Name - Omkar Pandit
UT EID - oap338
Subject - Introduction to Machine Learning.

Q 1) Decision Tree

a) Formula for Entropy $= -\sum_{i=1}^{n} P_i \log_2 P_i$

In our case, result predictor = Satisfied / not satisfied.

So, calculating probabilities of indivisual categories,

$P(\text{satisfied}) = P_1 = \frac{5}{8} = 0.625$

$P(\text{not satisfied}) = P_2 = \frac{3}{8} = 0.375$

$$\therefore \text{Entropy} = -\left((P_1 \log_2 P_1) + (P_2 \log_2 P_2)\right)$$
$$= -\left(\left(\frac{5}{8}\log_2\frac{5}{8}\right) + \left(\frac{3}{8}\log_2\frac{3}{8}\right)\right)$$
$$= -(-0.4237 - 0.5306.)$$
$$= \boxed{0.9543}$$

Now, calculating Entropy with split feature:

So, on clean & satisfied? together,

Satisfied = Yes = 5

Satisfied = No = 3

For satisfied = Yes,
clean = yes = 2
= No = 3

For satisfied = No.
clean = yes = 2
= No = 1

∴ ratios are,
clean (Yes) = 2/5
clean (No) = 3/5

clean (Yes) = 2/3
clean (No) = 1/3.

$$\in(\text{clean}/\text{satisfied}) = P_1(\text{Yes}) \cdot \in(2,2) + P_2(\text{No}) \cdot \in(3,1)$$

$$= \frac{4}{8}\left[-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right]$$

$$+ \frac{4}{8}\left[-\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)\right]$$

$$= \frac{1}{2}[1 + 0.81]$$

$$= \boxed{0.905}$$

∴ Information Gain = $\in(\text{satisfied}) - \in(\text{clean}/\text{satisfied})$

= ~~0.950~~

= 0.9543 − 0.905

= 0.0493 = ~~0.045~~ $\boxed{0.045}$.

Q 1) II)

Rent ⇒ 400 , 500 , 600 , 600 , 700 , 800 , 1000 , 1200)

Y Y Y Y N Y N N

a b c d e f g.

a) Find the Entropy at mid-point of (400, 500)

⟶ $\boxed{a}$.

No = 0 Yes = 1 ⇒ yes = 4 & No = 3

$$\text{Entropy (a)} = -\left\{ \frac{1}{8}\left[\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right] + \frac{7}{8}\left[\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right] \right\}$$

$$= \left\{ \frac{1}{8}(0) + \frac{7}{8}\left(0.571\log_2(0.571) + 0.424\log_2(0.424)\right) \right\}$$

$$= -\left\{\frac{7}{8}(-0.461 - 0.524)\right\}$$

$$= -7/8 \times (-0.985)$$

$$= \underline{0.8618} = \boxed{0.862}$$

$$\text{Information gain (a)} = 0.95 - 0.861$$

$$= \underline{0.084}$$

b) Entropy at mid-point (500, 600) - $\boxed{b}$.

Yes = 2 No = 0 $\leftarrow b \rightarrow$ Yes = 3 No = 3

$$\text{Entropy}(b) = -\left\{\frac{2}{8}\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] + \frac{6}{8}\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\frac{3}{6}\right]\right\}$$

$$= -\left\{\frac{1}{4}[0+0] + \frac{3}{4}\left[\tfrac{1}{2}\log_2(\tfrac{1}{2}) + \tfrac{1}{2}\log_2(\tfrac{1}{2})\right]\right\}$$

$$= -\{0 + 3/4(-1)\}$$

$$= -\{-0.75\}$$

$$= 0.75.$$

$\therefore$ Information gain (b) = 0.95 - 0.75 = $\boxed{0.20}$.

c) Entropy at mid-point (600, 600).

