**Name: Nikhil Jadhav (801075504)**

**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Tree**

**Part 1**

Consider the training examples shown in Table below for a binary classification

problem.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer ID | Gender | Car Type | 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) 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?

(g) Explain why Customer ID should not be used as the attribute test

condition even though it has the lowest Gini

**Answers**

1. Gini index for overall collection of training examples is as follows:
2. The Customer ID column value is unique for each row therefore, the Gini Index for each value is 0. Hence, the Gini Index for Customer ID is 0.
3. We have two genders values (Male, Female) in the given problem table. Therefore we will calculate Gini index for those individual values and then use it to find the Gini Index for Gender

Gini index for male:

Gini Index for female:

The Gini Index for Gender is:

1. We have three distinct values for Car Type attribute i.e. Family, Sports, Luxury. Therefore, we will calculate Gini Index for each unique value.

Gini Index for Family car:

Gini Index for Sports car:

Gini Index for Luxury car:

Total Gini Index for Car Type attribute

1. Using the method used in previous example

Gini Index for Small size shirt:

Gini Index for Medium size shirt:

Gini Index for Large size shirt:

Gini Index for Extra Large size shirt:

Total Gini Index for Size attribute:

1. Car Type has the least Gini Index compared to other attributes. Therefore, Car Type attribute is better.
2. Gini index is the statistical measure of distribution of one attribute W.R.T other attributes. Since, the ID column has unique entry for every new record using it as a measure for attribute test would not be feasible.