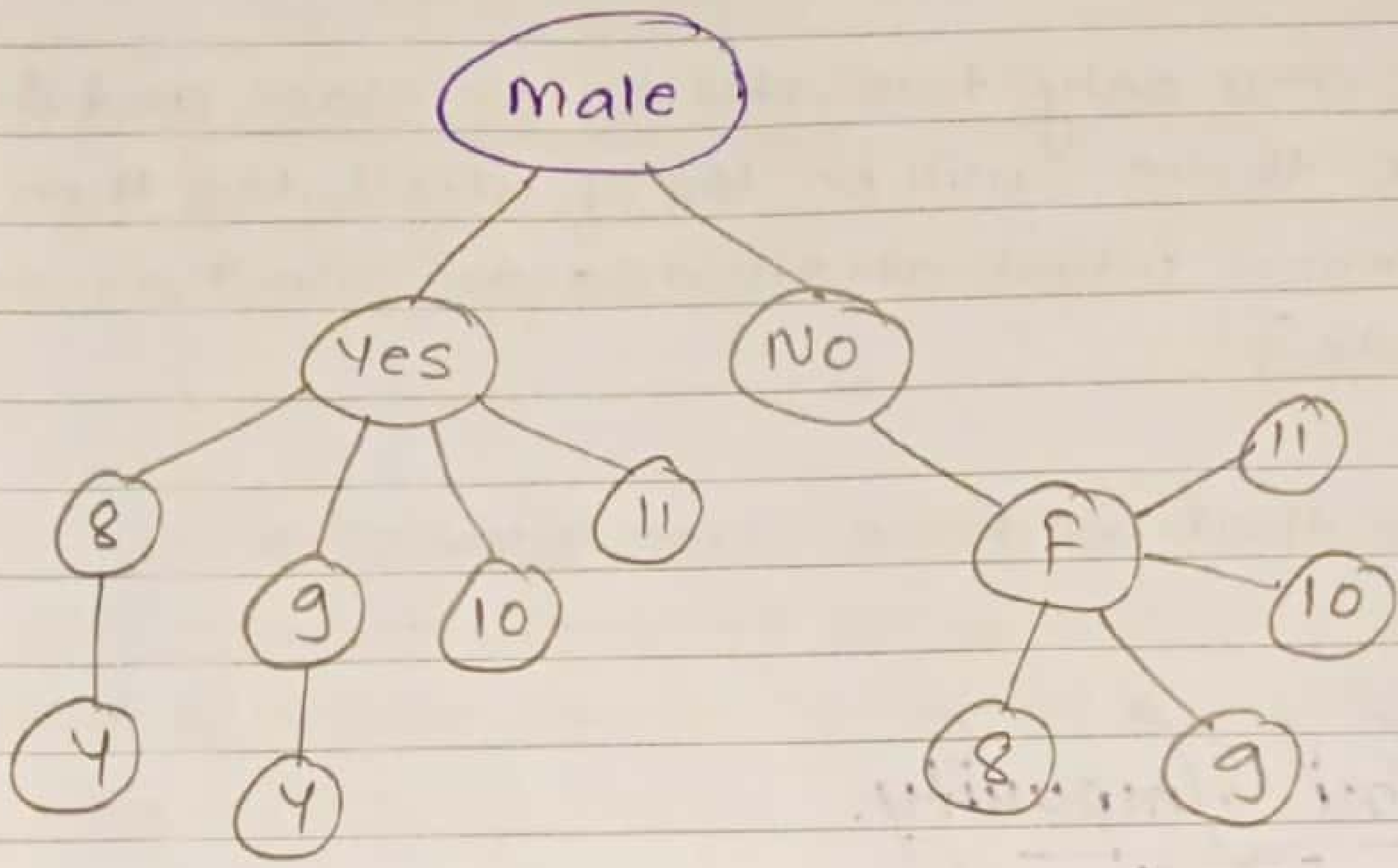
Decision Tree

Example: Table:01

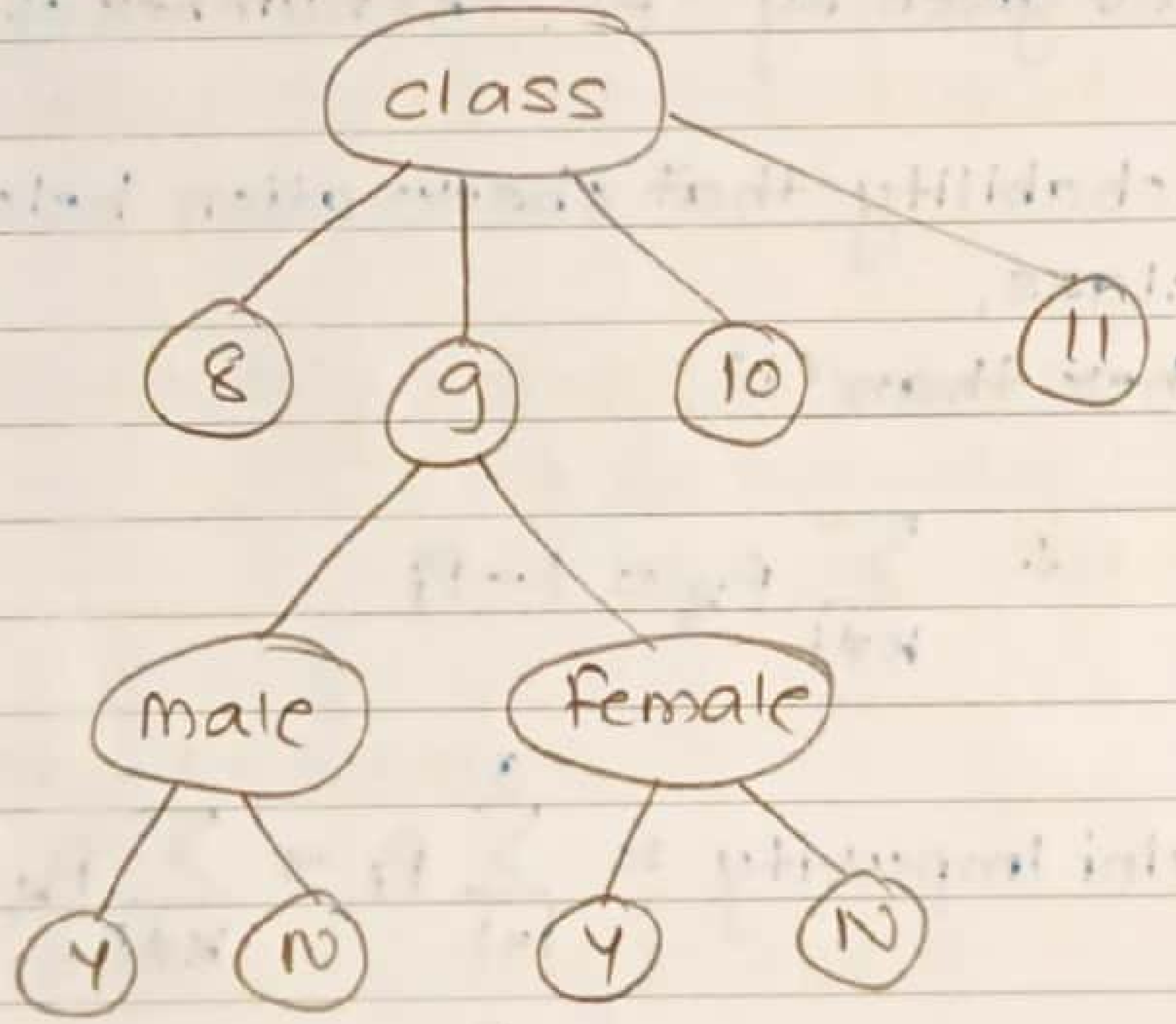
class	Gender	Stay in Hostel.
9	Male(M)	Y
10	Female(F)	N
8	F	Y
8	F	N
9	M	Y
10	M	N
11	F	Y
11	M	Y
8	F	Y
9	M	N
11	M	N
11	M	Y
10	F	N
10	M	Y
8	F	

predict.

If we make tree based on Male column:



Tree based on class column.



So, we can take any attribute as the root node. But which attribute (column) we have to take as root node?

Here, are only two attributes class and Gender. Suppose there will be 100 of attributes then how to know which attributes to select as root node?

For that we have some terms :→

Gini Impurity.

let the observation belongs to class 'i' and its probability be given by 'P_i'. k = number of classes

then, probability that observation belongs to any other class, other than 'i',

$$\therefore \sum_{k \neq i} P_k = 1 - P_i$$

then

$$\text{Gini Impurity} = \sum_{i=1}^{J} P_i \times \sum_{k \neq i} P_k$$

$$= \sum_{i=1}^{J} P_i(1 - P_i)$$

$$= \sum_{i=1}^{J} (P_i - P_i^2)$$

$$= \sum_{i=1}^{J} P_i - \sum_{i=1}^{J} P_i^2$$

$$= 1 - \sum_{i=1}^{J} P_i^2$$

$$\therefore \boxed{\text{Gini Impurity} = 1 - \sum_{i=1}^{J} P_i^2}$$

Gini Impurity is a measure of how often a randomly choosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.

It is calculated by multiplying the probability that a given observation is classified into the correct class and sum of all the probabilities when that particular observation is classified into the wrong class.

