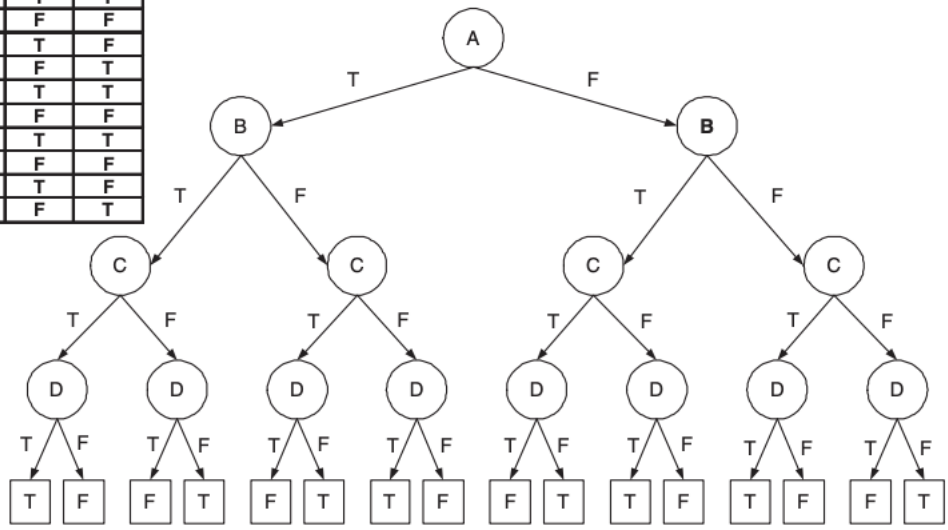
HW 2

Question 1:
Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?

| A | B | C | D | Class |
|---|---|---|---|---|
| T | T | T | T | T |
| T | T | T | F | F |
| T | T | F | T | F |
| T | T | F | F | T |
| T | F | T | T | F |
| T | F | T | F | T |
| T | F | F | T | T |
| T | F | F | F | F |
| F | T | T | T | F |
| F | T | T | F | T |
| F | T | F | T | T |
| F | T | F | F | F |
| F | F | T | T | T |
| F | F | T | F | F |
| F | F | F | T | F |
| F | F | F | F | T |




Can not be simplified.

Question 2:
Consider the training examples show in the below table 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.

| 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 |

(a). Gini = $1 - 2 \times 0.52 = 0.5$
(b). The gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0.
(c). The gini for Male is $1 - 2 \times 0.52 = 0.5$. The gini for Female is also 0.5. Therefore, the overall gini for Gender is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

Question 3: Consider the following dataset for a binary class problem.
Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

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

The contingency tables after splitting on attributes $A$ and $B$ are:

| | $A=T$ | $A=F$ |
|---|---|---|
| + | 4 | 0 |
| − | 3 | 3 |

| | $B=T$ | $B=F$ |
|---|---|---|
| + | 3 | 1 |
| − | 1 | 5 |

The overall entropy before splitting is:

$$E_{orig} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.9710$$

The information gain after splitting on A is:

$$\begin{aligned} E_{A=T} &= -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.9852 \\ E_{A=F} &= -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = 0 \\ \Delta &= E_{orig} - 7/10 E_{A=T} - 3/10 E_{A=F} = 0.2813 \end{aligned}$$

The information gain after splitting on B is:

$$\begin{aligned} E_{B=T} &= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \\ E_{B=F} &= -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.6500 \\ \Delta &= E_{orig} - 4/10 E_{B=T} - 6/10 E_{B=F} = 0.2565 \end{aligned}$$

Therefore, attribute $A$ will be chosen to split the node.

Question 4:
C4.5 rules 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 dataset that has a large difference in the class size (i.e., some classes are much bigger than others). Which method (between C4.5 rules and RIPPER) is better in terms of finding high accuracy rules for the small classes?

(a). The C4.5 rules algorithm generates classification rules from a global perspective. This is because the rules are derived from decision trees, which are induced with the objective of partitioning the feature space into homogeneous regions, without focusing on any classes. In contrast, RIPPER generates rules one-class-at-a-time. Thus, it is more biased towards the classes that are generated first.

(b). The class-ordering scheme used by C4.5rules has an easier interpretation than the scheme used by RIPPER.

Question 5:
Consider a binary classification problem with the following set of attributes and attribute values.
Air conditioner = {working, broken}
Engine = {Good, Bad}
Mileage = {High, Medium, Low}
Rust = {Yes, No}

Suppose a rule-based classifier produces the following rule set:
(a). Are the rules mutually exclusive?
(b). Is the rule set exhaustive?
(c). Is ordering needed for this set of rules?
(d). Do you need a default class for the rule set?

$$\text{Mileage} = \text{High} \longrightarrow \text{Value} = \text{Low}$$
$$\text{Mileage} = \text{Low} \longrightarrow \text{Value} = \text{High}$$
$$\text{Air Conditioner} = \text{Working}, \text{Engine} = \text{Good} \longrightarrow \text{Value} = \text{High}$$
$$\text{Air Conditioner} = \text{Working}, \text{Engine} = \text{Bad} \longrightarrow \text{Value} = \text{Low}$$
$$\text{Air Conditioner} = \text{Broken} \longrightarrow \text{Value} = \text{Low}$$

(a). No
(b). Yes
(c). Yes, because a test instance may trigger more than one rule.
(d). No, because every instance is guaranteed to trigger at least one rule.

Question 6:
Given the attached data, use Weka to explore the possible association rules as well as classify and predict data. Summarize, analyze and interpret the results.