Yes = 3, No = 0 $\leftarrow c \rightarrow$ yes = 2 No = 3.

$$\text{Entropy}(c) = -\left\{\frac{3}{8}\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] + \frac{5}{8}\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right]\right\}$$

$$= -\left\{\frac{5}{8}[-0.529 - 0.442]\right\}$$

$$= \frac{5}{8} \times (0.971)$$

$$= 0.6068 = \boxed{0.607}$$

$\therefore$ Information gain (c) = 0.95 - 0.607
= $\boxed{0.343}$.

d) Entropy at mid-point (600, 700). - $\boxed{d}$.

yes = 4, No = 0 ← d → yes = 1 No = 3.

$$Entropy = -\left\{ \frac{4}{8}\left[\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right]\right.\right.$$
$$\left.+ \frac{4}{8}\left[\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right]\right\}$$
$$= -\left\{\frac{4}{8}[0+0] + \frac{1}{2}[0.25(-2) + 0.75(-0.415)]\right\}$$
$$= -\left\{\frac{1}{2}[-0.5 - 0.311]\right\}$$
$$= -\tfrac{1}{2} \times [-0.811]$$
$$= \underline{0.4055}$$

∴ Information Gain (d) = 0.95 − 0.4055 = $\boxed{0.545}$.

e) Entropy at mid-point (700, 800) - $\boxed{e}$

yes = 4, No = 1 ← e → yes = 1, No = 2.

$$Entropy = -\left\{\frac{5}{8}\left[\frac{4}{5}\log_2\left(\frac{4}{5}\right) + \frac{1}{5}\left(\log_2\left(\frac{1}{5}\right)\right)\right]\right.$$
$$\left.+ \frac{3}{8}\left[\frac{1}{3}\log_2\left(\tfrac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]\right\}$$
$$= -\left\{\frac{5}{8}[(0.8)*(-0.322) + (0.2)(-2.322)] + 3/8[(0.33)(-1.599) + (0.67)(-0.577)]\right\}$$
$$= -\{5/8[-0.258 - 0.464] + 3/8[-0.528 - 0.38]\}$$
$$= -\{5/8(-0.722) + 3/8(-0.908)\}$$
$$= 0.45125 + 0.34275$$
$$= \underline{0.794}$$

∴ Information gain (e) = 0.95 − 0.794 = $\boxed{0.156}$.

f) Entropy at mid-point (800, 1000) - $\boxed{F}$

Yes = 5 No = 1 ← F → Yes = 0 No = 2

$$\text{Entropy} = -\left\{ \frac{6}{8}\left[\frac{5}{6}\log_2\left(\frac{5}{6}\right) + \frac{1}{6}\log_2(1/6)\right] + 2/8\left[\frac{0}{2}\log_2(0/2) + \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right]\right\}$$

$$= -\left\{\frac{3}{4}\left[5/6(-0.264) + 1/6(-2.582)\right] + 1/4[0+0]\right\}$$

$$= -\frac{3}{4}\left[0.833(-0.264) + 0.167(-2.582)\right]$$

$$= 3/4[0.219 + 0.431]$$

$$= 0.487$$

Information gain(f) = 0.95 - 0.487 = $\boxed{0.463}$

g) Entropy at mid-point (1000, 1200)

Yes = 5, No = 2 ← g → Yes = 0 No = 1

$$\text{Entropy} = -\left\{\frac{7}{8}\left[\frac{5}{7}\log_2\left(\frac{5}{7}\right) + \frac{2}{7}\log_2\frac{2}{7}\right] + \frac{1}{8}\left[\frac{0}{1}\log_2\frac{0}{1} + \frac{1}{1}\log_2\frac{1}{1}\right]\right\}$$

$$= -\left\{\frac{7}{8}\left[(0.714)(-0.486) + (0.285)(-1.807)\right) + 1/8[0+0]\right\}$$

$$= -\left\{\frac{7}{8}\left[-0.347 - 0.516\right]\right\}$$

$$= -\left\{7/8 \times (-0.863)\right\}$$

$$= \underline{0.755}$$

Information gain(g) = 0.95 - 0.755 = $\boxed{0.195}$

So, final list of all entropies, $\boxed{4}$.

Mid-point	Information Gain
a	0.089
b	0.200
c	0.343
d	$\boxed{0.545}$ ← highest.
e	0.156
f	0.463
g	0.195

∴ As information gain at mid-point (d) mid-point (600, 700) is highest = 0.545
so, lets split the tree at 650

Part of tree →

Rent

≤650 >650.

For further steps, to find following parts of tree, we again calculate entropy of features clean & Rent >650

Part I) For clean →

C_1 = yes C_2 = No.