Gini Impurity value lies between 0 and 1,

0 → no impurity

1 → random distribution

according to table 01,

class	Stay in Hostel	Total value
8	Y=2, N=1	3
9	Y=2, N=1	3
10	Y=1, N=3	4
11	Y=3, N=1	4
		14

P(Y): Probability of person staying in class 8 and who is living in hostel.

P(N): Probability of person in class 8 and who is not staying in hostel.

Class	Stay in Hostel	Total value	P(Y)	P(N)
8	Y=2, N=1	3	2/3	1/3
9	Y= 2, N=1	3	2/3	1/3
10	Y=1, N=3	4	1/4	3/4
11	Y=3, N=1	4	3/4	1/4

Now, we will calculate Gini Impurity for each and individual classes

$$Gini = 1 - \sum_{i=1}^{J} (P_i)^2$$

$Gini(8) = 1 - P(Y)^2 - P(N)^2 = 1 - (2/3)^2 - (1/3)^2$
$= 4/9.$

$Gini(9) = 1 - (2/3)^2 - (1/3)^2 = 1 - 4/9 - 1/9 = \underline{4/9}$

$Gini(10) = 1 - (1/4)^2 - (3/4)^2 = 1 - 1/16 - 9/16 = \underline{3/8}$

$Gini(11) = 1 - (3/4)^2 - (1/4)^2 = 3/8$

Now, Gini of entire class column will be

$$\text{Gini(class)} = \sum_{i=1}^{n} \frac{\text{no. of instances for class}}{\text{Total no. of instance}} \times \text{Gini (c)}$$

$$\text{Gini(entire class)} = \frac{n \cdot 8}{T} \cdot G(8) + \frac{n9}{T} \cdot G(9) + \frac{n10}{T} G(10) + \frac{n11}{T} G(11)$$

$$= \frac{3}{14} \cdot \frac{2}{3} + \frac{3}{14} \cdot \frac{4}{9} + \frac{4}{14} \cdot \frac{3}{8} + \frac{4}{14} \cdot \frac{3}{8}$$

$$= 0.66 + 0.44 + 0.375 + 0.375$$

$$= 0.404$$

This whole calculation is for class column only. Now we have to calculate gini for gender column

