**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 |

1. Compute the Gini index for the overall collection of training examples.

|  |
| --- |
| Class C0 has 10 counts  Class C1 has 10 counts  Total count = 20  Gini = 1 – [(10/20)^2 + (10/20)^2] = 1 – (0.25 + 0.25) = 0.5 |

1. Compute the Gini index for the Customer ID attribute.

|  |
| --- |
| First, computing Gini index for each customer id c,  Gini(c) = 1 – [(0/1)^2 + (1/1)^2] = 1 – 1 = 0  So we will have 0 as the Gini index for each customer id. So the weighted average of those will be,  Gini = (1/20) \* 0 + (1/20) \* 0 + …… + (1/20) \* 0 = 0 |

1. Compute the Gini index for the Gender attribute

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | Male | Female | Total | | Class C0 | 6 | 4 | 10 | | Class C1 | 4 | 6 | 10 | | Total | 10 | 10 | 20 |   Gini(female) = 1 – [(4/10)^2 + (6/10)^2] = 1 – (0.16 + 0.36) = 0.48  Gini(male) = 1 – [(6/10)^2 + (4/10)^2] = 1 – (0.36 + 0.16) = 0.48  Weighted sum of Gini Index: [(10/20) \* 0.48] + [(10/20) \* 0.48] = 0.48 |

1. Compute the Gini index for the Car Type attribute using multiway split.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Family | Sports | Luxury | Total | | Class C0 | 1 | 8 | 1 | 10 | | Class C1 | 3 | 0 | 7 | 10 | | Total | 4 | 8 | 8 | 20 |   Gini(family) = 1 – [(1/4)^2 + (3/4)^2] = 0.375  Gini(sports) = 1 – [(8/8)^2 + (0/8)^2] = 0  Gini(luxury) = 1 – [(1/8)^2 + (7/8)^2] = 0.21875  Weighted sum of Gini Index: [(4/20)\* 0.375] + [(8/20) \* 0] + [(8/20) \* 0.21875] = 0.1625 |

1. Compute the Gini index for the Shirt Size attribute using multiway split.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | Small | Medium | Large | Extra Large | Total | | Class C0 | 3 | 3 | 2 | 2 | 10 | | Class C1 | 2 | 4 | 2 | 2 | 10 | | Total | 5 | 7 | 4 | 4 | 20 | | Gini Index | 0.48 | 0.49 | 0.5 | 0.5 | - |   Weighted sum of Gini Index: [(5/20) \* 0.48] + [(7/20) \* 0.49] + [(4/20) \* 0.5] + [(4/20) \* 0.5] = 0.4915 |

1. Which attribute is better, Gender, Car Type, or Shirt Size?

|  |
| --- |
| Since Gini index indicates the degree of impurity, so the attribute that has smallest Gini index is the better attribute. Hence, the answer will be Car Type (Gini index: 0.1625). |

1. Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini

|  |
| --- |
| Customer ID should not be used as the attribute test condition because each attribute is unique. |