P_1 (satisfied) = 1/4 P(not satisfied) = 3/4

$$\text{Entropy} = -(0.25 \log_2(0.25) + 0.75 \log_2(0.75))$$
$$= -(0.25(-2) + 0.75(-0.41))$$
$$= -(0.5 + 0.31)$$
$$= -(-0.810)$$
$$= \boxed{0.810}.$$

So for observations with > 650.

Satisfied = Yes = 1 No = 2

Clean = Yes = 1 Clean = Yes = 2

No = 0 No = 1

$$\text{Entropy} = [P_1(\text{Yes}) \cdot E(1,2) + P_2(\text{No}) \cdot E(0,1)]$$
$$= \{ 3/4 (1/2 \log_2 1/2 + 1/2 \log_2 1/2]$$
$$+ 1/4 [0/1 \log_2 0/1 + 1 \cdot \log_2 1] \}$$
$$= -\{ 3/4 (-1) \}$$
$$= \underline{0.75}$$

$\therefore$ Information gain = 0.816 − 0.75 = $\boxed{0.066}$.

II) Rent →

700	800	1000	1200
N	Y	N	N

a (between 700 and 800), b (between 800 and 1000), c (between 1000 and 1200)

a) Entropy at mid-point (700, 800)

Yes = 0 No = 1 ← a → Yes = 1 No = 2

$$\text{Entropy} = -\{ 1/4 (0/1 \log_2 0/1 + 1/1 \log_2 (1/1)]$$
$$+ 3/4 (1/3 \log_2 (1/3) + 2/3 \log_2 (2/3)] \}$$
$$= -\{ 1/4 (0+0) + 3/4 (-0.53 - 0.39) \}$$
$$= -3/4 [-0.92]$$

$\therefore$ Entropy = $\underline{0.69}$

$\therefore$ Information gain (a) = 0.816 − 0.69 = $\boxed{0.126}$

b) Entropy at mid-point (800, 1000)

yes = 1, No = 1 $\xleftarrow{}$ b $\rightarrow$ yes = 0, No = 2.

$$\text{Entropy} = -\left\{ \tfrac{2}{4}\left[\tfrac{1}{2}\log_2(\tfrac{1}{2}) + \tfrac{1}{2}\log_2(\tfrac{1}{2}) \right] + \tfrac{2}{4}\left[\tfrac{0}{2}\log_2\tfrac{0}{2} + \tfrac{2}{2}\log_2\tfrac{2}{2} \right] \right\}$$

$$= -\left\{ \tfrac{1}{2}[-1] + \tfrac{1}{2}[0] \right\}$$

$$= \boxed{0.5}.$$

$\therefore$ Information gain (b) = $0.816 - 0.5$

$= \boxed{0.316}$.

c) Entropy at mid-point (1000, 1200).

yes = 1, No = 2 $\leftarrow$ c $\rightarrow$ yes = 0 No = 1

$$\text{Entropy} = -\left\{ \tfrac{3}{4}\left[\tfrac{1}{3}\log_2(\tfrac{1}{3}) + \tfrac{2}{3}\log_2(\tfrac{2}{3}) \right] + \tfrac{1}{4}\left[\tfrac{0}{1}\log_2(\tfrac{0}{1}) + \tfrac{1}{1}\log_2(\tfrac{1}{1}) \right] \right\}$$