Gender	Stay in Hostel	Total value	P(Y)	P(N)
Male	Y=5, N=3	8	5/8	3/8
Female	Y=3, N=3	6	3/6	3/6
		14		

$$\text{Gini(male)} = 1 - (5/8)^2 - (3/8)^2 = 1 - \frac{25}{64} - \frac{9}{64}$$

$$= 0.468$$

$$\text{Gini(Female)} = 1 - \left(3/6\right)^2 - \left(3/6\right)^2 = 0.5$$

$$\therefore \text{Gini(Gender column)} = \frac{8}{14} \times 0.468 + \frac{6}{14} \times 0.5$$

$$= 0.4817$$

$$\text{Gini(class column)} = 0.404$$

$\therefore$ out of Gender column and class column Gini of gender column is more.

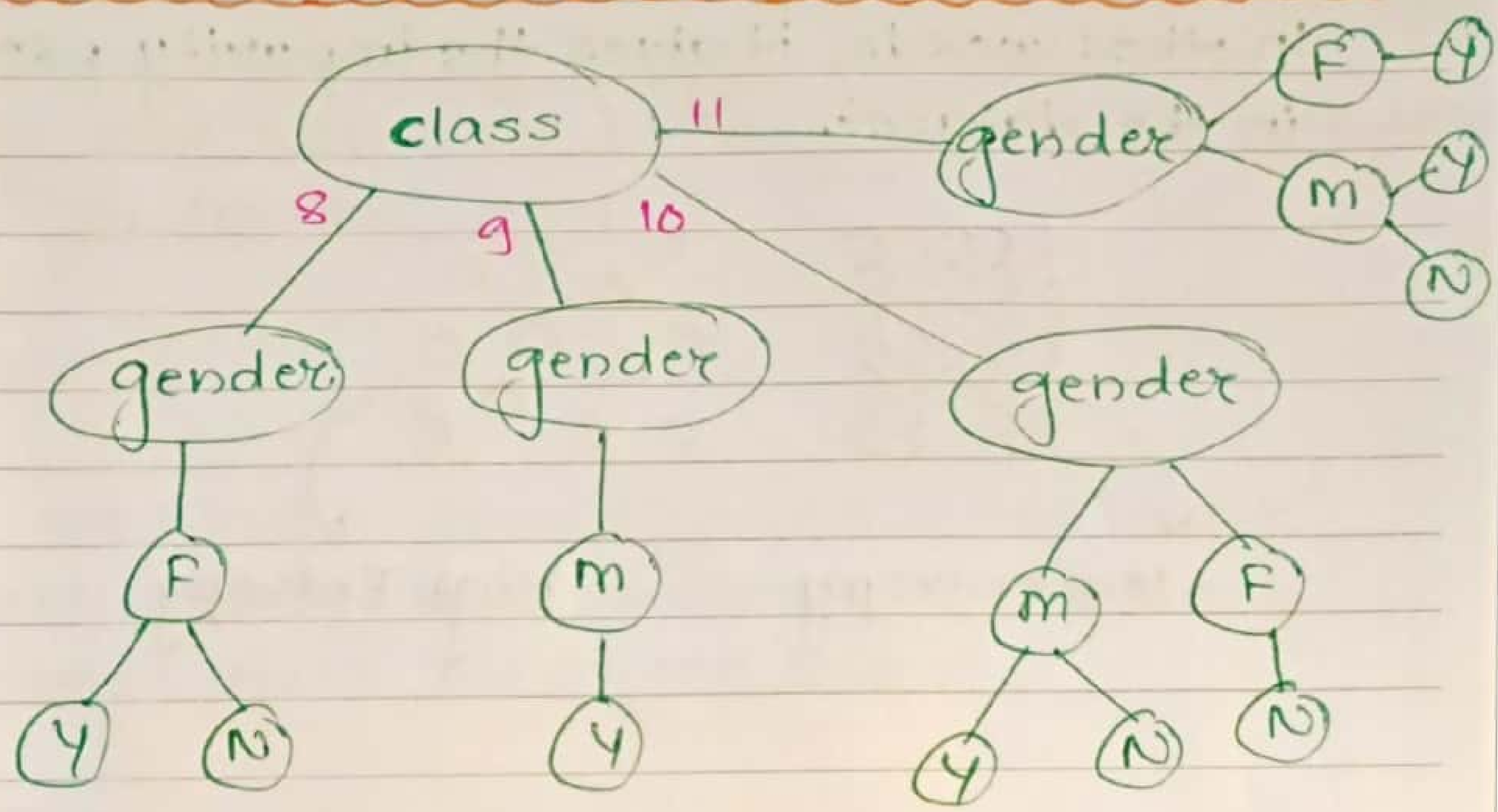
(Gini → Gini Impurity)

Here, Gini Impurity (Gender column) is more compared to Gini Impurity (class column). So, we have to take class column as the root node or parent node.

Note:
The node for which the Gini Impurity is least is selected as the root node to split.

