
DS503 Machine Learning for Data Science

**Assignment 02**

Name: Muhammad Qasim
Roll No: 19k-1612

| Refund | Status | Tax Income | cheat |
|---|---|---|---|
| Yes | Single | 115 | No |
| Yes | Single | 60 | No |
| Yes | Single | 40 | Yes |
| Yes | Married | 130 | Yes |
| Yes | Married | 80 | No |
| Yes | Divorced | 130 | Yes |
| Yes | Divorced | 80 | No |
| No | Single | 75 | No |
| No | Single | 65 | Yes |
| No | Married | 125 | Yes |
| No | Married | 80 | No |
| No | Divorced | 90 | Yes |
| No | Divorced | 75 | Yes |

C4.5:-

Discrete Tax Income:

range = 40 – 130

Buckets

low = 40 – 70 (Tax Income)

medium = 90 – 100 (Tax Income)

high = 100 – 130 (Tax Income)

| Refund | Status | Tax Income | cheat |
|---|---|---|---|
| Yes | Single | high | No |
| Yes | Single | low | No |
| Yes | Single | low | Yes |
| Yes | Married | high | Yes |
| Yes | Married | medium | No |
| Yes | Divorced | high | Yes |

| | | | |
|---|---|---|---|
| Yes | Divorced | medium | No |
| NO | Single | medium | NO |
| NO | Single | low | Yes |
| No | Married | high | Yes |
| NO | Married | medium | No |
| NO | Divorced | medium | Yes |
| NO | Divorced | medium | Yes |

(A) Entropy (~~CART~~):-

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

$$E(\text{cheat}) = -\frac{6}{13}\log_2\left(\frac{6}{13}\right) - \frac{7}{13}\log_2\left(\frac{7}{13}\right)$$

$$E(\text{cheat}) = -0.46 \log_2(0.46) - 0.53 \log_2(0.53)$$

$$E(\text{cheat}) = 1.0007$$

1st Iteration:-

(A) Refund:-

| | | cheat | | Total |
|---|---|---|---|---|
| | | Yes | N | |
| Refund | Yes | 3 | 4 | 7 |
| | NO | 4 | 2 | 6 |
| | | | | 13 |

$$E(\text{cheat}, \text{refund}) = P(\text{Yes}) * E(3,4) + P(\text{NO}) * E(4,2) \quad \text{—(i)}$$

$$E(3,4) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= -0.42\log_2(0.42) - 0.57\log_2(0.57)$$

$$\boxed{E(3,4) = 0.987} \quad -(a)$$

$$E(4,2) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right)$$

$$E(4,2) = -0.66 * \log_2(0.66) - 0.33 * \log_2(0.33)$$

$$\boxed{E(4,2) = 0.923} \quad -(b)$$

Subsitituting (a) and (b) in (i)

$$(i) \Rightarrow E(\text{cheat, refund}) = \frac{7}{13} * 0.987 + \frac{6}{13} * 0.923$$

$$= 0.53 * 0.987 + 0.46 * 0.923$$

$$\boxed{E(\text{cheat, refund}) = 0.947}$$

(B) status

| | | cheat | | Total |
|---|---|---|---|---|
| | | Yes | No | |
| | Single | 2 | 3 | 5 |
| status | Divorced | 3 | 1 | 4 |
| | Married | 2 | 2 | 4 |
| | | | | 13 |

(ii) $E(\text{cheat, status}) = P(\text{Single}) * E(2,3) + P(\text{Divorced}) * E(3,1) + P(\text{Married}) * E(2,2)$

$$E(2,3) = -\frac{2}{5} * \log_2\left(\frac{2}{5}\right) - \frac{3}{5} * \log_2\left(\frac{3}{5}\right)$$

$$= -0.4 * \log_2(0.4) - 0.6 * \log_2(0.6)$$

$$\boxed{E(2,3) = 0.970} \quad - (c)$$

$$E(3,1) = -\frac{3}{4} * \log_2\left(\frac{3}{4}\right) - \frac{1}{4} * \log_2\left(\frac{1}{4}\right)$$

$$= -0.75 * \log_2(0.75) - 0.25 * \log_2(0.25)$$

$$\boxed{E(3,1) = 0.811} \quad - (d)$$

$$E(2,2) = -\frac{2}{4} * \log_2\left(\frac{2}{4}\right) - \frac{2}{4} * \log_2\left(\frac{2}{4}\right)$$

$$= -0.5 * \log_2(0.5) - 0.5 \log_2(0.5)$$

$$\boxed{E(2,2) = 1} \quad - (e)$$

Substituting (c), (d) and (e) in (ii)

$$(ii) \Rightarrow E(\text{cheat, status}) = \frac{5}{13} * 0.970 + \frac{4}{13} * 0.81 + \frac{4}{13} * 1$$

$$= 0.38 * 0.970 + 0.30 * 0.81 + 0.30 * 1$$

$$= 0.36 + 0.243 + 0.30$$

$$\boxed{E(\text{Cheat, status}) = 0.90}$$

(c) Tax Income :-

| | | cheat | | |
|---|---|---|---|---|
| | | Yes | NO | Total |
| Tax Income | low | 2 | 1 | 3 |
| | medium | 2 | 4 | 6 |
| | high | 3 | 1 | 4 |
| | | | | 13 |