$$= -\left\{ \tfrac{3}{4}[-0.53 - 0.39] + \tfrac{1}{4}[0+0] \right\}$$

$$= -\tfrac{3}{4} \times (-0.92)$$

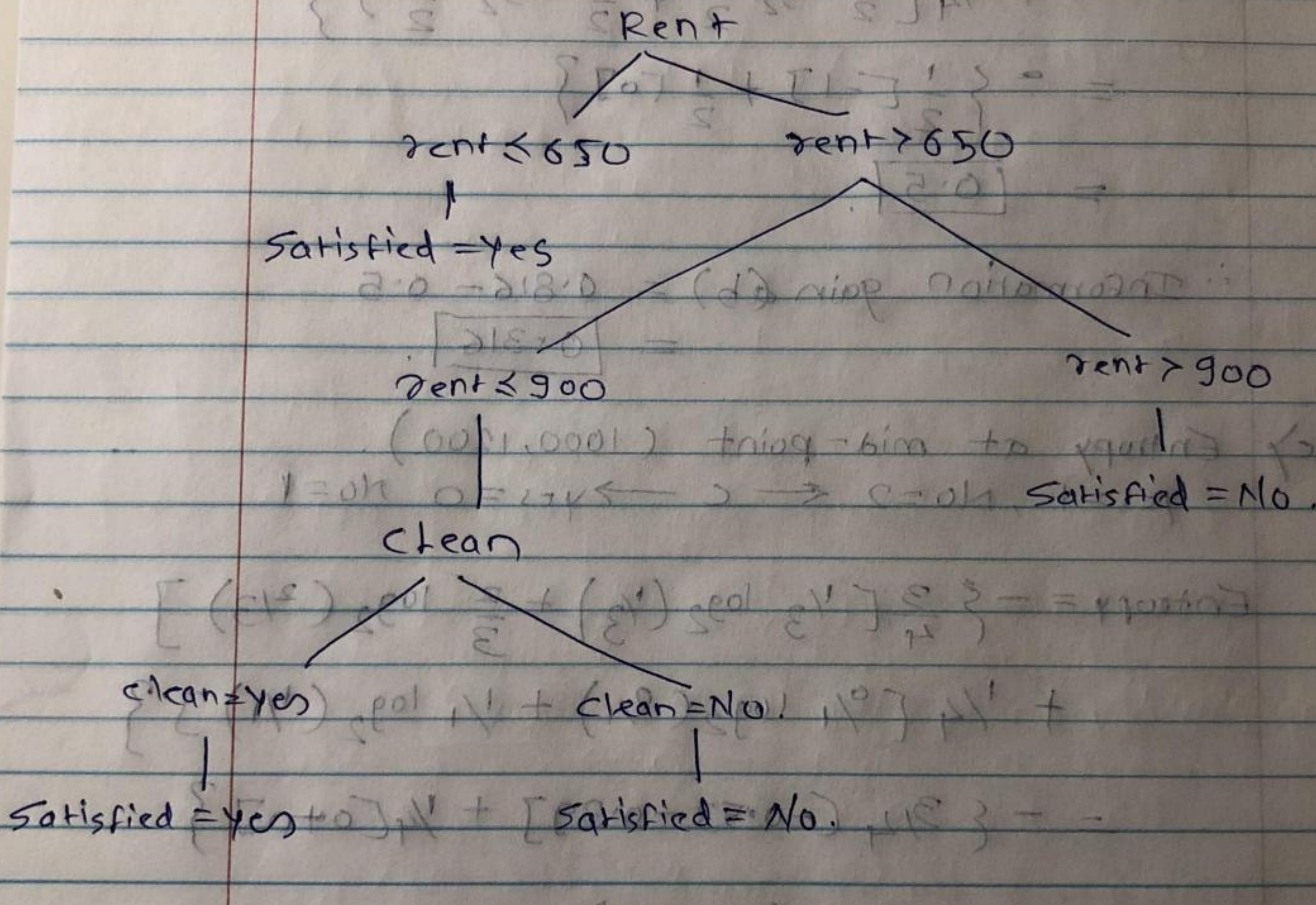
$$= \boxed{0.69}.$$

Information gain (c) = $0.816 - 0.69$

$= \boxed{0.126}$

∴ As entropy at point (b) mid-point of 800, 1000 is highest.
lets make a next split at <u>900</u>

∴ Final decision tree is,



Final Decision Tree.

Q 1) B) Prediction results for test data.

i) | Person | Rent | clean | satisfied? |
|---|---|---|---|
| P9 | 1100 | Yes | No. |

So, if we traverse the decision tree,
As Given rent = 1100
① First condition - rent > 650 as 1100 > 650
↳ ~~select right lea~~ traverse in right direction.
② Given rent = 1100
Second condition - rent > 900 & as 1100 > 900
↳ predicted value for satisfied? = $\boxed{No}$

∴ Actual given value = No.
predicted value = No.
In this case our prediction is right.

ii) | Person | Rent | clean | satisfied? |
|---|---|---|---|
| P10 | 500 | Yes | Yes |

So, if we traverse the decision tree,
①. Given rent = 500
first condition - rent ≤ 650 ∴ as 500 ≤ 650 True
↳ consider left leaf.
↳ predicted value for satisfied? = Yes.