Suppose we have 100 columns, we will calculate gini of every column and use that column as our root node which has less gini impurities.

Gini Impurity is an approach which works with categorical data. not continuous data.



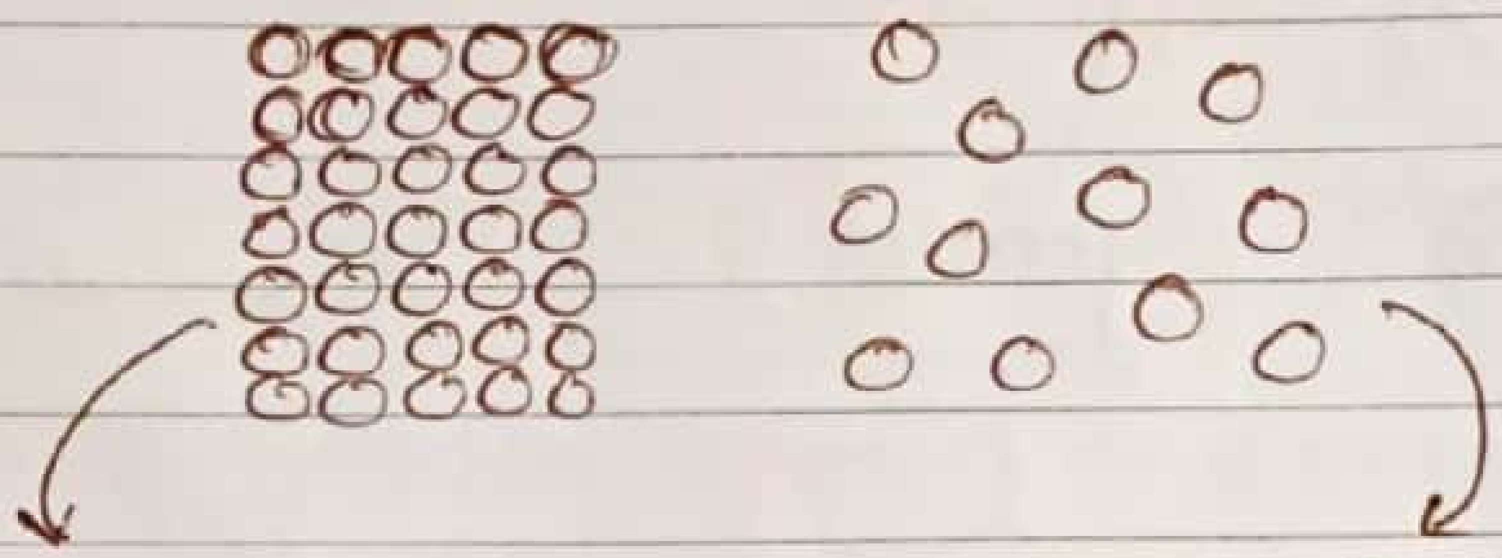
- If we have to predict that in class 11 there is a female, is she going to stay in hostel or not
  So, by above tree diagram we can predict that Yes.

- If a new student in class11 is male, then whether he is going to stay in hostel or not?
  → Yes, because most of time male students are staying in hostel of class11.

Entropy

Entropy is the measure of randomness in data.

In other words, it gives the impurity present in the dataset.



low entropy High Entropy

$$\text{Entropy}(E) = -\sum_{i=1}^{n} P.\log_2(P)$$

Information Gain.

Information Gain calculates the decrease in entropy after splitting a node.

It is a difference between entropies before and after the split.

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

E = Entropy

The more the Information gain, the more entropy is removed.

Based on dataset of Table 01 we calculate Entropy and based on Entropy we try to calculate Information Gain.

Explanation:-

we have 14 records and 2 class in our dataset. out of these 14 records, how many instances gives Yes as the output.

14 records : $n(Y) = 8$, $n(N) = 6$

calculating Entropy for Yes and No,

$$Entropy(L) = -P(Y) \cdot \log_2 P(Y) - P(N) \cdot \log_2 P(N)$$

$$= -\frac{8}{14} \cdot \log_2 \frac{8}{14} - \frac{6}{14} \cdot \log_2 \frac{6}{14}$$

$$= 0.9852$$

$$\therefore \quad E(L) = 0.985$$

Here we are able to calculate entropy of label column.

calculating Entropy for class column and gender column.

$$E(8) = -P(Y)\cdot\log_2 P(Y) - P(N)\cdot\log_2 P(N)$$
$$= -\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\cdot\log_2\left(1/3\right)$$
$$= \boxed{0.9182}$$

$$E(9) = -\frac{2}{3}\cdot\log_2\frac{2}{3} - \frac{1}{3}\cdot\log_2\frac{1}{3}$$
$$= \boxed{0.9182}$$

$$E(10) = -\frac{1}{4}\cdot\log_2\frac{1}{4} - \frac{3}{4}\cdot\log_2\frac{3}{4}$$
$$= \boxed{0.811}$$

$$E(11) = -\frac{3}{4}\cdot\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4}$$
$$= \boxed{0.811}$$

Information Gain from class column:

$$I(\text{class}) = \frac{\text{Total records of class 8}}{\text{Total no. of records}}\left[\text{Entropy of class 8}\right]$$

$+$

$$\frac{\text{Total records of class 9}}{\text{Total records of no. 9}} \left[\text{Entropy of class } 9\right]$$

$$+$$

$$\vdots$$

$$\therefore \quad I(\text{class}) = \left(\frac{3}{14} \cdot 0.918\right) + \left(\frac{3}{14} \times 0.918\right) + \left(\frac{4}{14} \times 0.811\right) + \left(\frac{4}{14} \times 0.811\right)$$

$I(\text{class}) = 0.8574$: Total Information Gain of class

Now,

Information Gain (IG) = $E_{Before} - E_{After}$

where,

E_{before} = E(Label column)

E_{after} = E(Label column)

$$\therefore \quad IG = 0.9852 - 0.8574$$
$$= 0.1278$$

0.1278 will be the total Information Gain means this is the total difference between the entropy of label column and entropy of class column.

