Assignment # 2

1. Consider the training examples shown in Table 1 for a binary classification problem.

(a) Compute the Gini index for the overall collection of training examples.
(b) Compute the Gini index for the Customer ID attribute.
(c) Compute the Gini index for the Gender attribute.
(d) Compute the Gini index for the Car Type attribute using multiway split.
e) Compute the Gini index for the Shirt Size attribute using multiway split.
(f) Which attribute is better, Gender, Car Type, or Shirt Size?

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

Table 1

2. Consider the training examples shown in Table 2 for a binary classification problem.

(a)What is the entropy of this collection of training examples with respect to the positive class?
(b) What are the information gains of $a1$ and $a2$ relative to these training examples?
(c) For $a3$, which is a continuous attribute, compute the information gain for every possible split.

(d) What is the best split (among $a1$, $a2$, and $a3$) according to the information gain?

Instance	a_1	a_2	a_3	Target 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	−

Table 2

(e) What is the best split (between $a1$ and $a2$) according to the classification error rate?

(f) What is the best split (between $a1$ and $a2$) according to the Gini index?

3. Consider the following data set shown in Table 3 for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Table 3

(a) Calculate the information gain when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

(b) Calculate the gain in the Gini index when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

4. Consider the following set of training examples shown in Table 4.

(a) Compute a two-level decision tree using the greedy approach described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

(b) Repeat part (a) using *X* as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

X	*Y*	*Z*	No. of Class C1 Examples	No. of Class C2 Examples
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

Table 4

5. C4.5rules is an implementation of an indirect method for generating rules from a decision tree. RIPPER is an implementation of a direct method for generating rules directly from data.

(a) Discuss the strengths and weaknesses of both methods.

(b) Consider a data set that has a large difference in the class size (i.e., some classes are much bigger than others). Which method (between C4.5rules and RIPPER) is better in terms of finding high accuracy rules for the small classes?