Theoretical Assignment 01

Name – **Susanket Sarkar**
Roll number : **20AE10041**

Question 01

The data given in the question:

X1	X2	X3		#(Y=1)	#(Y=2)	#(Y=3)
1	1	A		15	0	0
1	2	A		15	0	0
2	2	A		2	9	1
2	1	A		3	5	0
1	1	B		0	10	4
1	2	B		0	10	1
2	2	B		8	2	4
2	1	B		7	3	1
1	1	C		0	6	0
1	2	C		0	9	0
2	2	C		1	0	14
2	1	C		0	0	20
1	1	D		0	2	15
1	2	D		1	3	14
2	2	D		1	0	9
2	1	D		0	0	5

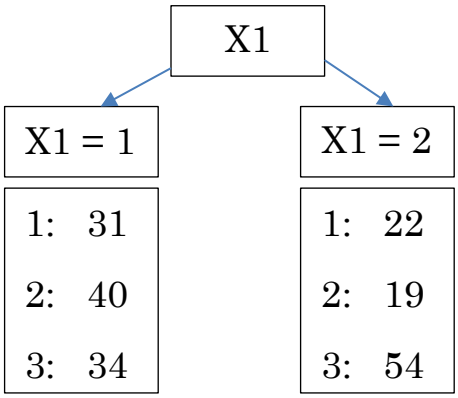
We calculate the information gain of each of the features:

For **X1**:

Weighted Entropy =

$$\frac{105}{200} \times \left(-\frac{40}{105} log_2\left(\frac{40}{105}\right) - \frac{65}{105} log_2\left(\frac{65}{105}\right)\right)$$
$$+\frac{95}{200} \times \left(-\frac{54}{95} log_2\left(\frac{54}{95}\right) - \frac{41}{95} log_2\left(\frac{41}{95}\right)\right)$$
$$= \frac{105}{200} \times 0.503 + \frac{95}{200} \times 0.469$$
$$= 0.972$$

Information Gain = 1 – 0.972 = **0.028**

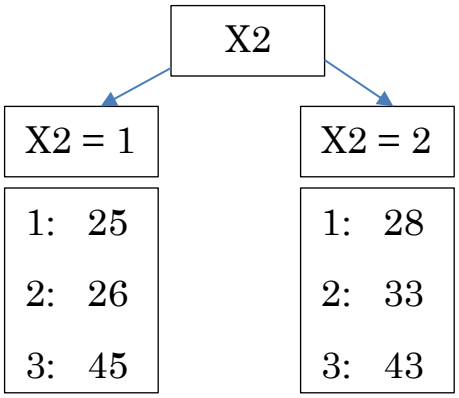


For X2:

Weighted Entropy =

$$\frac{96}{200}\times\left(-\frac{45}{96}log_2\left(\frac{45}{96}\right)-\frac{51}{96}log_2\left(\frac{51}{96}\right)\right)$$
$$+\frac{104}{200}\times\left(-\frac{43}{104}log_2\left(\frac{43}{104}\right)-\frac{61}{104}log_2\left(\frac{61}{104}\right)\right)$$
$$= 0.479 + 0.508$$
$$= 0.987$$

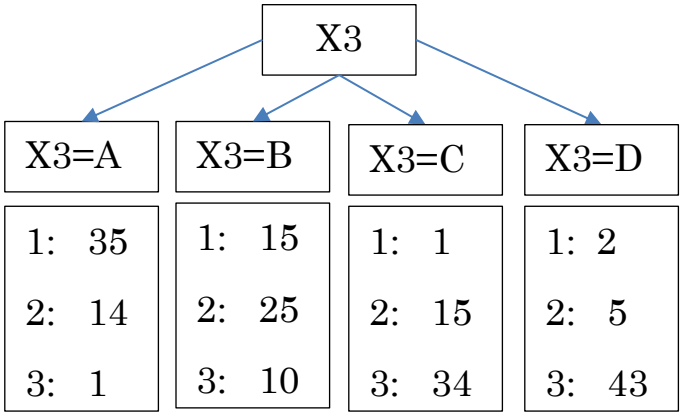
Information Gain = 1 – 0.987 = **0.013**



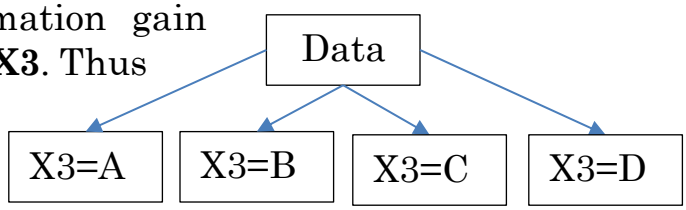
Similarly for **X3** we find:

Weighted Entropy = 0.8426

Information Gain = 1 – 0.8426 = **0.1574**



As we find out the Maximum Information gain among the three features (X1, X2, X3) is **X3**. Thus the **first split of the decision tree will be with the X3 feature**. The leaf nodes after the first split looks like:

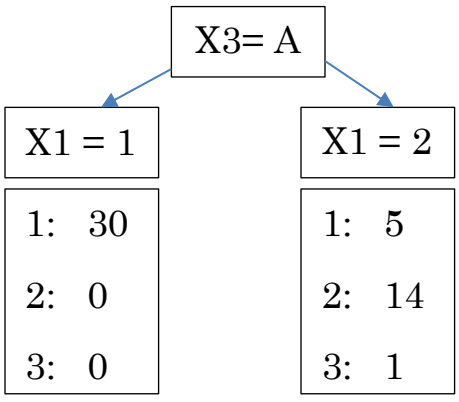


After the first split we have to find the feature for the second split. So we calculate the information gain in both the features X1 and X2.

We calculate the info. Gain for the split (X3 = A and X1 = 1) and (X3 = A and X1 = 2) :

Weighted Entropy for (X3 = A) =

$$\frac{30}{50}\times\left(-\frac{30}{30}log_2(1)-0\,log_2(0)\right)$$
$$+\frac{20}{50}\times\left(-\frac{14}{20}log_2\left(\frac{14}{20}\right)-\frac{6}{20}log_2\left(\frac{6}{20}\right)\right)$$
$$= \mathbf{0.3524}$$



Similarly we find the weighted entropy for (X3 = B,C,D) :

Weighted Entropy for (X3 = B) = **0.846**

Weighted Entropy for (X3 = C) = **0.1295**

Weighted Entropy for (X3 = D) = **0.5692**

Info. gain = $1 - 0.25 \times (0.3524 + 0.846 + 0.1295 + 0.5692) =$ **0.5257**

We calculate the info. Gain for the split (X3 = A and X2 = 1) and (X3 = A and X2 = 2) :

Weighted Entropy for (X3 = A) = 0.919

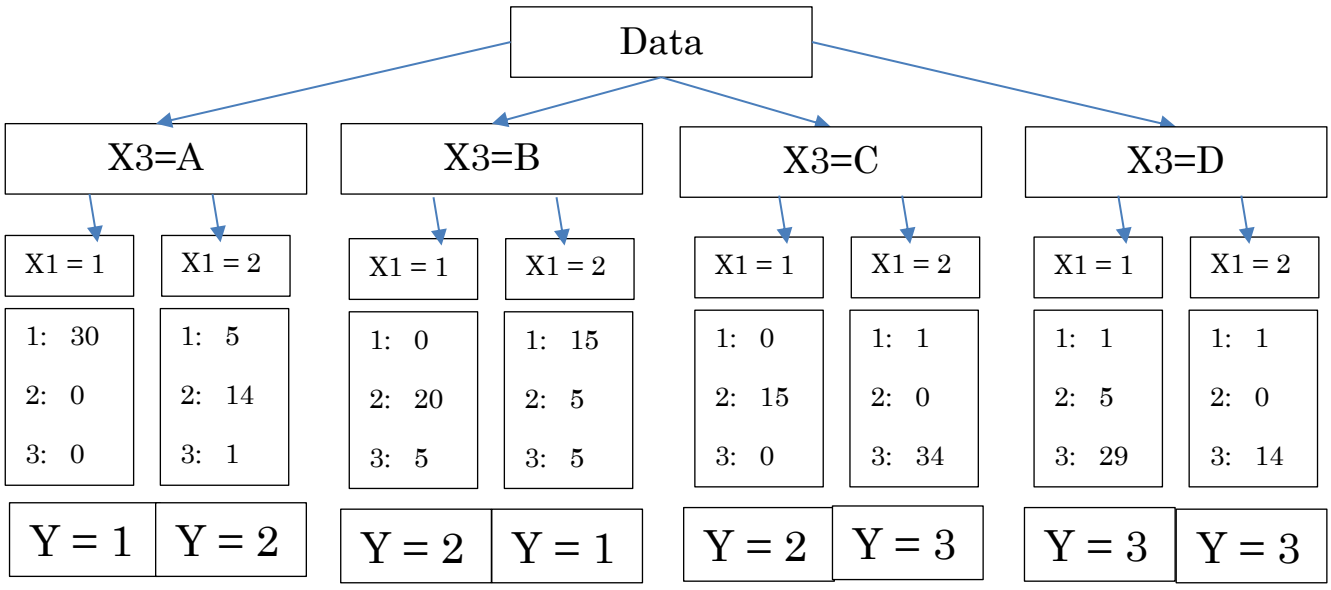
Weighted Entropy for (X3 = B) = 0.998

Weighted Entropy for (X3 = C) = 0.874

Weighted Entropy for (X3 = D) = 0.576

Info. gain = $1 - 0.25 \times (0.919 + 0.998 + 0.874 + 0.576) =$ **0.1583**

Clearly the information gain is higher for split with feature **X1**.After splitting the Decision tree looks like this.