Information Gain for Gender column:

$$\text{Entropy}(m) = -P(Y)\cdot\log_2 P(Y) - P(N)\cdot\log_2 P(N)$$
$$= -\frac{3}{8}\cdot\log_2\frac{3}{8} - \frac{5}{8}\cdot\log_2\frac{5}{8}$$
$$= 0.9544$$

$$\text{Entropy}(F) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\cdot\log_2\frac{3}{6}$$
$$= 1$$

$$\therefore \text{Entropy}(\text{Gender}) = \frac{8}{14}\times 0.9544 + \frac{6}{14}\times 1$$
$$= 0.9739$$

Information Gain (IG) = $E_{before} - E_{after}$

For Gender column,

$$= E(L) - E(G)$$
$$= 0.9852 - 0.9739$$
$$= \cancel{0.0127}\ 0.01127$$

$\therefore$ Information Gain(Gender) = 0.01127

Information Gain (class) = 0.1278

case 01:

Feature can be categorical } classification
Outcome can be categorical } problem

case 02:

Feature can be continuous } classification
outcome can be categorical } problem

case 03:

Feature can be continuous } Regression
outcome can be continuous } problem

Example:

x_1	x_2	x_3	x_4
1.1	7.5	2.5	A
2.2	8.8	5.5	B
3	9.2	6	A
3.6	5.1	6.7	A
5	5.4	7	B
5.8	2	8.9	B
8	1	9.1	A

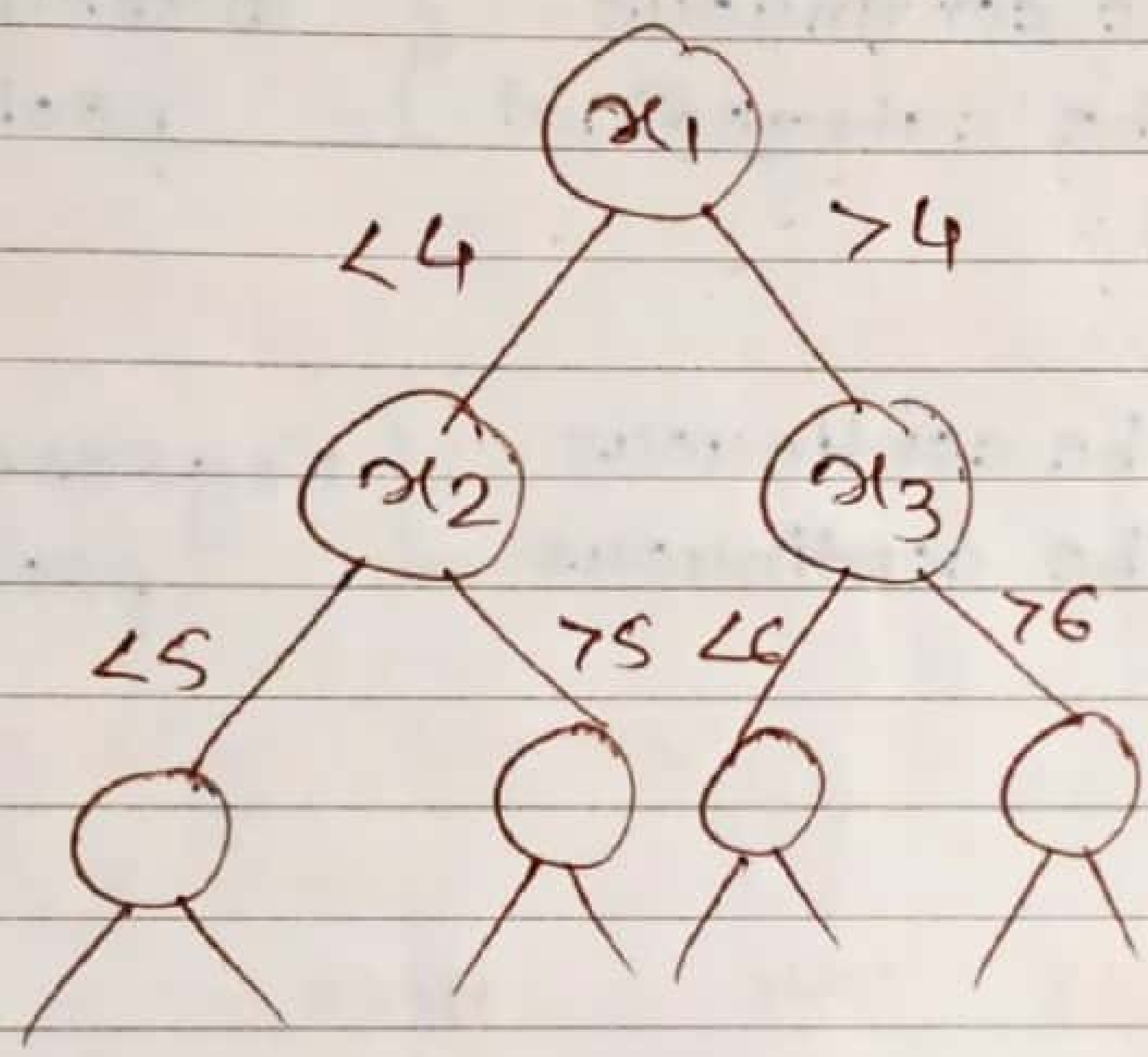
Here,
Feature → continuous
outcome → categorical

How to select the node in this case?

