Name : Muhammad Qasim
Roll No : 19k-1612
Course : Machine Learning
Assignment : 01

QNo2

	Refurned	status	Tax Incom	cheat
1	Yes	Single	125	No
2	Yes	Single	80	Yes
3	No	Single	75	No
4	No	Single	65	Yes
5	Yes	Single	60	No
6	No	Married	120	Yes
7	No	Married	180	No
8	Yes	Married	90	Yes
9	No	Single	95	Yes

F1, T

i) Navie Bayes:

⇒ K-Fold cross-Validation (N-N/K , N/K)
N = number of instances in datasets

$P(cheat = yes) = \frac{4}{6}$, $P(cheat = NO) = \frac{2}{6}$

$P(Refum = No \mid cheat = NO) = \frac{1}{2}$, $P(Refum = Yes \mid cheat = NO) = \frac{1}{2}$
$P(status = Single \mid cheat = NO) = \frac{1}{2}$, $P(status = Married \mid cheat = NO) = \frac{2}{4}$
$\mu = 70$, $\sigma^2 = 250$ (IF class = NO)

$\mu = 92.5$ (IF class = Yes)

$$\sigma^2 = \frac{(92.5-65)^2 + (92.5-120)^2 + (92.5+90)^2 + (92.5-95)^2}{4}$$

$$\sigma^2 = \frac{1525}{4} = 381.25$$

Let $X = [R = No, S = Single, 75]$

$P(X|C=NO) = P(R=NO | C=NO) \times P(Status = Single | C=Yes) \times P(65 | C=NO)$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(250)}} e^{\frac{(75-70)^2}{2(250)}}$$

$= 6.63 \times 10^{-3}$

Let $y = [R = Yes, Single, 80]$

$P(y|C=Yes) = P(R=Yes | C=Yes) \times P(Single | C=Yes) \times P(80 | C=Yes)$

$$= \frac{1}{4} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(381.25)}} e^{\frac{(80-92.5)^2}{2(92.5)}}$$

$= 3.76 \times 10^{-3}$

$$P(X|C=NO) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(250)}} e^{\frac{(80-70)^2}{500}}$$

$= 7.70 \times 10^{-3}$

$$P(X|C=Yes) = \frac{3}{4} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(381.25)}} e^{\frac{(75-92.5)^2}{2(92.5)}}$$

$= 7.99 \times 10^{-3}$

$\because P(X|C=Yes) > P(X|C=NO)$

$\therefore$ cheat = Yes for X

$$P(y \mid c=NO) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(250)}} \, e^{\frac{(80-70)^2}{500}}$$

$$= 7.70 \times 10^{-3}$$

$\because P(y \mid c=NO) > P(y \mid c=yes)$

$\therefore$ cheat $=$ No for y

Let $Z = [R=\text{Yes}, \text{single}, 125]$

$$P(Z \mid c=yes) = \frac{1}{4} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(381.25}} \, e^{\frac{(125-72.5)^2}{2(381.5)}}$$

$$= \boxed{0.094}$$

$$P[Z / c=NO] = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{\sqrt{2\pi(500)}} \, e^{\frac{(125-70)^2}{500}}$$

$$= 2.67$$

$\because P(Z \mid c=NO) > P(Z \mid c=Yes)$

$\therefore$ cheat $=$ No for Z

$$\text{Accuracy1} = \frac{TP+TN}{TP+TN+FP+FN} = \boxed{\frac{1}{3} = 0.33}$$

predict

Actual	False	True
False	1	1
True	1	0

precision1 : $\frac{TP}{TP+FP} = 0$, Recall1 $= \frac{TP}{TP+FN} = 0$

F1 Score1 $= 0$

For Second Fold:

Refumed	status	Income tax	cheat
NO	single	65	yes
Yes	Single	60	NO
NO	Married	120	Yes,