(iii) $E(\text{cheat}, \text{Tax Income}) = P(\text{low}) * E(2,1) + P(\text{medium}) * E(2,4) + P(\text{high}) * E(3,1)$

$$E(2,1) = -\frac{2}{3} * \log_2\left(\frac{2}{3}\right) - \frac{1}{3} * \log_2\left(\frac{1}{3}\right)$$

$$= -0.66 * \log_2(0.66) - 0.33 * \log_2(0.33)$$

$$\boxed{E(2,1) = 0.923} \quad - (f)$$

Substituting (b) (d) and (f) in (iii)

$$E(\text{cheat}, \text{Tax Income}) = \frac{3}{13} * 0.923 + \frac{6}{13} * 0.923 + \frac{4}{13} * (0.811)$$

$$= 0.213 + 0.426 + 0.249$$

$$\boxed{E(\text{cheat}, \text{Tax Income}) = 0.88}$$

Gain (cheat, refund) = 1 − 0.947

**Gain (cheat, refund) = 0.05**

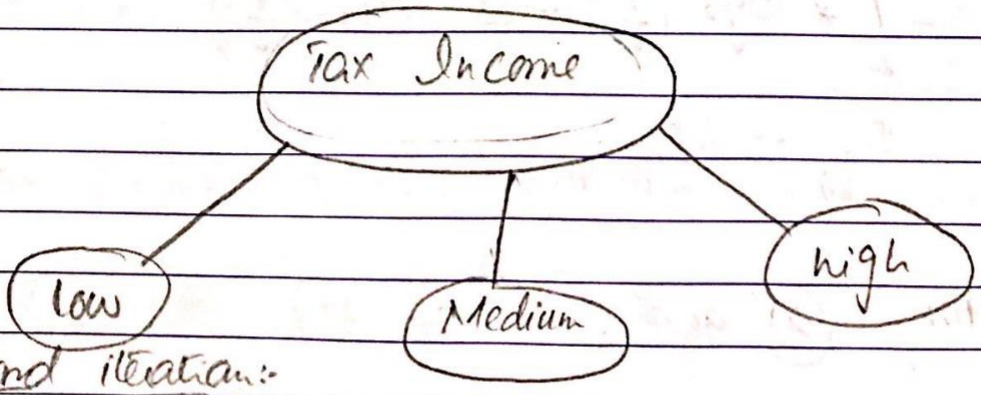
Gain (cheat, Status) = 1 − 0.90

**Gain (cheat, Status) = 0.10**

Gain (cheat, Tax Income) = 1 − 0.88

**Gain (cheat, Tax Income) = 0.12**

"Tax Income will be root node because of maximum Information gain"




Second iteration:-

Low table:-

| Refund | Status | cheat |
|---|---|---|
| Yes | Single | No |
| Yes | Single | Yes |
| No | Single | Yes |

$$E(cheat) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$\boxed{E(cheat) = 0.970}$$

(A) refund :-

| | | cheat | | Total |
|---|---|---|---|---|
| | | Yes | NO | |
| | Yes | 1 | 1 | 2 |
| refund | NO | 1 | 0 | 1 |
| | | | | 3 |

$$E(cheat, refund) = P(Yes) * E(1,1) + P(NO) * E(1,0) \quad -(i)$$

$$\boxed{E(1,0) = 0} \quad -(a)$$

$$E(1,1) = -\frac{1}{2} * \log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$\boxed{E(1,1) = 1} \quad -(b)$$

Subsituting (a) and (b) in (i)

$$(i) \Rightarrow E(cheat, refund) = \frac{2}{3} * 1 + \frac{1}{2} * 0$$

$$\boxed{E(cheat, refund) = 0.66}$$

(B) Status

Date:

cheat

| | | Yes | No | Total |
|---|---|---|---|---|
| Status | Married | 0 | 0 | 0 |
| | Single | 2 | 1 | 3 |
| | Divorced | 0 | 0 | 0 |
| | | | | 3 |

$E(\text{cheat, status}) = P(\text{Single}) * E(2,1) + P(\text{Married}) * 0 + P(\text{Divorced}) * 0$ — (ii)

using previous knowledge from iteration of in (i)

