UNIVERSITE DE SOUSSE

Institut Supérieur d'Informatique

et des Techniques de Communication
Hammam Sousse

جامعة سوسة

المعهد العالي للإعلامية وتقنيات الاتصال
بحمام سوسة

Compte Rendu

DATA Mining

Réalisé par :

Nada Boudhina

3DNI1

Année Universitaire : 2021-2022

Exercice

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

Partie 1 :

A/Gini index for the overall collecion of trainning examples

The Gini index for the overall collection is :

$$1 - (5/10)\ 2\ -(5/10)2 = 0.5.$$

B/Gini index for the Customer ID attribute

The Gini index for the Customer ID is 0. Because we dont have any data repeated ; any new customer will have new ID .

C/the Gini index for the Gender attribute

C0 : $1-(6/10)^2-(4/10)^2=0.48$

C1 : $1-(6/10)^2-(4/10)^2=0.48$

The Gini index is s $0.5 \times 0.48 + 0.5 \times 0.48 = 0.48$.

D/ The Gini index for the car Type attribute using multiway split

The gini for Family is $1-(1/4)^2 -(3/4)^2 =0.375$

The gini for sports is $1-(8/8)^2 -(0/8)^2 =0$

The gini for luxury is $1-(1/8)^2 -(7/8)^2 =0.2188$

The Gini index is $(4/20)*0.375+(8/20)*0.2188=0.1615$

E/The Gini index for the shirt size attribute using multiway split

The gini for Small shirt size is $1-(3/5)^2 -(2/5)^2 =0.48$

The gini for Meduim shirt size is $1-(3/7)^2 -(4/7)^2 =0.48$

The gini for Large shirt size is $1-(2/4)^2 -(2/4)^2 =0.5$

The gini for Extra Large shirt size is $1-(2/4)^2 -(2/4)^2 =0.5$

The Gini index is

$(3+2)/20*0.48+(3+4)/20*0.48+(2+2)/20*0.5+(2+2)/20*0.5=0.4914$.

F /Which attribute is better ?

The car type attribute is better because it has the lowest Gini index.

G / Explain why customer ID should not be used as the attribute test condition even though it has the lowest Gini

The attribute cannot be used for prediction because new customers are assigned to new Customer IDs.

Partie 2 :

A / The entropy of this collection of trainning examples with respect to the class attribute

The entropy of the training examples is $-4/9 \log2(4/9) - 5/9 \log2(5/9) = 0.9911$.

B/The information gains of a1 and a2 relative to these training examples

The entropy for a1 is

$(4/9)[-(3/4)\log^2(3/4)-(1/4)\log^2(1/4)]$

+

$(5/9)[-(1/5)\log^2(1/5)-(4/5)\log^2(4/5]=0.7616$

The information gain for a1 is 0.9911-0.7616=0.2294

The entropy for a2 is

$(5/9)[-(2/5)\log^2(2/5)-(3/5)\log^2(3/5)]$

+

$(2/9)[-(2/4)\log^2(2/4)-(2/4)\log^2(2/4]=0.9839$

The information gain for a2 is 0.9911-0.9839=0.0072

C/For a3,which is a continous attribute

a 3	Class label	Split point	Entropy	Info gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

The best split for a3 occurs at split point equals to 2.

D/the best split (among a1,a2,a3) according to the information gain

a1

E/ The best split (between a1 and a2) according to the misclassification error rate

The error rate for a1 is 2/9 and that for a2 is 4/9 so that a1 is the best split attribute.

F/ The best split (between a1 and a2) according to the Gini index

For attribute a1 ,the gini index is

$(4/9)[1-(3/4)^2-(1/4)^2] + (5/9)[1-(1/5)^2 - (4/5)^2]=0.3444$

For attribute a2 ,the gini index is

$(5/9)[1-(2/5)^2-(3/5)^2] + (2/9)[-(2/4)^2-(2/4)^2]=0.4889$

Since the gini index for a1 is smaller, it produces the better split