The predictions for the table of data are as follows:

X1	X2	X3	$\hat{Y}$	Y = 1	Y = 2	Y = 3	Points correctly classified	Points misclassified
1	1	A	1	15	0	0	15	0
1	2	A	1	15	0	0	15	0
2	1	A	2	2	9	1	9	3
2	2	A	2	3	5	0	5	3
1	1	B	2	0	10	4	10	4
1	2	B	2	0	10	1	10	1
2	1	B	1	8	2	4	8	6
2	2	B	1	7	3	1	7	4
1	1	C	2	0	6	0	6	0
1	2	C	2	0	9	0	9	0
2	1	C	3	1	0	14	14	1
2	2	C	3	0	0	20	20	0
1	1	D	3	0	2	15	15	2
1	2	D	3	1	3	14	14	4
2	1	D	3	1	0	9	9	1
2	2	D	3	0	0	5	5	0

Looking at the table we find that :

Total points correclty classifed: **171**

Total points incorreclty classifed: 29

Accuracy: $\frac{171}{200}$ = **0.855** (85.5%)

Question 02

From the table given in the question we get the probability distribution as,

	Y = 1	Y = 2	Y = 3
X1 = 1	31/43	34/45	34/82
X1 = 2	12/43	11/45	48/82
X2 = 1	15/43	12/45	39/82
X2 = 2	28/43	33/45	43/82
X3 = A	32/43	9/45	1/82
X3 = B	8/43	22/45	9/82
X3 = C	1/43	9/45	34/82
X3 = D	2/43	5/45	38/82

According to the above table the prior distribution is :

$$P(Y = 1) = \frac{43}{43 + 45 + 82} = 0.253$$
$$P(Y = 2) = \frac{45}{43 + 45 + 82} = 0.265$$
$$P(Y = 3) = \frac{82}{43 + 45 + 82} = 0.482$$

We use this probability distribution to find out the labels of the points:

a. X1 = 2, X2 = 1, X3 = A
b. X1 = 2, X2 = 1, X3 = B
c. X1 = 1, X2 = 1, X3 = C
d. X1 = 2, X2 = 1, X3 = D

Also we know that for Naïve Bayes Classifier,
P(Y=y | X1, X2, X3) = K*P(Y=y)*P(X1|Y=y)* P(X2|Y=y)* P(X3 |Y=y)

For point X1 = 2, X2 = 1, X3 = A:

P(Y=1 | X1 = 2, X2 = 1, X3 = A) $= K_1 \times 0.253 \times \frac{31}{43} \times \frac{15}{43} \times \frac{32}{43}$

$= K_1 (0.0473)$

P(Y=2 | X1 = 2, X2 = 1, X3 = A) $= K_1 \times 0.265 \times \frac{11}{45} \times \frac{12}{45} \times \frac{9}{45}$

$= K_1 (0.0035)$

P(Y=3| X1 = 2, X2 = 1, X3 = A) $= K_1 \times 0.482 \times \frac{48}{82} \times \frac{39}{82} \times \frac{1}{82}$

$= K_1 (0.00163)$

We know that,

P(Y=1|X1,X2,X3) + P(Y=2|X1,X2,X3) + P(Y=3|X1,X2,X3) = 1

Thus , K_1 * (0.0473 + 0.0035 + 0.00163) = 1

$K_1 = 19.07$

Thus for point X1 = 2, X2 = 1, X3 = A:

Prediction: **Y = 1**

Confidence = 19.07*0.0473 = **0.902**

For point X1 = 2, X2 = 1, X3 = B:

P(Y=1 | X1 = 2, X2 = 1, X3 = B) $= K_2 \times 0.253 \times \frac{12}{43} \times \frac{15}{43} \times \frac{8}{43}$

$= K_2 (0.00458)$

P(Y=2 | X1 = 2, X2 = 1, X3 = B) $= K_2 \times 0.265 \times \frac{11}{45} \times \frac{12}{45} \times \frac{22}{45}$

$= K_2 (0.00845)$

P(Y=3| X1 = 2, X2 = 1, X3 = B) $= K_2 \times 0.482 \times \frac{48}{82} \times \frac{39}{82} \times \frac{9}{82}$

$= K_2 (0.01473)$

We know that,

P(Y=1|X1,X2,X3) + P(Y=2|X1,X2,X3) + P(Y=3|X1,X2,X3) = 1

Thus , K_2 = (0.00458+0.00845+0.01473) = 1

$K_2 = 36.02$

Thus for point X1 = 2, X2 = 1, X3 = B:

Prediction: **Y = 3**

Confidence = 36.02*0.01473 = **0.5036**

For point X1 = 1, X2 = 1, X3 = C:

$P(Y=1 \mid X1 = 2, X2 = 1, X3 = C) = K3 \times 0.253 \times \frac{31}{43} \times \frac{15}{43} \times \frac{1}{43}$

$= K_3 (0.00147)$

$P(Y=2 \mid X1 = 2, X2 = 1, X3 = C) = K3 \times 0.265 \times \frac{34}{45} \times \frac{12}{45} \times \frac{9}{45}$