$E(2,1) = 0.923$

(ii) $E(\text{cheat, status}) = \frac{3}{3} * 0.923$

$E(\text{cheat, status}) = 0.923.$

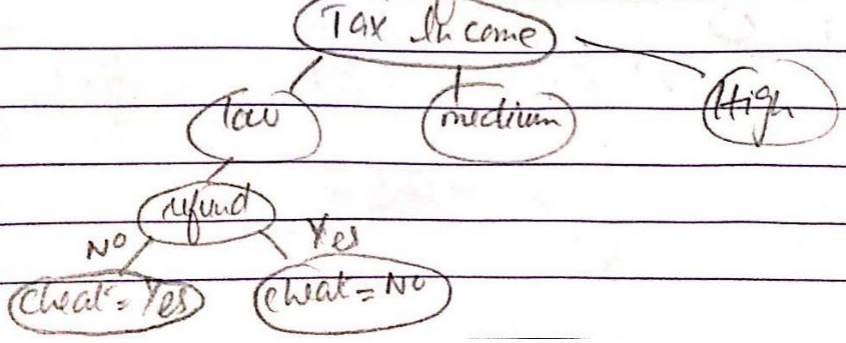
$\text{Gain (cheat, refund)} = 0.970 - 0.66$

$\text{Gain(cheat, refund)} = 0.31$

$\text{Gain (cheat, Status)} = 0.970 - 0.923$

$\text{Gain (cheat, status)} = 0.047$

"Second node will be refund in low branch".



Weigh table :-

| refund | Status | cheat |
|---|---|---|
| Yes | Single | No |
| Yes | Married | Yes |
| No | Married | yes |
| Yes | Divorced | Yes |

$$E(\text{cheat}) = -\frac{3}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= 0.5 + 0.5 + 0.5$$

$$\boxed{E(\text{cheat}) = 1.5}$$

(A) Refund :-

| | | Yes | No | Total |
|---|---|---|---|---|
| refund | Yes | 2 | 1 | 3 |
| | No | 1 | 0 | 1 |
| | | | | 4 |

(i) $E(\text{cheat, refund}) = P(\text{Yes}) * E(2,1) + P(\text{No}) * E(1,0)$

$E(1,0) = 0$

$E(2,1) = 0.923$ (using prior knowledge)

(i) $E(\text{cheat, refund}) = \frac{3}{4} * 0.923 + \frac{1}{4} * 0$

E(cheat, refund) = 0.692

(ii) Status:-

| | | Yes | No | Total |
|---|---|---|---|---|
| Status | Single | 0 | 1 | 1 |
| | Married | 2 | 0 | 2 |
| | Divorced | 1 | 0 | 1 |
| | | | | 4 |

E(cheat, status) = P(Married) * E(2,0) + P(Single) * E(0,1)
+P(Divorced) * E(1,0) — (ii)

E(2,0) = E(0,1) = E(1,0) = 0

(ii)

