**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Tree**

**Part 1**

**NAME: Ashish Bisht**

**Student ID: 801168390**

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**.

Ans:

Gini index is the difference of 1 and the sum of the fraction of records belonging to j at given node k,

Gini = 1 - p(C0|Class)2 - p(C1|Class)2

= 1 - (10/20)2 - (10/20)2

 = 0.5

**(b) Compute the Gini index for the Customer ID attribute.**

Ans: The customer ID produces it so we can have 20 Gini values that look like,

Gini(1) = 1 - (1/1)2 - (0/1)2

= 0

**(c) Compute the Gini index for the Gender attribute**

Ans:

Gini(Male) = 1 - p(C0|M)2 - p(C1|M)2

= 1 - (6/10)2 - (4/10)2

 = 0.48

Gini(Female) = 1 - p(C0|F)2 - p(C1|F)2

= 1 - (6/10)2 - (4/10)2

= 0.48

Gini(Gender) = [(T(Male)/T(Male + female)]\*Gini(Male) + [(T(Female)/T(Male + Female)]\*Gini(Female)

= (10/20)\*0.48 + (10/20)\*0.48

= 0.48

**(d) Compute the Gini index for the Car Type attribute using multiway**

**split.**

Ans:

Gini(Family) = 1 - p(C0|Family)2 - p(C1|Family)2

= 1 - (1/4)2 - (3/4)2

= 0.375

Gini(Sports) = 1 - p(C0|Sports)2 - p(C1|Sports)2

= 1 - (8/8)2 - (0/8)2

= 0

Gini(Luxury) = 1 - p(C0|Luxury)2 - p(C1|Luxury)2

= 1 - (1/8)2 - (7/8)2

= 0.2188 

TGini(Car Type) = [(T(Family)/T(Car Type)]\*Gini(Family) + [(T(Sports)/T(Car Type)]\*Gini(Sports) + [(T(Luxury)/T(Car Type)]\*Gini(Luxury)

= (4/20)\*0.375 + (8/20)\*0 + (8/20)\*0.2188

= 0.1625

**(e) Compute the Gini index for the Shirt Size attribute using multiway**

**split.**

Ans:

Gini(Small) = 1 - p(C0|Small)2 - p(C1|Small)2

= 1 - (3/5)2 - (2/5)2

= 0.48 

Gini(Medium) = 1 - p(C0|Medium)2 - p(C1|Medium)2

= 1 - (3/7)2 - (4/7)2

= 0.4898 

Gini(Large) = 1 - p(C0|Large)2 - p(C1|Large)2

= 1 - (2/4)2 - (2/4)2

= 0.5

Gini(Extra Large) = 1 - p(C0|Extra Large)2 - p(C1|Extra Large)2

= 1 - (2/4)2 - (2/4)2

= 0.5

TGini(Shirt Size) = [(T(Small)/T(Shirt Size)]\*Gini(Small) + [(T(Medium)/T(Shirt Size)]\*Gini(Medium) + [(T(Large)/T(Shirt Size)]\*Gini(Large) + [(T(Extra Large)/T(Shirt Size)]\*Gini(Extra Large)

= (5/20)\*0.48 + (7/20)\*0.4898 + (4/20)\*0.5+ (4/20)\*0.5

= 0.4914

**(f) Which attribute is better, Gender, Car Type, or Shirt Size?**

Ans:   
Gender = 0.48

Car Type = 0.1625

Shirt Size = 0.4914.

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

**condition even though it has the lowest Gini**

Ans: Customer ID s is distinctive so it cannot be used as a predictive attribute.

As it splits it into all the possible nodes which will not need any kind of predictive behaviour, it is of no use.