$= K_3 (0.00947)$

$P(Y=3 \mid X1 = 2, X2 = 1, X3 = C) = K3 \times 0.482 \times \frac{34}{82} \times \frac{39}{82} \times \frac{34}{82}$

$= K_3 (0.03941)$

We know that,

$P(Y=1|X1,X2,X3) + P(Y=2|X1,X2,X3) + P(Y=3|X1,X2,X3) = 1$

Thus , $K_3 = (0.00147+0.00947+0.03941) = 1$

$K_3 = 19.86$

Thus for point X1 = 1, X2 = 1, X3 = C:

Prediction: **Y = 3**

Confidence = 19.86*0.03941 = **0.7827**

For point X1 = 2, X2 = 1, X3 = B:

$P(Y=1 \mid X1 = 2, X2 = 1, X3 = D) = K4 \times 0.253 \times \frac{12}{43} \times \frac{15}{43} \times \frac{2}{43}$

$= K_4 (0.001145)$

$P(Y=2 \mid X1 = 2, X2 = 1, X3 = D) = K4 \times 0.265 \times \frac{11}{45} \times \frac{12}{45} \times \frac{5}{45}$

$= K_4 (0.001919)$

$P(Y=3 \mid X1 = 2, X2 = 1, X3 = D) = K4 \times 0.482 \times \frac{48}{82} \times \frac{39}{82} \times \frac{38}{82}$

$= K_4 (0.062186)$

We know that,

$P(Y=1|X1,X2,X3) + P(Y=2|X1,X2,X3) + P(Y=3|X1,X2,X3) = 1$

Thus , $K_4 = (0.001145+0.001919+0.062186) = 1$

$K_4 = 15.32$

Thus for point X1 = 2, X2 = 1, X3 = B:

Prediction: **Y = 3**

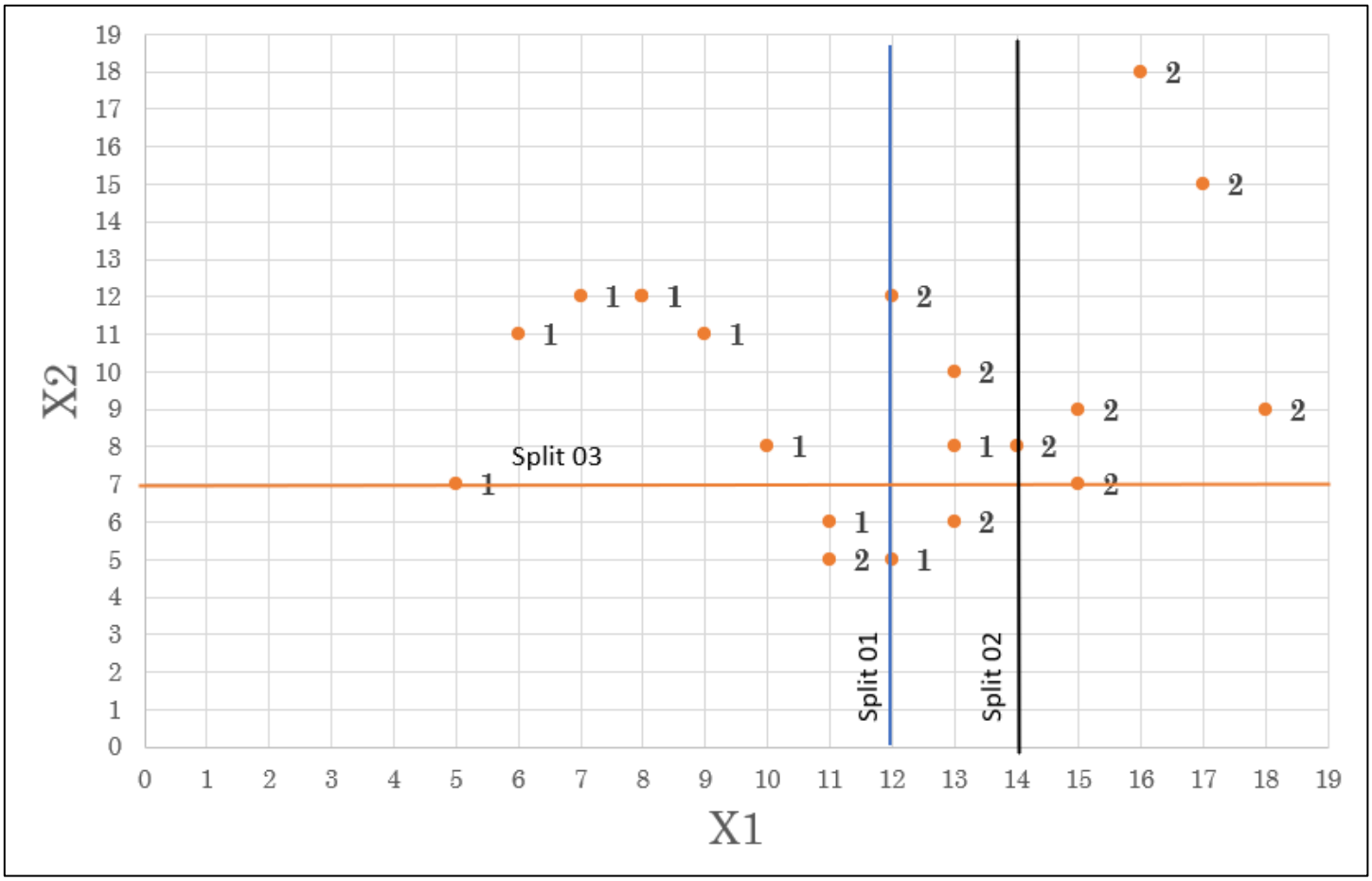
Confidence = 15.32*0.062186 = **0.9527**