The concept here is we have to create a threshold }.

Suppose we take threshold for $\boxed{x_1 = 4}$ ($\approx$ average)

Similarly, for $\boxed{x_2 = 5}$, $\boxed{x_3 = 6.}$



we will divide until we get the leaf node.

Another dataset :-

This is Regression Problem.

x_1	x_2	y
1	10	5
2	20	10
3	30	15
4	40	20
5	50	25
6	60	30

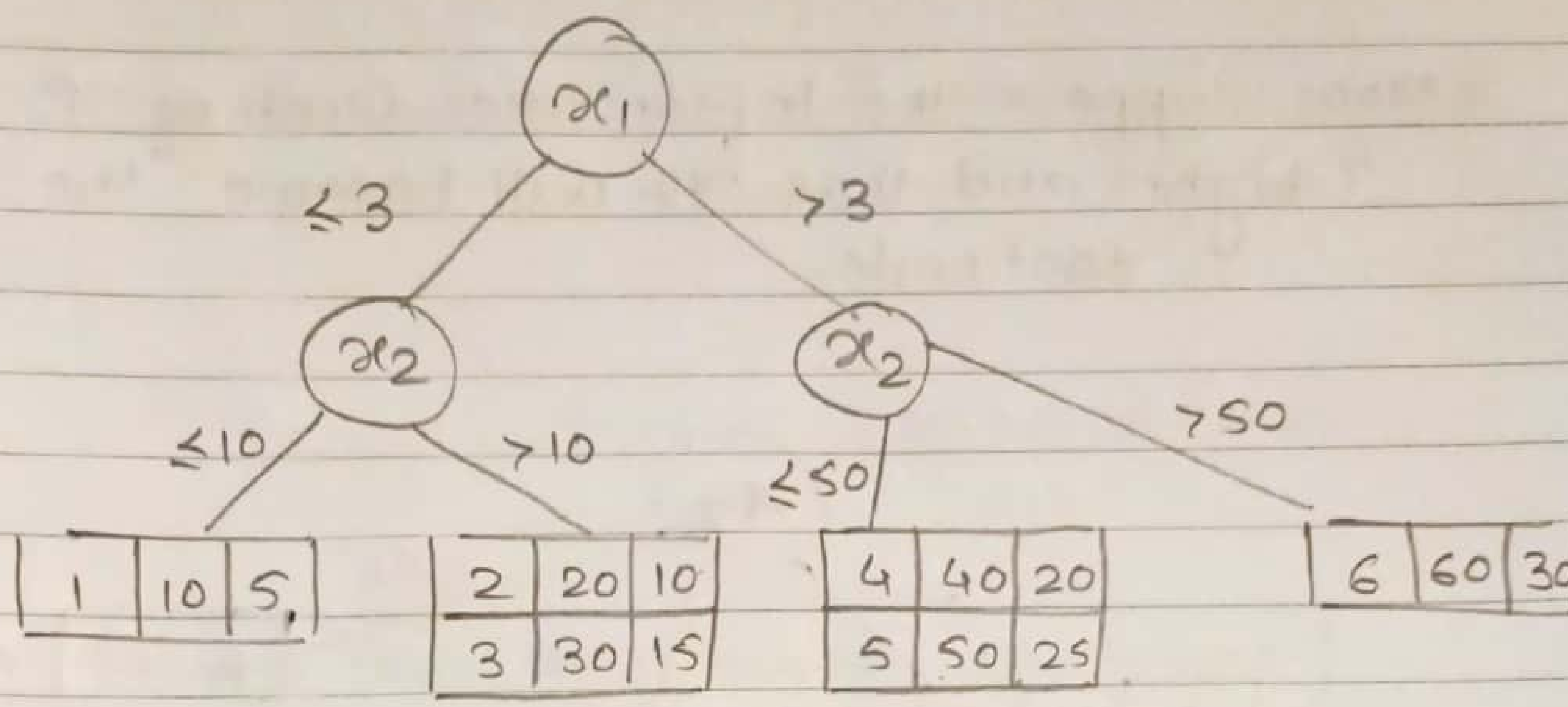
Suppose the Information Gain x_1 is high.

so, x_1 will become root node.

