Exercice :

1. Consider the training examples shown in Table 3.5 for a binary classification problem.

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

Gini_index = 1- $\sum_{i=0}^{c-1} P_i(t)^2$

= 1 - $P(C_0)^2 - P(C_1)^2$

= 1- 1/4 -1/4

=1/2

b. Compute the Gini index for Gender the attribute.

	M	F
C0	6	4
C1	4	6

Gini_index(M) = 1 - $P(C_0)^2 - P(C_1)^2$
= 1 – (6/10)² - (4/10)²
= 0.28

Gini_index(F) = 1 - $P(C_0)^2 - P(C_1)^2$
= 1 – (4/10)² - (6/10)²
= 0.28

Gini_index = 10/20 * Gini_index(M) + 10/20 * Gini_index(F)
=0.48

c. Compute the Gini index for Car_Type the attribute using multiway split.

	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
	4	8	8

Gini_index(Family) = 1 - $P(C_0)^2 - P(C_1)^2$
= 1 – (1/4)² - (3/4)²
= 0.375

Gini_index(Sport) = 0

Gini_index(Luxury) = 1 – (1/8)² - (7/8)²
= 0.218

Gini_index = 0.075 + 0.0872
= 0.1622

d. Compute the Gini index for the attribute using multiway

	Small	Meduim	Large	Extra_Large
C0	3	3	2	2
C1	2	4	2	2
	5	7	4	4

$$Gini\_index(Small) = 1 - (3/5)^2 - (2/5)^2$$
$$= 0.48$$

$$Gini\_index(Meduim) = 1 - (3/7)^2 - (4/7)^2$$
$$= 0.489$$

$$Gini\_index(Large) = 1 - (2/4)^2 - (2/4)^2$$
$$= 0.5$$

$$Gini\_index(Extra\_Large) = 1 - (2/4)^2 - (2/4)^2$$
$$= 0.5$$

$$Gini\_index = 5/20*0.48 + 7/20 * 0.489 + 4/20* 0.5 + 4/20 *0.5$$
$$= 0.491$$

e. the better attribute is Car_Type.

f. The Custumer_Id should not be used as the attribute test condition even through it has the lowest Gini , because it doesn't represent any new information , there is not a changeable numbers