Test data

Figure 1.2

$P(yes) = \frac{3}{6}$, $P(C=NO) = \frac{3}{6}$

$P(Ref=No \mid c=Yes) = \frac{1}{3}$, $P(Ref=Yes \mid c=Yes) = 2/3$

$P(Ref=NO \mid C=NO) = 2/3$, $P(Ref=Yes \mid C=NO) = 1/3$

$P(single \mid c=Yes) = 2/3$, $P(single \mid C=NO) = 1/3$

$P(married \mid c=Yes) = 1/3$, $P(married \mid C=NO) = 2/3$

if $C=NO \rightarrow \mu = 93.3$, $\sigma^2 = 505.5$

if $C=Yes \rightarrow \mu = 88.33$, $\sigma^2 = 38.88$

Let $X = \{No, Single, 65\}$

$$P(X \mid C=NO) = 2/3 \times 1/3 \times \frac{1 e^{\frac{(65-93.3)^2}{2150.5}}}{\sqrt{2\pi(505.5)}}$$

$$= 8.707 \times 10^{-3}$$

$$P(X \mid C=yes) = 1/3 \times 2/3 \times \frac{e^{\frac{(65-38.88)^2}{2(38.88)}}}{\sqrt{2\pi(38.88)}}$$

$$= 15.58 > P(X \mid C=NO)$$

$\therefore$ cheat = Yes for X

Let $y = (No, Single, 60)$

$$P(y|c=No) = \frac{2}{3} , \frac{1}{3} \times \frac{e^{\frac{(60-93.3)}{2(505.5)}}}{\sqrt{2\pi(505.5)}}$$

$$= \boxed{0.0118}$$

$$P(y|c=Yes) = \frac{2}{3} \times \frac{2}{3} \times \frac{1 \, e^{\frac{(60-88.33)^2}{2(38.88)}}}{\sqrt{2\pi(38.88)}}$$

$$= \boxed{863.721} > P(y|c=No)$$

$\therefore$ cheat = yes for y

Let $Z = [No, married, 120]$

$$P(Z|c \neq Yes) = \frac{1}{\cancel{3}} \times \frac{1}{\cancel{3}} \times 2841.58$$

$$P(Z|c=No) = \frac{2}{3} \times \frac{2}{3} \times (P(120|C=No))$$

$$= 1446.83$$

Accuracy$^2 = \frac{2}{3}$

A \ P	F	T
F	0	1
T	0	2

precision$_2$ $= \frac{TP}{TP+FP} = \boxed{0.66}$

Recall2 $= \frac{TP}{TP+FN} = \frac{2}{2} = 1$

F1 Score2 $= \frac{2 \times P \times 1}{P+1} = 0.795$

For Third Fold:

Refumed	status	IncomeTax	cheat
No	Married	80	No
Yes	Married	90	Yes
No	Single	95	Yes

Test data For Fold3

$P(Yes) = \frac{1}{2}$, $P(NO) = \frac{1}{2}$

$P(Ref=Yes \mid c=Yes) = \frac{1}{3}$, $P(Ref=Yes \mid c=NO) = 2/3$
$P(Ref=NO \mid c=Yes) = 2/3$, $P(Ref=NO \mid C=NO) = 1/3$
$P(single \mid c=Yes) = 2/3$, $P(single \mid C=NO) = 3/3 = 1$
$P(married \mid c=yes) = 1/3$, $P(married \mid c=NO) = 0/3 = 0$

IF $c=Yes \rightarrow \mu = 88.33$, $\sigma^2 = 538.88$
IF $c=NO \rightarrow \mu = 86.66$, $\sigma^2 = 772.22$

Let $X = (No, married, 80)$
$P(X \mid c=Yes) = \frac{2}{3} \times \frac{1}{3} \, P(80 \mid c=yes)$

$P(X \mid C=NO) = 0$
$\therefore$ cheat = Yes for X

Let $y = (Yes, married, 90)$
$P(y \mid c=Yes) = 0$
$\therefore$ cheat = Yes for y

Let $Z = (No, Single, 95)$

$P(Z|c=yes) = \frac{2}{3} \times \frac{2}{3} \times P(95|c=yes) = 7.95 \times 10^{-3}$

$P(Z|c=NO) = \frac{1}{3} \times 1 \times P(95/c=NO) = 5.005 \times 10^{-3}$

∴ cheat = yes for Z

Accuracy3 = $\frac{2}{3}$

recall3 = 1

F1Score3 = 0.795

		P	
		F	T
	F	0	1
	T	0	2

Navie Bayes

$\boxed{\begin{array}{l} \text{Accuracy} = 0.44 \\ \text{F1 Score} = 0.53 \end{array}}$ Ans

ii) KNN (3, CV)

Assuming $K=1$

Yes = 1 , NO = 0 , Single = 0 , Married = 1

"Vectorization"

For Training data:

Figure 1.1

	Refund	status	Income Tax	cheat
1	0	0	65	Yes
2	1	0	60	No
3	0	1	120	Yes
4	0	1	80	No
5	1	1	90	Yes
6	0	0	95	Yes

Training

$K=1$

				$\hat{y}$ predicts	y Actual
①	1	0	125	Yes	NO
②	1	0	80	NO	Yes
③	0	0	75	NO	NO

Testing

$[1 \quad 0 \quad 125 \quad NO]$

① $\sqrt{(1-0)^2+(0-0)^2+(125-65)^2} = 60$ Yes

② $\sqrt{(1-1)^2+(0-0)^2+(125-60)^2} = 65$ No.