E(cheat, status) = 0

Gain (cheat, refund) = 1.5 - 0.692

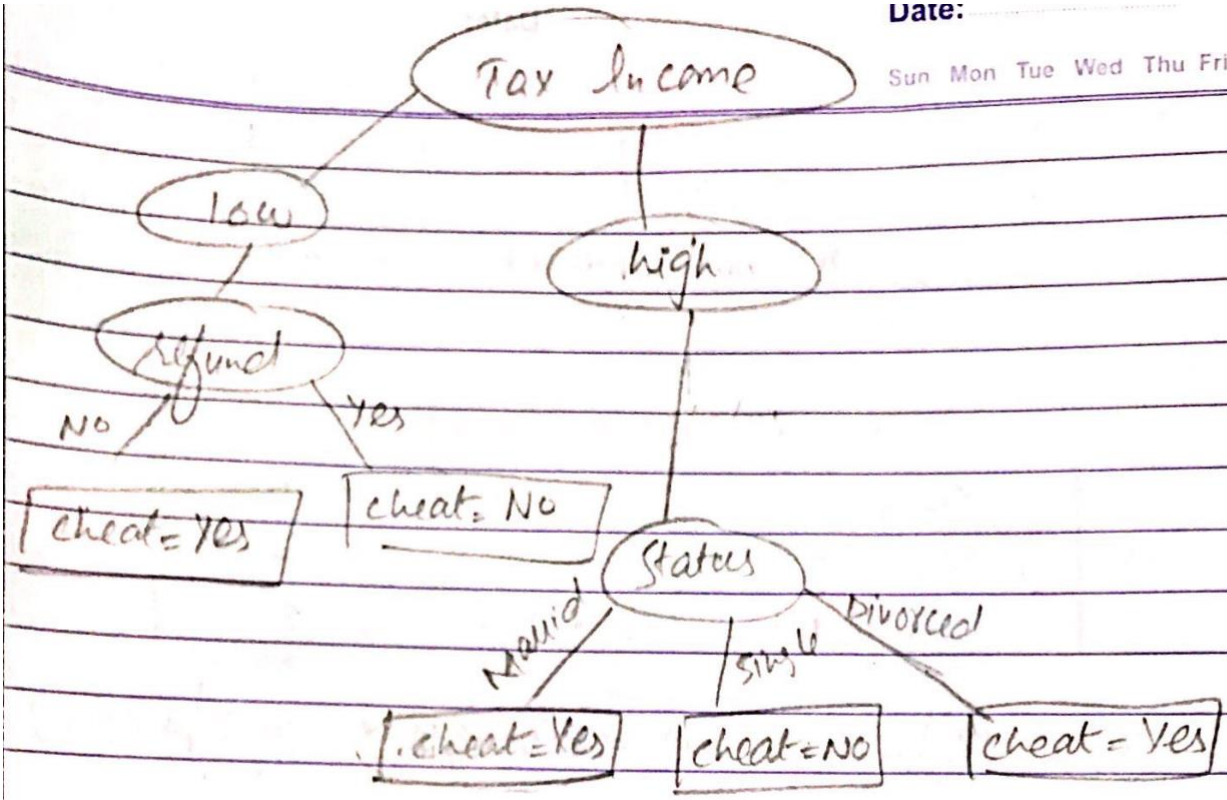
Gain (cheat, refund) = 0.808

Gain (cheat, status) = 1.5

Status will be second node in medium branch due to maximum IG.

Date:

Sun Mon Tue Wed Thu Fri Sat




medium table:-

| refund | Status | cheat |
|---|---|---|
| Yes | Married | NO |
| Yes | Divorced | NO |
| No | Single | NO |
| NO | Married | NO |
| No | Divorced | Yes |
| No | Divorced | Yes |
| | | |

$E(cheat) = \frac{4}{6} \times \log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.923$

(A) refund:

| | | Yes | NO | Total |
|---|---|---|---|---|
| | Yes | 0 | 2 | 2 |
| refund | NO | 2 | 2 | 4 |
| | | | | 6 |

(i) · $E(\text{cheat, refund}) = P(\text{Yes})\,E(0,2) + P(\text{No}) * E(2,2)$

using Prior knowledge from previous iteration

$E(0,2) = 0$

$E(2,2) = 1$

(i) $\Rightarrow E(\text{cheat, refund}) = \frac{2}{6} * 0 + \frac{4}{6} * 1$

$$\boxed{E(\text{cheat, refund}) = 0.66}$$

(iii) Status:-

| | | Yes | No | Total |
|---|---|---|---|---|
| | Single | 0 | 1 | 1 |
| Status | Married | 0 | 2 | 2 |
| | Divorced | 2 | 1 | 3 |
| | | | | 6 |

(ii) $\Rightarrow E(\text{cheat, refund}) = P(\text{single}) * E(0,1) + P(\text{married}) * E(0,2) + P(\text{Divorced}) * E(2,1)$