∴ Actual given value = Yes.
predicted value = Yes.

— prediction is right

iii)

Person	Rent	clean	satisfied?
P11	1300	No	Yes.

So, Now traverse the decision tree,

1. Given rent = 1300
   first condition - rent > ~~900~~ 650 as 1300 > ~~900~~ 650 True.
   ↳ traverse in right direction.
2. Given rent = 1300
   second condition = rent > 900 as 1300 > 900 True.
   predicted value for satisfied? = No.

Predicted value = No
Actual value = Yes.

iv)

Person	Rent	clean	satisfied?
P12	550	No	Yes.

Traverse the decision tree;

1. Given rent = 550
   First condition rent ≤ 650 as 550 ≤ 650 True.
   predicted value for satisfied? = Yes.

∴ predicted value = Yes
Actual value = Yes

Q 17 c)

$$\text{Precision} = \frac{\text{True positive}}{\text{True +ve} + \text{False +ve}}$$

$$= \frac{2}{2+0} = 1 = \boxed{100\%}$$

$$\text{Recall} = \frac{\text{True +ve}}{\text{True +ve} + \text{False -ve}}$$

$$= \frac{2}{2+1} = \frac{2}{3} = \boxed{66.66\%}$$

d) Confusion matrix -

	~~True +ve~~	
	~~False~~ -ve	~~False~~ +ve

	Actual	
Predicted.	True +ve	False +ve
	False -ve	True -ve

	Actual	
Predicted.	2	0
	1	1

Q2 a) Naive Bayes.

$P(C_1) = P(\text{satisfied}) = 5/8 = 0.625$

$P(C_2) = P(\text{not satisfied}) = 3/8 = 0.375.$

$P(x_1|C_1) = P(\text{rent} / \text{satisfied})$

Given observations → 400, 500, 600, 600, 800

$$\text{mean}(\mu) = \frac{400+500+600+600+800}{5} = \boxed{580}.$$

$$\text{std. dev.}(\sigma) = \sqrt{\frac{(180)^2+(80)^2+(20)^2+(20)^2+(220)^2}{5-1}}$$

$= \boxed{148.32}.$

Variance $= \boxed{21,998.82}$

$P(x_1|C_2) = P(\text{rent} \mid \text{Not satisfied})$

Given observations – 700, 1000, 1200

$$\text{mean} = \frac{700+1000+1200}{3} = \boxed{966.66}.$$

$$\text{std dev}(\sigma) = \sqrt{\frac{(266.66)^2 + (33.34)^2 + (233.34)^2}{3-1}}$$

$$= \boxed{251.66}$$

$$\text{variance } (\sigma^2) = (251.66)^2 = \boxed{63,332.75}$$

3) I) For $P_g = x_1 = \boxed{1100}$

$P(x_1 | c_1) = P(P_g | c_1)$

$$= P(\text{rent} / \text{satisfied}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(P_g - \mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\sqrt{2 \times 3.14 \times 21998.82}} \times e^{-\frac{(1100-580)^2}{2 \times 21998.82}}$$

$$= \frac{1}{371.69} \times e^{-\frac{(270400)}{2 \times 21,998}} = \frac{0.0021}{371.69}$$

$$= \boxed{0.0000057}$$

$$P(C_1 / P_g) = 0.0000057 \times (5/8)$$
$$= \boxed{0.0000035}$$

II) $$P(P_g / c_2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(P_g - \mu)^2}{2\sigma^2}}$$

$$P(P9/C2) = \frac{1}{\sqrt{2 \times 3.14 \times (63.332.75)}} \times e^{-\frac{(1100-966.66)^2}{2 \times 63.332.75}}$$

$$= \frac{1}{630.65} e^{-0.14}$$

$$= \boxed{0.001378}$$

$$P(C2 \mid P9) = 0.001378 \times 3/8$$
$$= \boxed{0.00048}$$

$\therefore$ As $P(P9 \mid C_2) > P(P9/C1)$

$\therefore$ For $P_9 \rightarrow$ predicted outcome = Not satisfied.
Actual outcome = Not satisfied.