③ $\sqrt{(1-0)^2+(0-1)^2+(125-120)^2} = \boxed{5.19 \quad Yes}$

④ $\sqrt{(1-0)^2+(0-1)^2+(125-80)^2} = 45.02$ NO

⑤ $\sqrt{(1-1)^2+(0-1)^2+(125-90)^2} = 35.01$ Yes

⑥ $\sqrt{(1-0)^2+(0-0)^2+(125-95)^2} = 30.01$ Yes

Test data 2

[1 0 80 yes]

① $\sqrt{(1-0)^2 + (0-0)^2 + (80-65)^2} = 15.03$

② $\sqrt{(1-1)^2 + (0-0)^2 + (80-60)^2} = 20$

③ $\sqrt{(1-0)^2 + (0-1)^2 + (80-120)^2} = 40.02$

④ $\sqrt{(1-0)^2 + (0-1)^2 + (80-80)^2} = 2$ NO

⑤ $\sqrt{(1-1)^2 + (0-1)^2 + (80-90)^2} = 10.04$

⑥ $\sqrt{(1-0)^2 + (0-0)^2 + (80-95)^2} = 15.033$

Test dat 3: [0-0 0 75 NO]

① $\sqrt{(0-0)^2 + (0-0)^2 + (75-65)^2} = 10$

② $\sqrt{(0-\cancel{1})^2 + (0-0)^2 + (75-60)^2} = 15.03$

③ $\sqrt{(0-0)^2 + (0-1)^2 + (75-120)^2} = 45.0111$

④ $\sqrt{(0-0)^2 + (0-1)^2 + (75-80)^2} = 5.09$ NO

⑤ $\sqrt{(0-1)^2 + (0-1)^2 + (75-90)^2} = 15.06$

⑥ $\sqrt{(0-0)^2 + (0-0)^2 + (7)-95)^2} = 20$

Accuracy $= \frac{1}{3}$

precision = 0

Recall = 0

F1 Score = 0

Actual \ predict	F	T
F	1	1
T	1	0

K = 2 fold

Table 1.1

	Refund	status	IncomeTax	cheat
1	1	0	125	No
2	1	0	80	Yes
3	0	0	75	No
4	0	1	80	NO
5	1	1	90	Yes
6	0	0	95	NO

Training

				y	$\hat{y}$
T1 =	0	0	65	Yes	NO
T2 =	1	0	60	NO	
T3 =	0	1	120	Yes	

T1 Loss

1 $= \sqrt{(0-1)^2+(0-0)^2+(65-125)^2} = 60$

2 $= \sqrt{(0-1)^2+(0-0)^2+(65-80)^2} = 15.03$

3 $= \sqrt{(0-0)^2+(0-0)^2+(65-75)^2} = 10$ NO

4 $= \sqrt{(0-0)^2+(0-1)^2+(65-80)^2} = 15.03$

5 $= \sqrt{(0-1)^2+(0-1)^2+(65-90)^2} = 25.03$

6 $= \sqrt{(0-0)^2+(0-0)^2+(65-95)^2} = 30$

Test 2 Losses:		Test3 Losses:	
1	65	1	5.19 · NO
2	20	2	40.02
3	15.03 NO	3	45.011
4	20.04	4	40
5	30.014	5	30.01
6	35.01	6	25.01

Accuracy: $\frac{1}{3}$

precision = 0

Recall = 0

F1-Score = 0

		P	
		F	T
A	F	1	0
	T	2	0

⇒ K = 3 Fold:

Table 2.2

	Refund	status	Income/tax	cheat
1	1	0	125	No
2	1	0	80	yes
3	0	0	75	No
4	0	0	65	yes
5	1	0	60	No
6	0	1	120	yes