x_1 threshold := 3

x_2 threshold := 30

Scanned with CamScanner



Now, calculate $y - \hat{y}$

y : Actual value
$\hat{y}$: Predicted value

For

1	10	5

$y = 5 \quad \hat{y} = 5$

For

2	20	10
3	30	15

$y = 10 \quad \hat{y} = \left(\frac{10+15}{2}\right) = \underline{12.5}$

$y = 15 \quad \hat{y} = 12.5$

For

4	40	20
5	50	25

$y = 20 \quad \hat{y} = \frac{20+25}{2} = 22.5$

$y = 25, \; \hat{y} = 22.5$

For

6	60	30

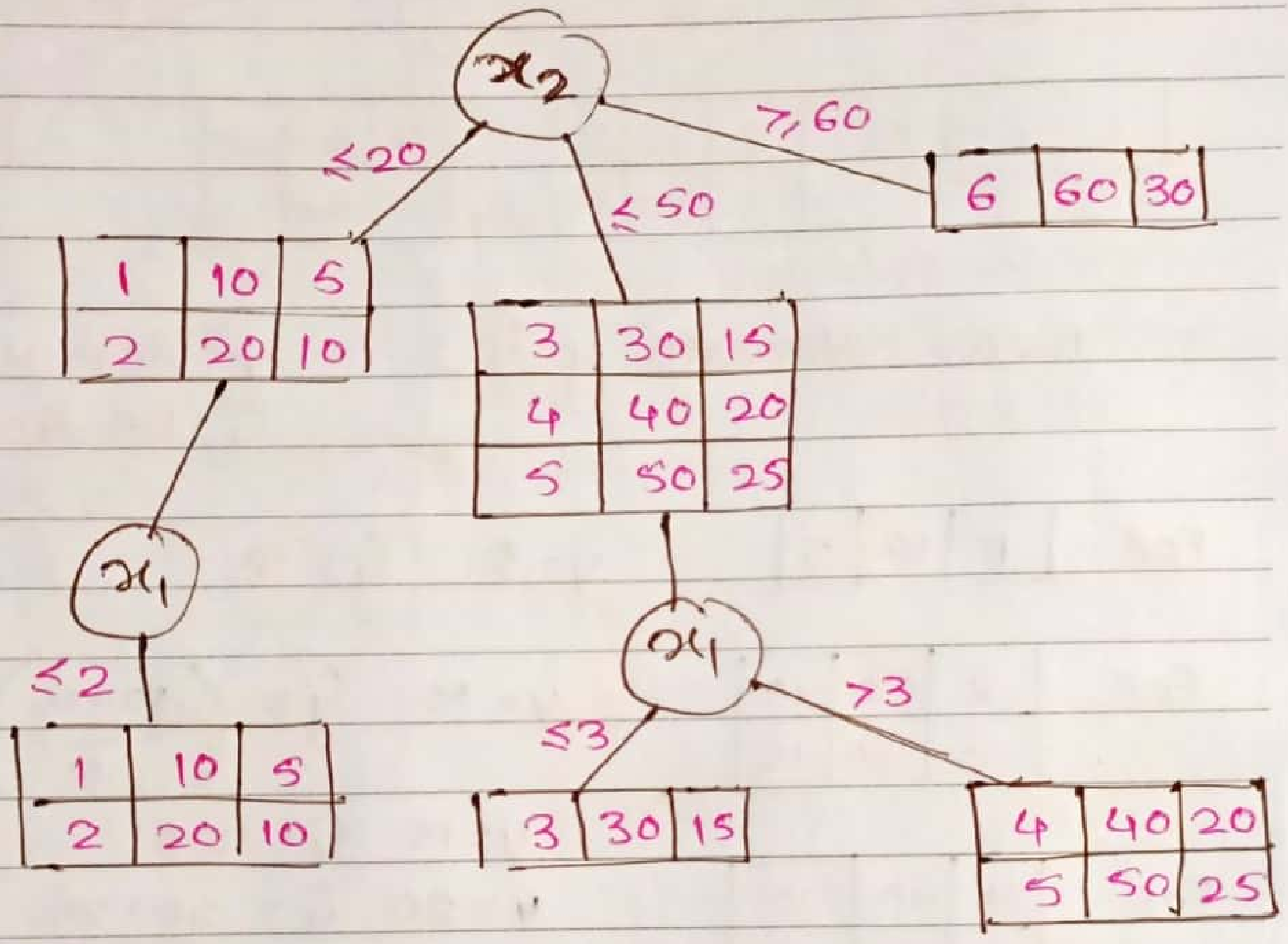
$y = 30, \; \hat{y} = 30$

Total Residual $= \Sigma (y - \hat{y})^2$

$$\therefore = (5-5)^2 + (10-12.5)^2 + (15-12.5)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$$

Scanned with CamScanner

Now, suppose the Information Gain of x_2 is high. and thus x_2 will become the root node.



Now, test the tree $(y-\hat{y})^2$

For

1	10	5
2	20	10

$y=5 \quad \hat{y} = 7.5 \quad \frac{10+5}{2}$

$y=10 \quad \hat{y} = 7.5$

For

3	30	15

$y=15, \ \hat{y}=15$

For

4	40	20
5	50	25

$y= 20, \ \hat{y} = 22.5$

$y = 25, \ \hat{y} = 22.5$

$$\text{Total Residuals} = (5-7.5)^2 + (10-7.5)^2 + (15-15)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$$

$$= 6.25 + 6.25 + 6.25 + 6.25$$

$$= \underline{\underline{25}}$$

Note: Take the situation which has maximum residual.

Tree Pruning

Tree pruning is the method of trimming down a full tree to reduce the complexity and variance in the data.