using prior knowledge from previous iterations

$E(0,2) = E(0,1) = 0$

$E(2,1) = 0.923$

(ii) $\Rightarrow \boxed{E(\text{cheat, refund}) = 0 + 0 + 0.46 = 0.46}$

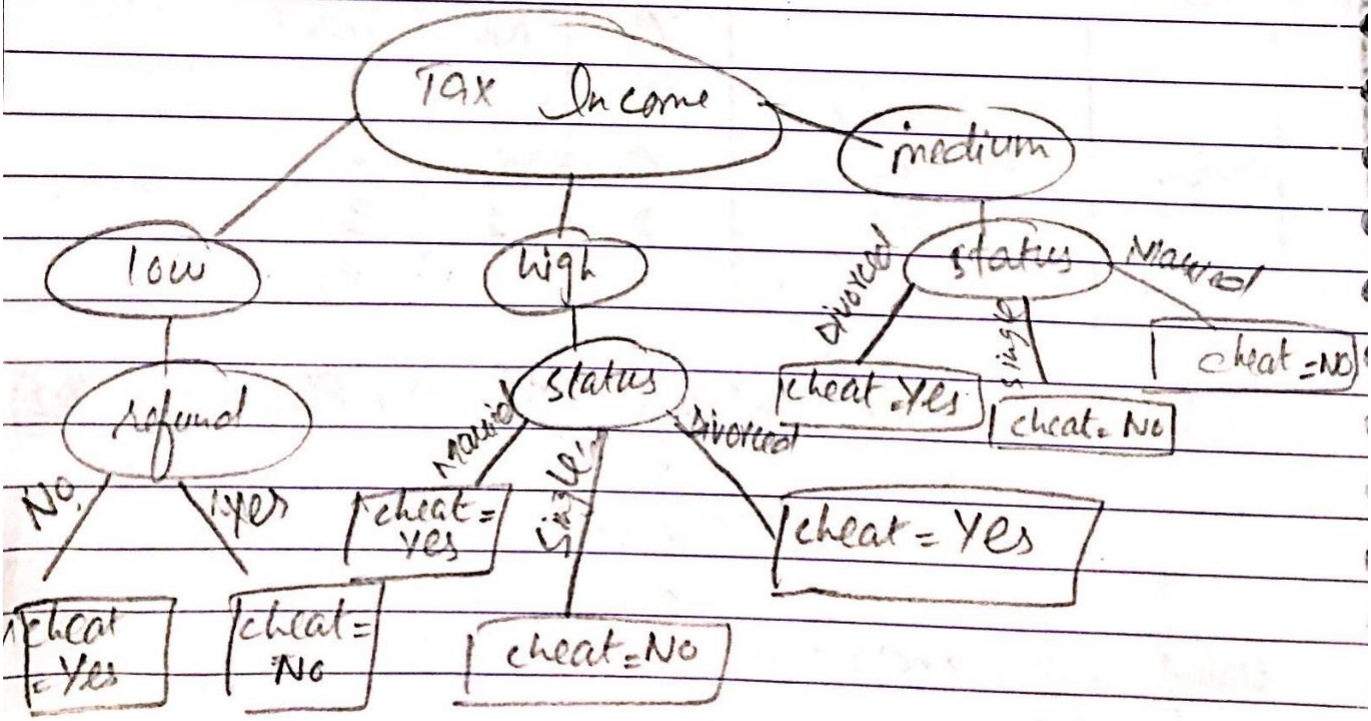
Gain(cheat, refund) = 0.923 - 0.66

Gain(cheat, refund) = 0.263

Gain (cheat, Status) = 0.923 - 0.46

Gain (cheat, Status) = 0.463

Status will be second node in medium branch due to maximum IG.




Due to max_depth = 2 we adjust some records into branches based on maximum occurences.

(2) Calculate:

Sun Mon Tue Wed Thu Fri Sat

$$\text{Gini index (Attribute = value)} = 1 - \sum_{i=1}^{n} (pi)^2$$

$$\text{Gini index (Attribute)} = \sum_{V=\text{value}} P_v \times \text{Gini}(V)$$

1st Iteration:

(A) refund:

| | | Yes | No | Total |
|---|---|---|---|---|
| | Yes | 3 | 4 | 7 |
| refund | No | 4 | 2 | 6 |
| | | | | 13 |
| | | | | |

$$\text{Gini (refund = Yes)} = 1 - \left[\left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right] = 1 - 0.18 - 0.32 = 0.5$$

$$\text{Gini (refund = No)} = 1 - \left[\left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right] = 1 - 0.44 - 0.11 = 0.45$$

$$\text{Gini (refund)} = \frac{7}{13} * (0.5) + \frac{6}{13} * (0.45)$$

$$= 0.26 + 0.20$$

$$\boxed{\text{Gini (Refund)} = 0.46}$$

(B) Status

| | | Yes | NO | Total |
|---|---|---|---|---|
| Status | Single | 2 | 3 | 5 |
| | Married | 3 | 1 | 4 |
| | Divorced | 2 | 2 | 4 |
| | | | | 13 |

$$Gini(Status = Single) = 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right] = 1 - 0.16 - 0.36$$

$$= 0.48$$

$$Gini(Status = Married) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 1 - 0.56 - 0.062$$

$$= 0.378$$

$$Gini(Status = Divorced) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 1 - 0.25 - 0.25$$

$$= 0.5$$

$$Gini(Status) = \frac{5}{13} * 0.48 + \frac{4}{13} * 0.37 + \frac{4}{13} * 0.5$$

$$= 0.18 + 0.113 + 0.153$$

$$\boxed{Gini(Status) = 0.446}$$

(c) Tax Income.

| | | Yes | No | total |
|---|---|---|---|---|
| | low | 2 | 1 | 3 |
| Tax Income | medium | 2 | 4 | 6 |
| | high | 3 | 1 | 4 |
| | | | | 13 |

$$\text{Gini}(\text{Tax Income} = \text{low}) = 1 - \left[\left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] = 1 - 0.44 - 0.11$$

$$= 0.49$$

$$\text{Gini}(\text{Tax Income} = \text{medium}) = 1 - \left[\left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right]$$

$$= 1 - 0.11 - 0.44$$

$$= 0.45$$

$$\text{Gini}(\text{Tax Income} = \text{high}) = 1 - \left[\left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right]$$