For P_{10}

I) $P(P_{10} / C_1) = \frac{1}{371.69} e^{-\frac{(500-580)^2}{2 \times 21998.82}}$

$$= \frac{e^{-0.145}}{371.69} = \boxed{0.00231}$$

$$P(C1 / P_{10}) = 0.00231 \times 0.625$$
$$= \boxed{0.0014}$$

II) $P(P_{10} / C_2) = \frac{1}{630.65} e^{-\frac{(500-966)^2}{2 \times 63,332.75}}$

$$= \frac{e^{-1.714}}{630.65} = \boxed{0.00028}$$

$$\therefore P(C2 / P_{10}) = 0.00028 \times 0.375$$
$$= \boxed{0.00010}$$

Hence, as $P(P_{10}|c_1) > P(P_{10}|c_2)$
Predicted = yes = satisfied.
Actual = yes = satisfied.

For P_{11} = $x_1 = 1300$

I) $P(P_{11}|c_1) = \cancel{\frac{1}{630.65}} \frac{1}{371.69} e^{-\frac{(1300-580)^2}{2\times 21,998.82}}$

$= \frac{e^{-11.78}}{371.69} = 2.06 \times e^{-8}$

$P(c_1/P_{11}) = (2.06 \times \cancel{10}\, e^{-8}) \times 5/8$
$= 1.39\, e^{-8} \simeq 1.4 e^{-8}$

II) $P(P_{11}/c_2) = \frac{1}{630.65} e^{-\frac{(1300-966)^2}{2\times 63,332.75}}$

$= \boxed{0.00066}$

$P(C_2|P_{11}) = 0.00066 \times 0.375$
$= \boxed{0.000246}$.

As, $P(P_{11}|c_2) > P(P_{11}|c_1)$
predicted = Not satisfied.
Actual = satisfied.

For P_{12} $x_1 = 550$

$P(P_{12}|c_1) = \frac{1}{371.69} e^{-\frac{(550-580)^2}{2\times 21998.82}}$

$= \frac{e^{-(0.0204)}}{371.69} = \boxed{0.00263}$.

$P(C_1|P_{12}) = 0.00263 \times (5/8)$
$= \boxed{0.001643}$.

$P(P_{12}|C_2) = \frac{1}{630.65} e^{-\frac{(550-966)^2}{2 \times 63332.75}}$

$= \boxed{0.0004}$.

$P(C_2|P_{12}) = 0.0004 \times 3/8$
$= \boxed{0.00015}$

$\therefore$ As, ~~$P(C_1|P_1$~~ $P(P_{12}|C_1) > P(C_2|P_{12})$

$\therefore$ predicted = satisfied.
Actual = satisfied.

Q 2) c) Precision $= \frac{TP}{TP+FP} = \frac{2}{2+0} = 1$

$= \boxed{100\%}$.

Recall $= \frac{\text{True +ve}}{\text{True +ve + False -ve}} = \frac{2}{2+1}$

$= 2/3 = \boxed{66.66\%}$.

Q 27 d) Confusion matrix.

	Actual	
Predicted.	2	0
	1	1

Q 3) Describe the difference between a generative model and a discriminative model.

	Generative model	Discriminative model.
①	Generative models focus on how data was generated in order to catagorize it.	Discriminative models don't care about data generation, they simply catagorize the data.
②	If you have input data x and you want to catagorize it into labels y, generative model uses joint probability distribution, given as $P(x, y)$.	If you have input data X and you want to catagorize it into labels 'Y', discriminative model uses conditional probability distribution, given as $P(Y\|X)$.
③	As generative models first learn and then catagorize, they have some discriminative properties.	As discriminative models only focus on catogorization, they don't have generative properties.
④	Generative models estimate class-conditional pdfs and prior probabilities.	Discriminative models directly estimate posterior probabilities.
⑤	Generative since different sampling methods generate synthetic data points.	Discriminative models focus their computational resources only on a given task, so, gives better performance.
	Ex Naive Bayes, sigmoidal belief networks	Ex logistic regression, suport vector machines, Traditional neural networks.