2. a.

C_0	C_1
10	10

$GINI = 1 - \sum_{i=1}^{n} p_i^2 = 1 - p_1^2 - p_2^2 - \cdots - p_n^2$

$GINI = 1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = 0.5$

b. Customer ID

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:1	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0	C_0:0
C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:0	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1	C_1:1

$GINI = \frac{1}{20}(1 - (\frac{1}{1})^2 - (\frac{0}{1})^2) \cdot 10 + \frac{1}{20}(1 - (\frac{0}{1})^2 - (\frac{1}{1})^2) \cdot 10 = 0$

c. Gender

M:	C_0: 6	F:	C_0: 4
	C_1: 4		C_1: 6

$GINI = \frac{10}{20}(1 - (\frac{6}{10})^2 - (\frac{4}{10})^2) + \frac{10}{20}(1 - (\frac{4}{10})^2 - (\frac{6}{10})^2) = 0.48$

d. Car Type

Family:	C_0: 1	Sports:	C_0: 8	Luxury:	C_0: 1
	C_1: 3		C_1: 0		C_1: 7

$GINI = \frac{4}{20}(1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) + \frac{8}{20}(1 - (\frac{8}{8})^2 - (\frac{0}{8})^2) + \frac{8}{20}(1 - (\frac{1}{8})^2 - (\frac{7}{8})^2)$

$= 0.1625$

e. Shirt Size

S:	C_0: 3	M:	C_0: 3	L:	C_0: 2	XL:	C_0: 2
	C_1: 2		C_1: 4		C_1: 2		C_1: 2

$GINI = \frac{5}{20}(1 - (\frac{3}{5})^2 - (\frac{2}{5})^2) + \frac{7}{20}(1 - (\frac{3}{7})^2 - (\frac{4}{7})^2) + \frac{4}{20}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) + \frac{4}{20}(1 - (\frac{2}{4})^2 - (\frac{2}{4})^2)$

$\simeq 0.4914\overline{3}$

f. Car Type. Because the gini index of Car Type split is the smallest, which means the purest class.

g. There is no repitition in the Customer ID, and every ID is unique. Thus, we cannot get any useful information from splitting by Customer ID.

3. a. +: 4

-: 5

$-(\frac{4}{9})\cdot\log_2(\frac{4}{9}) - (\frac{5}{9})\log_2(\frac{5}{9}) \approx 0.9911$

$\text{Entropy} = \sum_{i=1}^{n} -p_i \log_2(p_i) = -p_1\log_2(p_1) - \cdots - p_n\log_2(p_n)$

b.

	a_1			a_2	
	+	−		+	−
T:	3	1	T:	2	3
F:	1	4	F:	2	2

c.

a_3	+	−
1	1	0
6	1	0
5	0	2
4	1	0
7	1	1
3	0	1
8	0	1

5.

	A: +	−	B: +	−
T	4	3	3	1
F	0	3	1	5

+: 4

−: 6

level 1 splitting

7. X:

	C_1	C_2
0	60	60
1	40	40

a) error rate $\frac{60+40}{200}$

Y:

	C_1	C_2
0	40	60
1	60	40

$\frac{40+40}{200}$

Z:

	C_1	C_2
0	30	70
1	70	30

$\frac{30+30}{200}$ lowest error rate

level 2 splitting on Z

Z=0:

X:

	C_1	C_2
0	15	45
1	15	25

Y:

	C_1	C_2
0	15	45
1	15	25

$Z=1$:

X:	C_1	C_2
0	45	15
1	25	15

Y:	C_1	C_2
0	25	15
1	45	15

b) $X=0$:

Y:	C_1	C_2
0	5	55
1	55	5

Z:	C_1	C_2
0	15	45
1	45	15

$X=1$

Y:	C_1	C_2
0	35	5
1	5	35

Z	C_1	C_2
0	15	25
1	25	15