$$= 1 - 0.562 - 0.0625$$

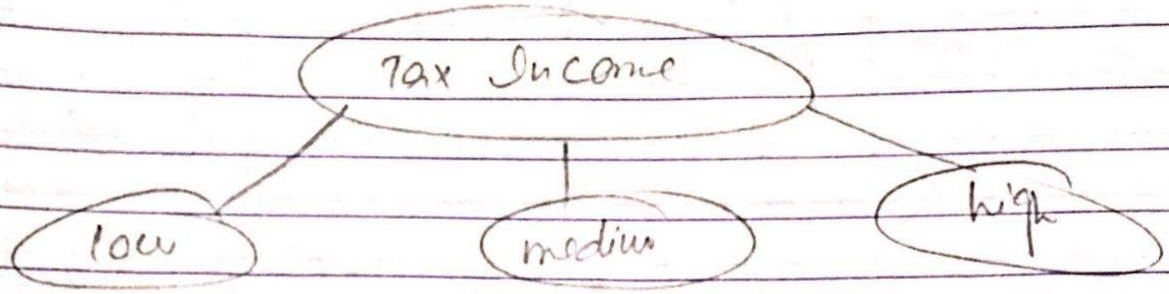
$$= 0.375$$

$$\text{Gini}(\text{Tax Income}) = \frac{3}{13}(0.49) + \frac{6}{13} \times (0.45) + \frac{4}{13} \times (0.375)$$

$$= 0.11 + 0.20 + 0.115$$

$$\text{Gini}(\text{Tax Income}) = 0.425$$

Since Tax Income has lowest gini index it will be root node.




## Second Iteration :-

(1) For low branch :- (using table from previous pages).

(A) Refund.

| | | Yes | No | Total |
|---|---|---|---|---|
| refund | Yes | 1 | 1 | 2 |
| | No | 1 | 0 | 1 |
| | | | | 3 |

$$\text{gini}(\text{refund} = \text{Yes}) = 1 - \left[\left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] = 1 - 0.44 - 0.11$$

$$= 0.78$$

$$\text{gini}(\text{refund} = \text{No}) = 1 - \left[\left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)\right]$$

$$= 0$$

$$\text{gini}(\text{refund}) = \frac{2}{3} * (0.78) + 0$$

$$\boxed{\text{gini}(\text{refund}) = 0.52}$$

B) Status:-

| | | Yes | No | Total. |
|---|---|---|---|---|
| Status | Single | 2 | 1 | 3 |
| | Married | 0 | 0 | 0 |
| | Divorced | 0 | 0 | 0 |
| | | | | 3 |

$$\text{gini}(\text{Status} = \text{Single}) = 1 - \left[\left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

$$= 0.45$$

$$\text{gini}(\text{Status} = \text{Married}) = 1 - \left[\left(\frac{0}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right]$$

$$= 1$$

$$\text{gini}(\text{Status} = \text{Divorced}) = 1 - \left[\left(\frac{0}{3}\right)^2 - \left(\frac{0}{3}\right)^2\right]$$

$$= 1$$

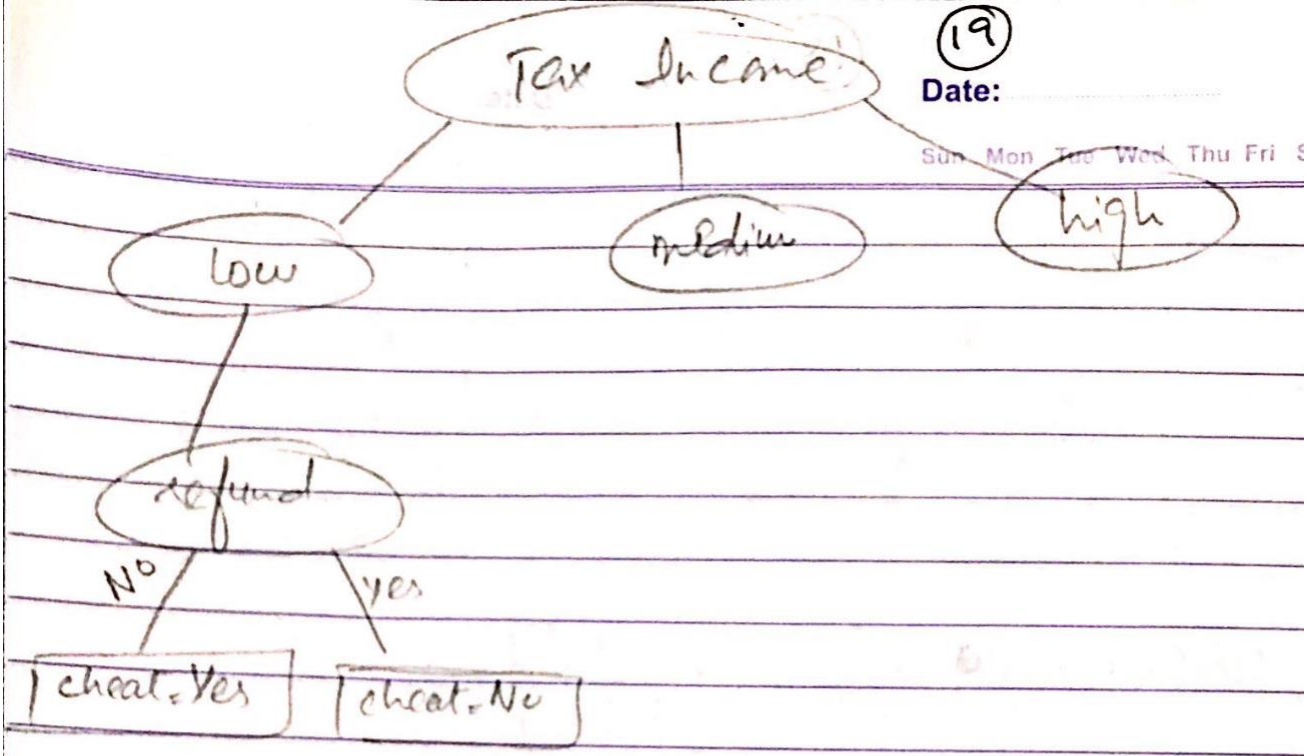
$$\text{gini}(\text{Status}) = \frac{3}{3} * 0.45 + 0 + 0$$