} Training

			y	$\hat{y}$
0	1	80	No	yes
1	1	90	yes	yes
0	0	95	yes	yes

①	$\sqrt{(0-1)^2+(1-0)^2+(80-125)^2} = 45$	$\sqrt{(1-1)^2+(1-0)^2+(90-125)^2} = 35$	$\sqrt{(0-1)^2+(0-0)^2+(95-125)^2} = 3$
②	" $= 10.41$	$= 10.4$	$= 15.03$
③	" $= 5.09$	$= 15.06$	$= 20$
④	" $= 15.03$	$= 25.03$	$= 30$
⑤	" $= 20$	$= 30.01$	$= 35.01$
⑥	" $= 40$	$= 30.01$	$= 25.01$

Accuracy3 $= \frac{2}{3}$

precision3 $= \frac{2}{2+0} = 1$

Recall3 $= \frac{2}{3} = 0.66$

F1-Score3 $= 2 \times \frac{P \times R}{P \times P} = 0.792$

Actual	Predict F	Predict T
F	0	1
T	0	2

KNN

Accuracy = 0.44
F1-Score = 0.264

④ C4.5 : (upto one Level)

⇒ K=1 fold = 18

Refund	Status	IncomeTax	cheat
No	S	65	Yes
yes	S	60	No
No	M	120	yes
No	M	80	No
yes	M	90	yes
No	S	95	yes

Training data

⇒ Using Information Gain to get best Split

① Refund → "No"

Info[3, 1] = entropy [3/4 , 1/4]

$= -\log_2(3/4) - \log_2(1/4)$

= 0.881 bits

② Refund → "yes"

Info(1, 1] = entropy[1/2 , 1/2]

= 1 bits

Info [(3,1] , [1,1]] = 0.881 (4/6) + (2/6) (1)

= .92 bits

Status Single:

Info([2,1]) = entropy(2/3, 1/3)
= 0.918 bits

Status Married:

Info([2,3]) = entropy(2/ ..)
= 0.918 bits

Info([2,1],[2,3]) = $\cancel{2}$ (1/$\cancel{2}$) 0.918
= 0.918 bits

=> For Income Tax:

1. Income ≥ 90

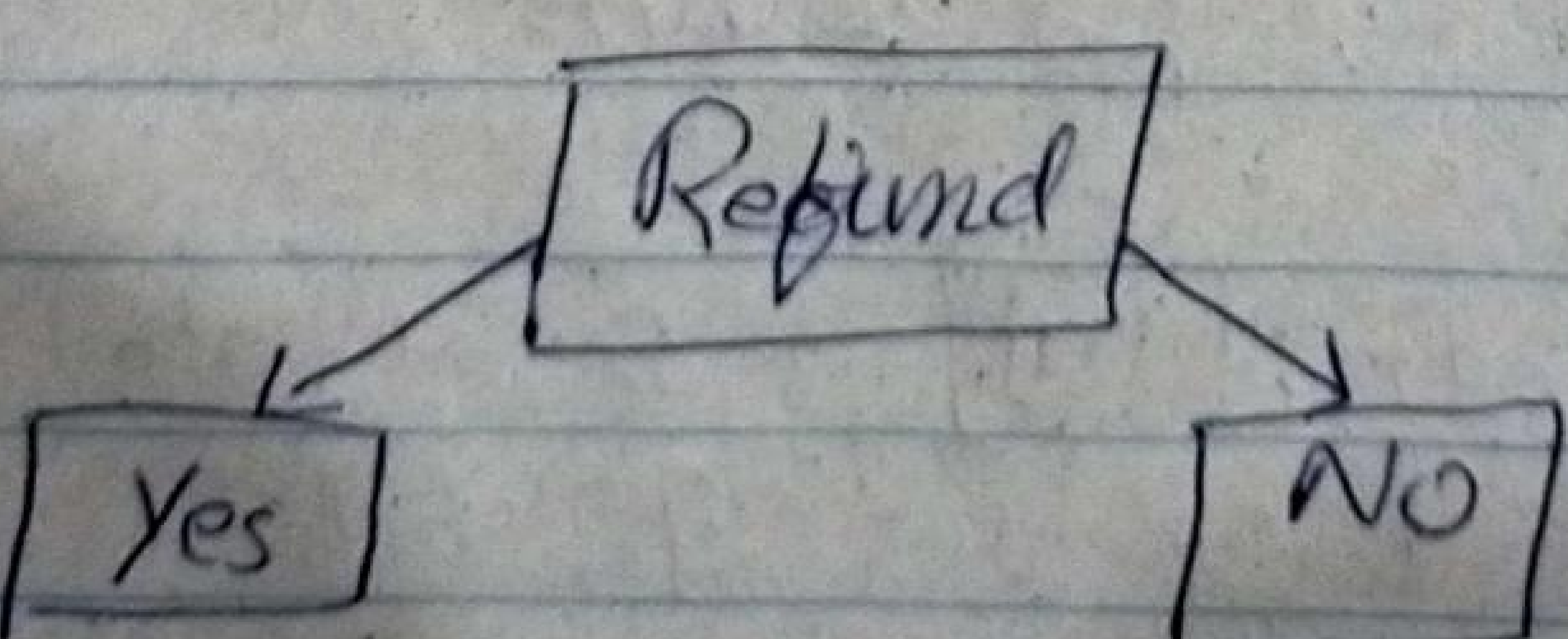
Info([3,0]) = 0 bits

2. Income < 90

Info([1,2]) = 0.918 bits

Info[(1,2),(3,0)] = (1/3) 0.918
= 0.306 bits

∴ Refund is Root and Node:



For Test data ① : [Yes single , 125 , No]

From tree → cheat = No

For Test data ② : [Yes single , 80 , Yes]
From Tree → cheat = NO
For Test dat ③ : [No single 75 No]
From Tree → cheat = Yes

Accuracy : $\frac{1}{3}$

Actual \ predict	F	T
F	1	1
T	1	0

Precision = 0
Recall = 0
F1-Score = 0

Table 4.1 page No # 10

K-2 fold : 2 :

Refund → "Yes"
Info([2,1]) = entropy(2/3 , 1/3)
= 0.918 bits
Refund → "No"
Info[(1,2)] = 0.918 bits

Info[(2,1],[1,2]) = 0.98 bits

Status → "Single"

Info([[1,1]]) = entropy[1/2 , 1/2]

= 1 bits

Info([1,1]) = 1 bits

Info[[1,1], [1,1]] = 1 bits

Incom Tax → ≥ 90

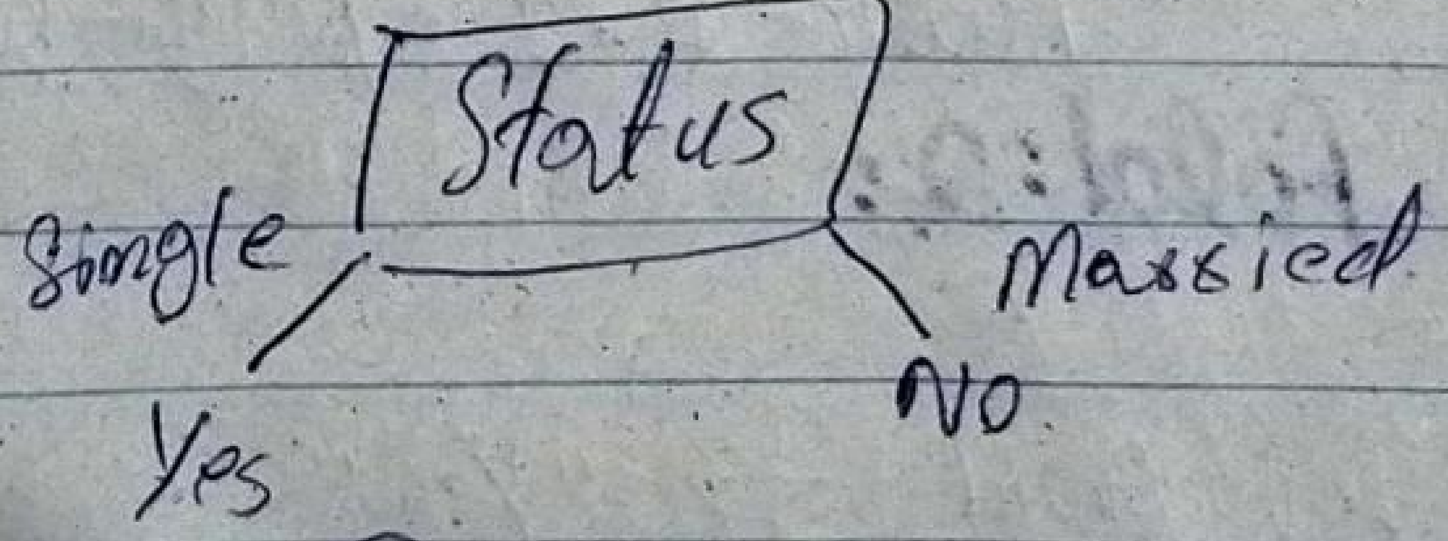
Info([2,1]) = 0.918 bits

Incom Tax → < 90

Info[1,2]) = 0.918 bits

Info[[2,1],[1,2]] = 0.918

Status is root Node:



for test data 1 = [No single 65, Yes]

cheat = yes

for test data 2 = [Yes single 60 No]

cheat = No

for test data 3 = [No married 120 yes]

cheat = No

Accuracy2 = 1/3 , precision = 0.5

Recall2 = 1

F1_Score2 = 0.66

		P	
		F	T
A	F	0	1
	T	1	1

K-3 Fold-3

Table 2.2 on page No 11

Entropy (cheat) = Entropy (1/2 , 1/2)
= 1 bits

Refund → "Yes"
Info[[1, 2]] = 0.918 bits

Refund → "No"
Info[[2, 1]] = 0.918 bits

Info[[1, 2], [2, 1]] = 0.918 bits

⇒ Status → "Sing"
Info[[2, 3]] = entropy[2/5 , 3/5]
= 0.970

Status → "Married"
Info[1, 0] = 0 bits

Info[[2, 3], [1, 0]] = (5/6) 0.970 + 0
= 0.808

Gain = 1 - 0.808 = 0.191

① Income Tax → ≥ 90
Info([1, 1]) = 1 bits

② Income Tax → < 90
Info([2, 2]) = 1 bits

Info[[1, 1], [2, 2]] = 1/2 ¹ + 1/2 ¹
= 1

Gain = 1 - 1 = 0 bit
This status is Root Node

$$\begin{array}{ccc} & \text{Status} & \\ \swarrow^{\text{Married}} & & \searrow^{\text{Single}} \\ \text{Yes} & & \text{No} \end{array}$$

for test data (1) [No Married 80 No]
cheat → Yes
for test data (2) [Yes Married 90 Yes]
cheat → Yes
for test data (3) [No Single 80 · Yes]
cheat → No

A \ P	F	T
F	0	1
T	1	1

Accuracy = $\frac{1}{3}$

precision = $\frac{1}{1+1} = \frac{1}{2}$

Recall = $\frac{1}{1+0} = 1$

F1_Score = 0.66

C4.5 (upto One Level)
Accuracy : 0.33
F1_Score : 0.44

iv CART : (up to One Level)

K-1:

for First Fold:

Assuming KNN or C4.5 Tables:

For Refund

		Cheat yes	Cheat NO	
Refund	yes	1	1	2
	NO	3	1	4
		4	2	6

$\text{Gini}(\text{Refund} = \text{yes}) = 1 - (1/2)^2 - (1/2)^2$
$= 0.5$

$\text{Gini}(\text{Refund} = \text{No}) = 1 - (1/4)^2 - (1/4)^2$
$= 0.375$

$\text{Gini}(\text{Refund}) = (2/6)(0.5) + (0.375)(4/6)$
$= 0.466$

For Status:

	Cheat yes	Cheat NO	
S	2	1	3
M	2	1	3
	4	2	6

$\text{Gini}(S \to \text{single}) = 1 - (2/3)^2 - (1/3)^2$
$= 0.44$

$\text{Gini}(\text{status} \to \text{Married})$
$\rightarrow = 1 - (2/3)^2 - (1/3)^2$
$= 0.44$

$\text{Gini}(\text{status}) = (1/2)(0.44) + (1/2)(0.44) = \boxed{0.44}$

⇒ For Income Tax

cheat

	Yes	No	
≥90	3	0	3
<90	1	2	3
	4	2	6

$Gini(income \geq 90) = 1 - (3/3)^2 - (0)^2$
$= 0$

$Gini(income < 90) = 1 - (1/3)^2 - (2/3)^2$
$= 0.44$

$Gini(IncomeTax) = 0.44(1/2) = 0.22$

income has low index ∴ it is root

Income Tax
≥90 / \ <90
yes No

For Yes data:

① [Yes Single 125 No]
cheat = Yes
② [Yes Single 80 Yes]
cheat = No
③ [No Single 96 No]
cheat No

For Second Fold:

Reffering 3rd KNN training data:

For Refunds

		Cheat Yes	Cheat No	
Refund		2	1	3
		1	2	3
		3	3	6

$$\text{Gini (Refund=Yes)} = 1-(2/3)^2-(1/3)^2$$
$$=0.44$$
$$\text{Gini (Refund=No)} = 1-(1/3)^2-(2/3)^2$$
$$=0.44$$
$$\text{Gini (Refund)} = 0.44$$

For Status

		Cheat Yes	Cheat No	
Status	S	2	3	5
	M	1	0	1
		3	3	6

$$\text{Gini (status = Single)} = 1-(2/5)^2-(3/5)^2$$
$$=0.48$$
$$\text{Gini (status = Married)} = 1-(1/1)^2$$
$$=0$$
$$\text{Gini (Status)} = 0.48(5/6)+0 = 0.4$$

For Income Tax:

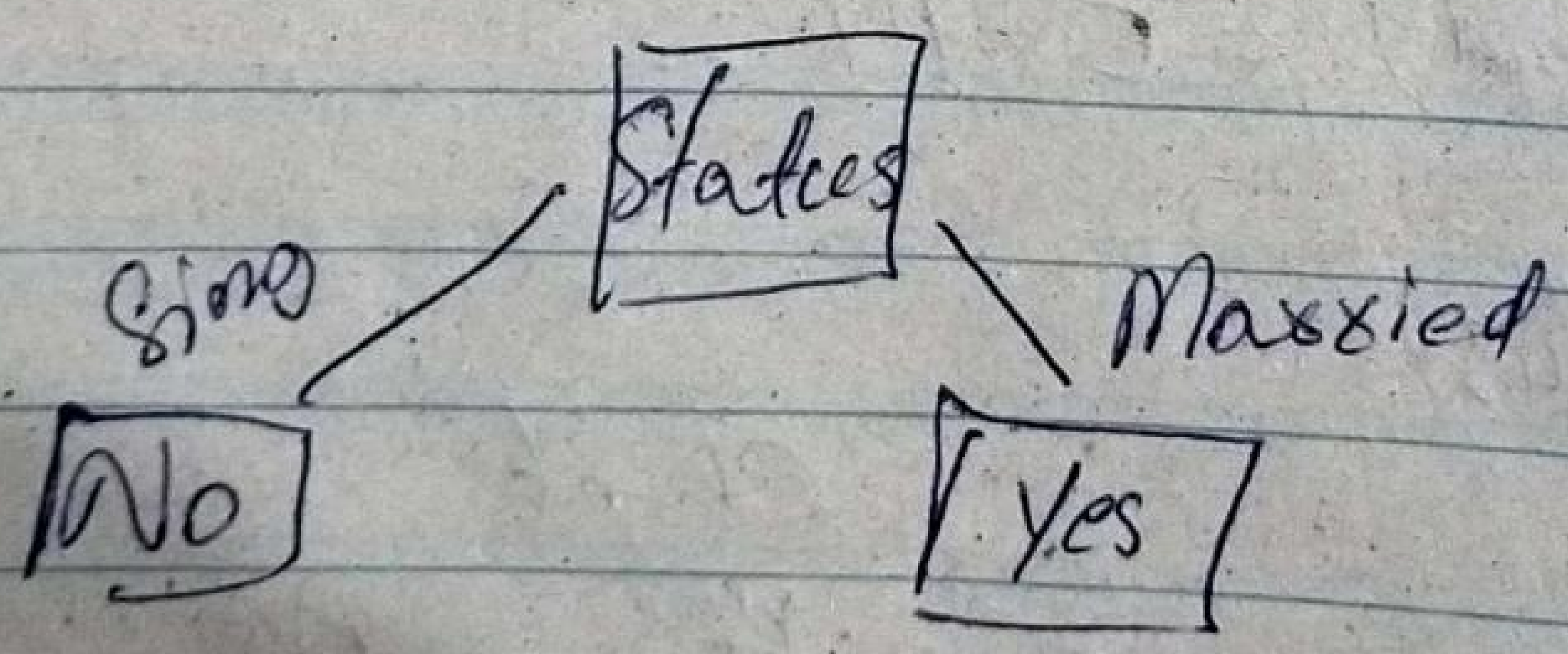
		cheat Yes	No	
Income	≥ 90	1	1	2
	< 90	2	2	4
		3	3	6

$$Gini(income \geq 90) = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(income < 90) = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(income) = \boxed{0.5}$$

Status is root Node:



Test data

① [No , Married 80, No] → Yes
② [Yes , Married 90 Yes] → Yes
③ [No . Single 95 Yes] → No

Accuracy: $\frac{1}{3}$

F1_Score: 0.66

		P F	P T
A	F	0	1
	T	1	1

K-3 Fold:3:

Second Traing data for KNN

8 For Refund

		Cheat Yes	Cheat No	
Refund	Yes	2	1	3
	No	1	2	3
		3	3	6

Gini (Refund) = 0.44

For Status 8

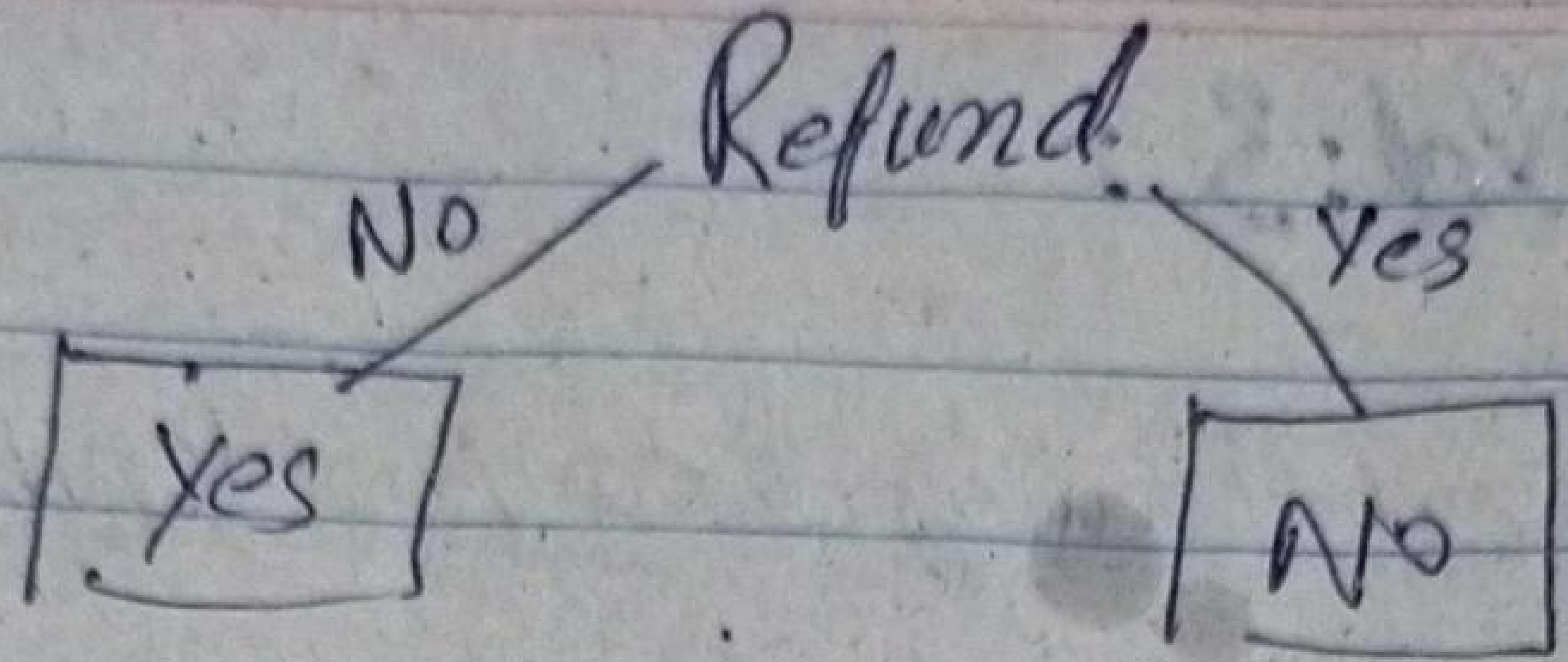
		Cheat Yes	Cheat No	
STATUS	M	1	1	2
	S	2	2	4
		6	6	6

Gini (status) = 0.5

For Income Tax

	Cheat Yes	Cheat No	
≥90	2	1	3
<90	1	2	3
	3	3	6

$\boxed{\text{Gini (Income) = 0.44}}$



For Test data

① [No Single 65 yes] → No
② [Yes Married 60 No] → yes
③ [No Married 120 yes] → No

	P: F	P: T
F	0	1
T	2	0

CART
Accuracy = 0.22
F1-Score = 0.20