Question 03:

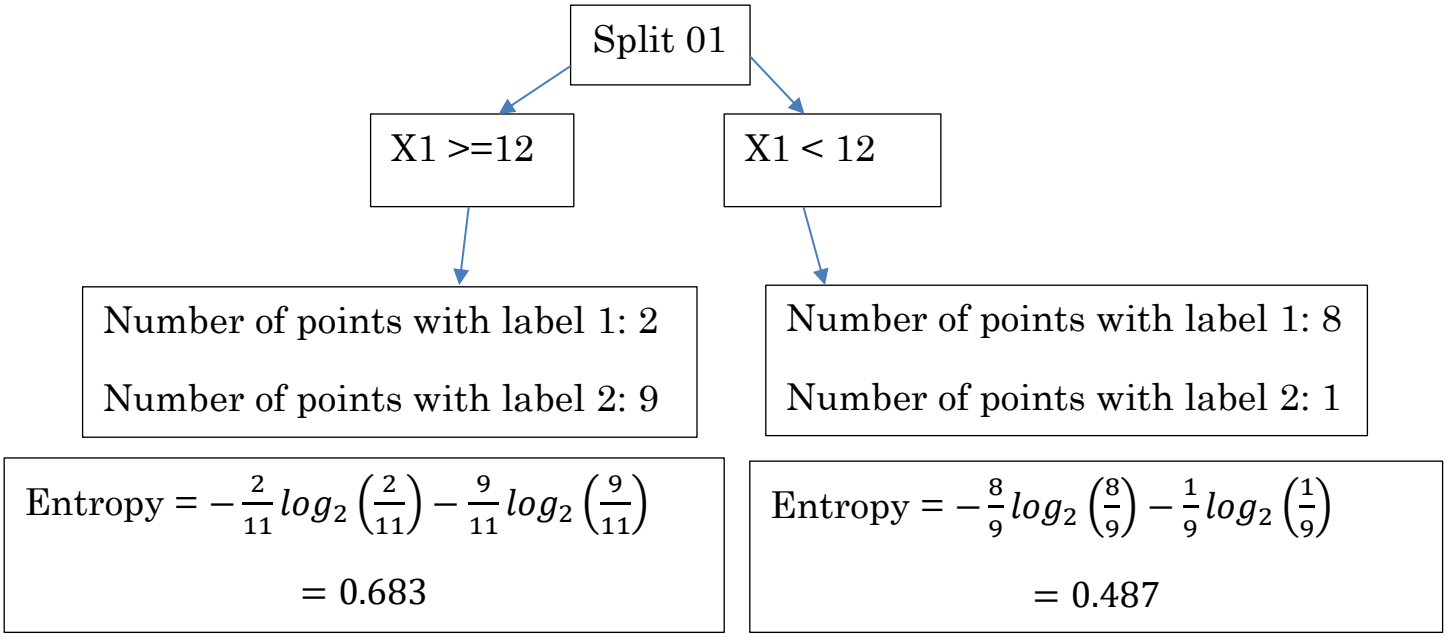
Given the data,

	X1	X2	Y		X1	X2	Y
1	5	7	1	**11**	13	6	2
2	7	12	1	**12**	14	8	2
3	12	5	1	**13**	17	15	2
4	10	8	1	**14**	15	9	2
5	6	11	1	**15**	13	10	2
6	13	8	1	**16**	11	5	2
7	8	12	1	**17**	16	18	2
8	9	11	1	**18**	15	7	2
9	11	6	1	**19**	12	12	2
10	8	12	1	**20**	18	9	2

Plotting the datapoints we get:

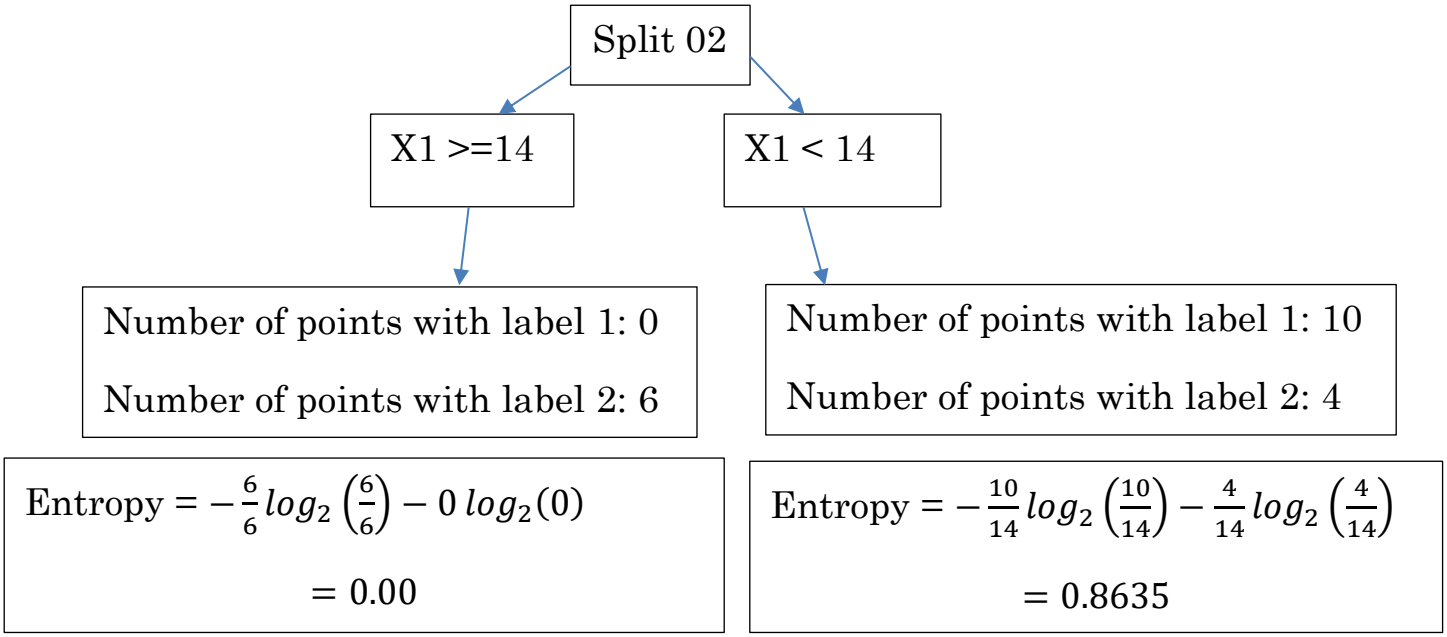


For Split 01,



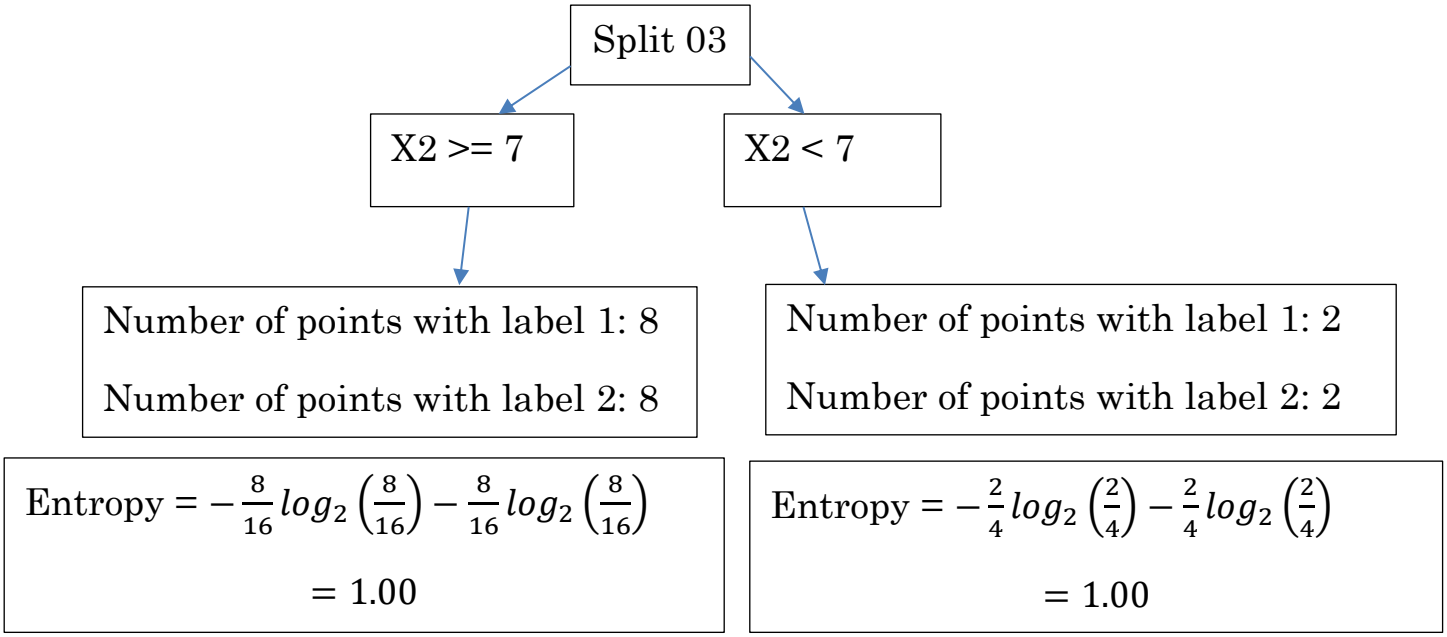
Weighted Entropy for Split 01: $\frac{11}{20}\times 0.683+\frac{9}{20}\times 0.487 = 0.5948$
Information Gain for the split : $1 - 0.6274 =$ **0.4052.**