$$\boxed{\text{gini}(\text{Status}) = 0.45}$$

Since refund has lowest gini index it be the second node in low branch.

Date:






For high branch = (using table from previous pages).

(A) Refund

| | | Yes | No | Total |
|---|---|---|---|---|
| Refund | Yes | 2 | 1 | 3 |
| | No | 1 | 0 | 1 |
| | | | | 4 |

$$gini(Refund=Yes) = 1 - \left[\left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

$$= 0.45$$

$$gini(Refund=No) = 1 - \left[\left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right]$$

$$= 1 - 1 = 0$$

$$gini(Refund) = \frac{3}{4} \times 0.45 + 0 = 0.33$$

(B) Status :

| | | Yes | No | Total |
|---|---|---|---|---|
| Status | Single | 0 | 1 | 1 |
| | Married | 2 | 0 | 2 |
| | Divorced | 1 | 0 | 1 |
| | | | | 4 |

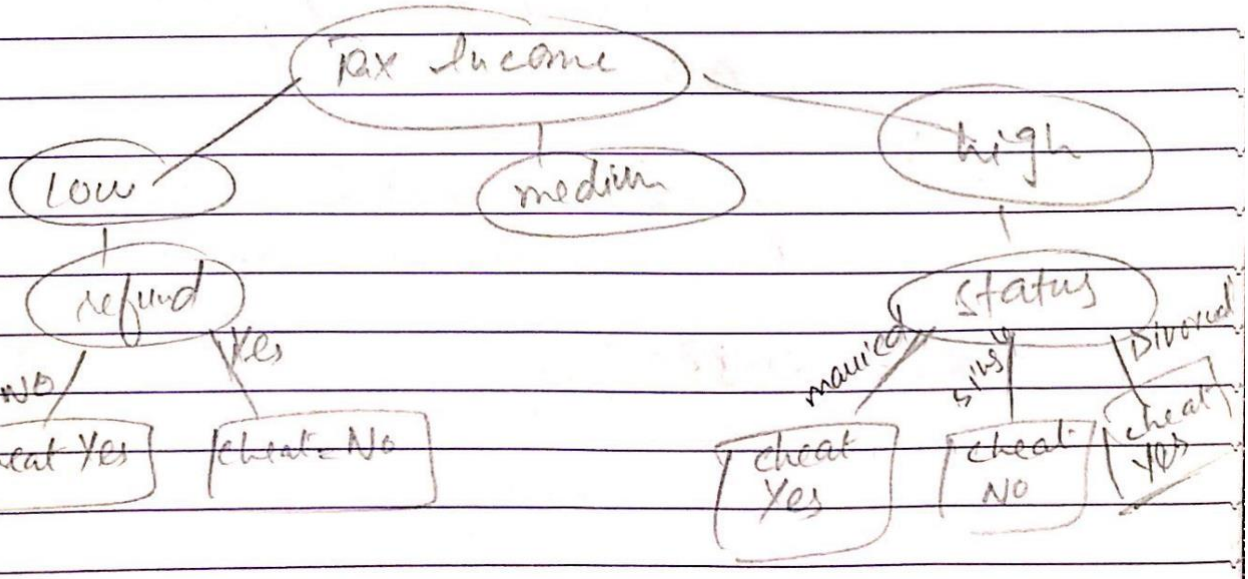
$$\text{gini}(\text{Status} = \text{Single}) = 1 - \left[\left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2\right]$$

$$= 0$$

$$\text{gini}(\text{Status} = \text{Married}) = \text{gini}(\text{Divorced}) = 0$$

$$\text{gini}(\text{Status}) = \frac{1}{4} * 0 + \frac{2}{4} * 0 + \frac{1}{4} * 0 = 0$$

Since Status has lowest gini index it will be Second node in high branch.



