34	62	96
[0.3 - 0.4)	25	99	124
[0.2 - 0.3)	11	105	116
[0.1 - 0.2)	9	122	131
[0 - 0.1)	3	150	153
Total	500	700	1200

a)

True Positives = 118+94+81+74+51 = 418
False Positives = 13+24+35+42+48 = 162

True Negatives = 700 - 162 = 538
False Negatives = 500 - 418 = 82

	Predicted True	**Predicted False**
Actual True	418	82
Actual False	162	538

Accuracy = (TP + FP)/(Total)
= (418 + 538)/1200
= .7967

TPR = TP/(TP + FN)
= 418/500

= 0.8360

FPR = FP/(FP + TN)
= 162/700
= 0.2314

b)

True Positives = 118+94+81 = 293
False Positives = 13+24+35 = 72

True Negatives = 700 - 72 = 628
False Negatives = 500 - 293 = 207

	Predicted True	**Predicted False**
Actual True	293	207
Actual False	72	628

Accuracy = (TP + TN)/(Total)
= (293 + 628)/1200
= 0.7675

TPR = TP/(TP + FN)
= 293/500
= .586

FPR = FP/(FP + TN)
= 72/700
= 0.1029

We observe that both the values, TPR and FPR, reduce when change the threshold to 0.7 from 0.5

c)

The table below shows TPR and FPR values for different thresholds, similar to the calculations above.

Threshold	**Cumulative +**	**Cumulative -**	**TPR**	**FPR**
0.9	118	13	0.236	0.0186
0.8	212	37	0.424	0.0529
0.7	293	72	0.586	0.103

0.6	367	114	0.734	0.163
0.5	418	162	0.836	0.231
0.4	452	224	0.904	0.32
0.3	477	323	0.954	0.461
0.2	488	428	0.976	0.611
0.1	497	550	0.994	0.786
0	500	700	1	1

Plot: FPR vs TPR (ROC Curve)