For Split 02,



Weighted Entropy for Split 02: $\frac{6}{20}\times 0.00+\frac{14}{20}\times 0.8635 = 0.6044$
Information Gain for the split : $1 - 0.6274 =$ **0.3955.**

For Split 03,



Weighted Entropy for Split 03: $\frac{8}{16}\times 1+\frac{2}{4}\times 1= 1$
Information Gain for the split : 1 – 1 = **0.**

Comparing the values of Information gain for the three split we see that Split 01 has the lowest Entropy and Higher Info. Gain. **Hence the most desirable split is Split 01 (X1>=12 and X1<12).**

Question 04

We know that the likelihood of a feature is:

$$P(X_i|Y) = \frac{1}{\sigma\sqrt{2x_i}} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

Where,
$\bar{x}$ is the mean of x
σ is the variance of x.

We know that for a Naïve Bayes Classification,

$$p(Y = y|X_1, X_2) = K_i \times P(Y = y) \times P(X_1|Y = y) \times P(X_2|Y = y)$$

For the data which has a label of Y = 1,
$\bar{x}_1$ = 8.9 and $\sigma(x_1)$ = 2.601
$\bar{x}_2$ = 9.2 and $\sigma(x_2)$ = 2.699
And the value of P(Y=1) = 0.5.

Thus the value of $p(Y = 1|X_1, X_2)$ is:

ID	X1	X2	Y	$P(X_1\|Y = 1)$	$P(X_2\|Y = 1)$	$p(Y = 1\|X_1, X_2)$
1	5	7	1	0.050	0.106	K1*0.003
2	7	12	1	0.117	0.086	K2*0.005
3	12	5	1	0.075	0.044	K3*0.002
4	10	8	1	0.140	0.134	K4*0.009
5	6	11	1	0.082	0.118	K5*0.005
6	13	8	1	0.044	0.134	K6*0.003
7	8	12	1	0.144	0.086	K7*0.006
8	9	11	1	0.153	0.118	K8*0.009
9	11	6	1	0.111	0.073	K9*0.004
10	8	12	1	0.144	0.086	K10*0.006
11	13	6	2	0.044	0.073	K11*0.002
12	14	8	2	0.022	0.134	K12*0.002
13	17	15	2	0.001	0.015	K13*0.000
14	15	9	2	0.010	0.147	K14*0.001
15	13	10	2	0.044	0.141	K15*0.003
16	11	5	2	0.111	0.044	K16*0.002
17	16	18	2	0.004	0.001	K17*0.000
18	15	7	2	0.010	0.106	K18*0.001
19	12	12	2	0.075	0.086	K19*0.003
20	18	9	2	0.000	0.147	K20*0.000

For the data which has a label of Y = 2,
$\bar{x}_1$ = 14.4 and $\sigma(x_1)$ = 2.221,
$\bar{x}_2$ = 9.9 and $\sigma(x_2)$ = 4.067

And the value of P(Y=2) = 0.5,

Thus the value of $p(Y = 2|X_1, X_2)$ is:

ID	X1	X2	Y	$P(X_1\|Y = 2)$	$P(X_1\|Y = 2)$	$P(Y = 2\|X_1, X_2)$
1	5	7	1	0.000	0.076	K1*0.0000009
2	7	12	1	0.001	0.086	K2*0.0000300
3	12	5	1	0.100	0.047	K3*0.0023781
4	10	8	1	0.025	0.088	K4*0.0011101
5	6	11	1	0.000	0.095	K5* 0.0000067
6	13	8	1	0.147	0.088	K6*0.0064750
7	8	12	1	0.003	0.086	K7* 0.0001214
8	9	11	1	0.009	0.095	K8*0.0004421
9	11	6	1	0.056	0.062	K9*0.0017236
10	8	12	1	0.003	0.086	K10*0.0001214
11	13	6	2	0.147	0.062	K11*0.0045603
12	14	8	2	0.177	0.088	K12*0.0077709
13	17	15	2	0.091	0.045	K13*0.0020228
14	15	9	2	0.173	0.096	K14*0.0082874
15	13	10	2	0.147	0.098	K15*0.0072192
16	11	5	2	0.056	0.047	K16*0.0013211
17	16	18	2	0.139	0.014	K17*0.0009356
18	15	7	2	0.173	0.076	K18*0.0065867
19	12	12	2	0.100	0.086	K19*0.0043001
20	18	9	2	0.048	0.096	K20*0.0023111

We also know that,
P(Y=1 | X1,X2) + P(Y=2 | X1,X2) = 1.