For medium branch:-

using table from previous pages.

(A) refund:-

| | | Yes | No | Total |
|---|---|---|---|---|
| refund | Yes | 0 | 2 | 2 |
| | No | 2 | 2 | 4 |
| | | | | 6 |

$$\text{gini}(\text{refund\_Yes}) = 1 - \left[\left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2\right]$$

$$= 0$$

$$\text{gini}(\text{refund} = \text{No}) = 1 - \left[\left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right]$$

$$= 0.5$$

$$\text{gini}(\text{refund}) = \frac{2}{6} * 0 + \frac{4}{6} * 0.5$$

$$\boxed{\text{gini}(\text{refund}) = 0.33}$$

(B) Status:

| | | Yes | No | Total |
|---|---|---|---|---|
| Status | Single | 0 | 1 | 1 |
| | Married | 0 | 2 | 2 |
| | Divorced | 2 | 1 | 3 |
| | | | | 6 |
| | | | | |

gini (status = single) = gini (Status = married) = 0

$$\text{gini}(\text{Status} = \text{Divorced}) = 1 - \left[\left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

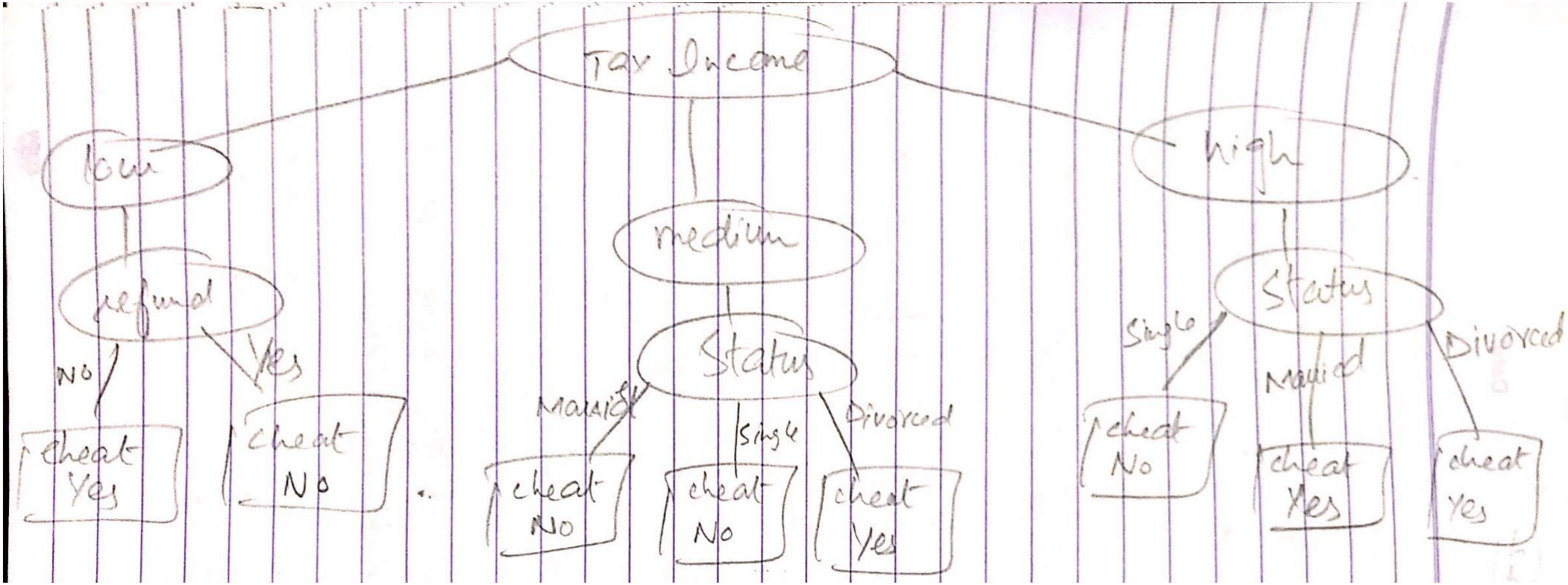
$$= 0.45$$

$$\text{gini}(\text{Status}) = \frac{1}{6} \times 0 + \frac{2}{6} \times 0 + \frac{3}{6} * 0.45$$

$$\boxed{\text{gini}(\text{Status}) = 0.225}$$

Since Status has lowest gini index it be second node in medium branch.

Some instance are adjusted based on maximum occurance criteria due to max depth constraint, can't grow tree further.


Tax Income
low
medium
high
refund
Status
Status
No
Yes
Single
Married
Divorced
Married
Single
Divorced
cheat Yes
cheat No
cheat No
cheat No
cheat Yes
cheat No
cheat Yes
cheat Yes