Thus, $K_i \times \left(\left(P(Y = 1) * P(X_1|Y) * P(X_2|Y)\right) + \left(P(Y = 2) * P(X_1|Y) * P(X_2|Y)\right)\right) = 1$

Thus we obtain the values of K_i and value of $P(Y = k|X_1, X_2)$.

i	K_i	P(Y=1|X1,X2)	P(Y=2|X1,X2)	$Confidence$
1	189.17	0.9997	0.0003	0.9997
2	98.07	0.9941	0.0059	0.9941
3	123.79	0.4112	0.5888	0.5888
4	47.63	0.8943	0.1057	0.8943
5	102.45	0.9986	0.0014	0.9986
6	52.97	0.3140	0.6860	0.6860
7	78.68	0.9809	0.0191	0.9809
8	52.59	0.9535	0.0465	0.9535
9	86.57	0.7016	0.2984	0.7016
10	78.68	0.9809	0.0191	0.9809
11	80.89	0.2622	0.7378	0.7378
12	53.92	0.1620	0.8380	0.8380
13	246.10	0.0044	0.9956	0.9956
14	55.49	0.0802	0.9198	0.9198
15	48.31	0.3025	0.6975	0.6975
16	132.97	0.6487	0.3513	0.6487
17	533.67	0.0014	0.9986	0.9986
18	70.36	0.0732	0.9268	0.9268
19	66.20	0.4307	0.5693	0.5693
20	214.05	0.0106	0.9894	0.9894

From the calculation of the confidence we observe that the least value of confidence is for **Point 19** and the confidence is **0.5693**.

.

Question 05

(i) We are given N datapoints of the form (x_i, y_i, w_i).

For the question the Loss Function is defined as:
$\boldsymbol{\mathcal{L}}(Loss) = \sum_{i=1}^{N} w_i(y_i - x_i a - b)^2 = W\|Y - A^T X\|_2^2$ ---------- Eq(1)

where $\boldsymbol{A} = [a_1, a_2 \dots a_N, b]$, $\widehat{\boldsymbol{Y}} = [\widehat{y_1}, \cdots \hat{y}_N, 1]$ and $\boldsymbol{W} = [w_1, \cdots w_n, 1]$

Where a and b are defined as follows:

$$\hat{y}_i = ax_i + b = A^T \cdot X$$

We have to optimize the loss function (**objective function**) as follows:

$$arg\,min\left[W\left\|Y - \hat{Y}\right\|_2^2\right]$$

To minimize the loss function we find $\frac{\partial \mathcal{L}}{\partial A} = 0$.
But from the equation Eq(1) we get :

$$\frac{\partial \mathcal{L}}{\partial A} = -2WX^T(Y - A^T X)$$

$$= -2WX^T Y + 2WX^T XA$$

Equating this expression to 0,

$$WX^T Y = W(X^T X)A$$

$$X^T Y = W^{-1}W(X^T X)A$$

$$X^T Y = (X^T X)A$$

$$A = (X^T X)^{-1}X^T Y$$

Thus the solution of the objective function is not dependent on the weight vector (W).

(ii) We are given N datapoints of the form (x_i, y_i)

For the question the loss function is defined as:

$$\mathcal{L}(Loss) = W\|Y - A^T X\|_2^2 + \lambda\|A - V\|_2^2 \text{ ----------Eq(2)}$$

where, $V = [v_1, v_2 \ldots v_N, 1]$ is a N+1 dimensional known vector and $\|A - V\|_2^2$ represents an L2 norm or the Euclidian distance between the points in vectors **A** and **V.** $\boldsymbol{\lambda}$ is a constant real number.

We have to optimize the loss function (**objective function**) as follows:

$$arg\min\left[\left[W\left\|Y - \hat{Y}\right\|_2^2\right] + \lambda\|A - V\|_2^2\right]$$

To minimize the loss function we find $\frac{\partial \mathcal{L}}{\partial A} = 0$.
But from the equation Eq(2) we get

$$\frac{\partial \mathcal{L}}{\partial A} = -2X^T(Y - X^T A) + 2\lambda(A - V)$$

$$= -2X^T Y + 2X^T XA + 2\lambda(A - V)$$

Equating this to zero,

$$X^T Y = (X^T X + \lambda I)A - \lambda V$$

where, I is the Identity matrix.

$$X^T Y + \lambda V = (X^T X + \lambda I)A$$

$$A = (X^T X + \lambda I)^{-1}(X^T Y + \lambda V)$$

Thus the solution to the objective equation is :

$$\boldsymbol{A = (X^T X + \lambda I)^{-1}(X^T Y + \lambda V)}$$