Just as we regularised linear regression, we can also regularise the decision tree model by adding a new term.

non-overlapping regions

subtree

$$\boxed{\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|}$$

non-negative tuning parameter

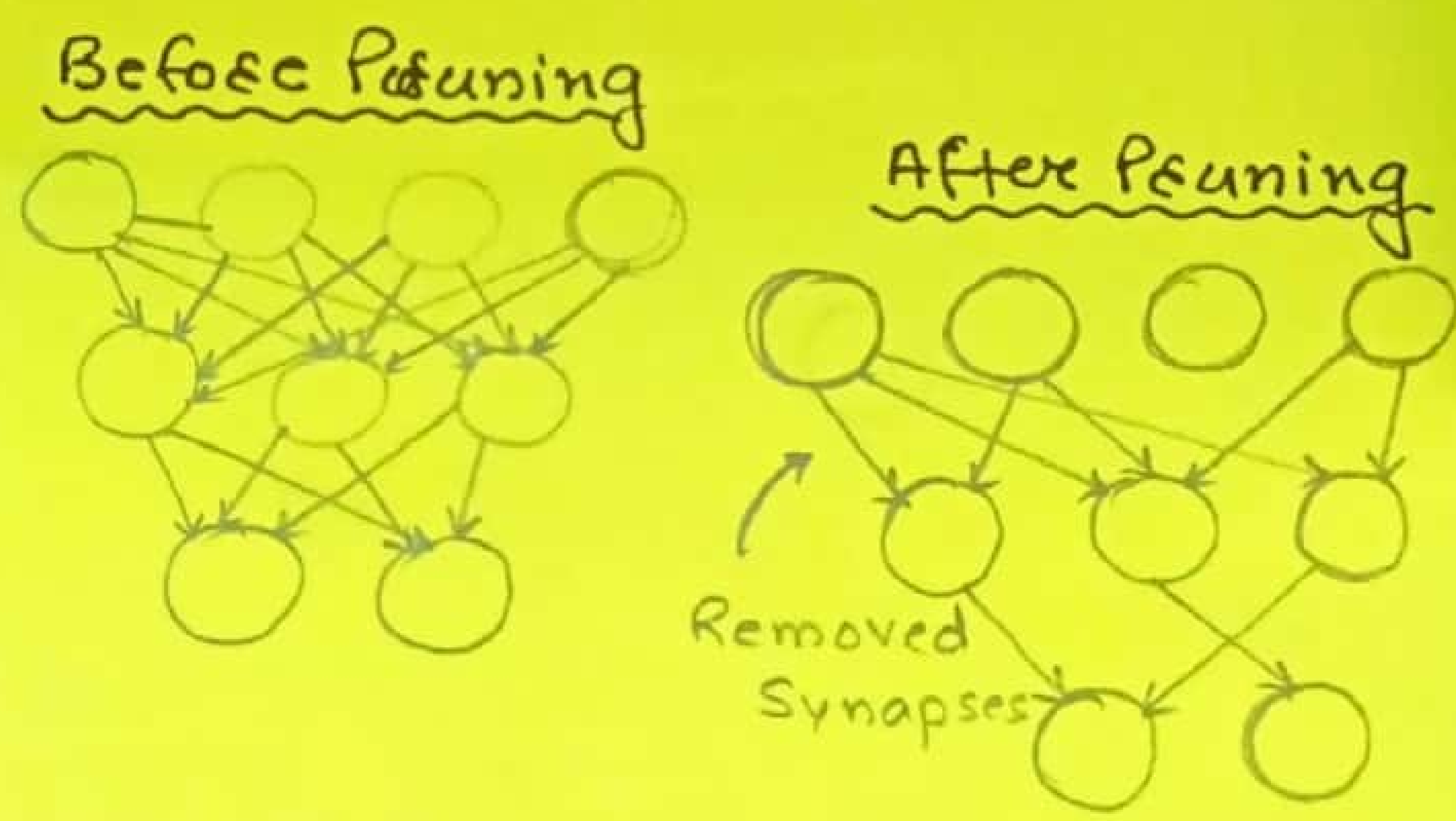
$$\text{Total Residuals} = (5-7.5)^2 + (10-7.5)^2 + (15-15)^2 + (20-22.5)^2 + (25-22.5)^2 + (30-30)^2$$

$$= 6.25 + 6.25 + 6.25 + 6.25$$

$$= \underline{\underline{25}}$$

Note: Take the situation which has maximum residual.

Tree Pru



* also we can remove neuron.

· Tree prun ... n.
a full tree to ... e
in the da

Just as we ...
also requ
adding a new term.

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

non-overlapping regions

subtree

non-negative tuning parameter

where, y_{R_m} → mean of all the response variable in the region 'm'

where,

T = subtree which is a subset of the full tree To

α = non-negative tuning parameter which penalises the MSE with an increase in tree length.

By using cross-validation such values of α and T are selected for which our model gives the lowest test error rate.

This is how the decision tree regression model works.

Post Pruning:

also known as backward pruning.

- Is the process where the decision tree is generated first and then the non-significant branches are removed.

- Cross validation set of data is used to check the effect of pruning and test whether expanding a node will make an improvement or not.

- If any improvement is there then we continue by expanding that node else if there is reduction in accuracy then the node not be expanded and should be converted into leaf node.

- This technique is used when decision tree will have very large depth and will show overfitting of model.

Pre-Pruning: also known as forward pruning.

- stops the non-significant branches from generating.

- It uses a condition to decide when should it terminate splitting of some of the branches prematurely as the tree is generated.

- can be done using Hyper-parameter tuning.
- overcome the overfitting issue.

Scanned with CamScanner

Different Algorithms for Decision Tree.

· ID3: (Iterative Dichotomiser)

- used to construct decision tree for classification.

- It uses Information Gain as the criteria for finding the root nodes and splitting them.

- It only accepts categorical attributes.

· C4.5

- It is an extension of ID3 algorithm, and better than ID3 as it deals both continuous and discrete values.

- It is also used for classification purpose.

· CART (Classification and Regression Algorithm):

· It uses gini impurity as the default calculation for selecting root notes however one can use "entropy" for criteria as well.

- It works on both regression as well as classification problems.

Entropy and Gini Impurity can be used reversibly. It does'n't affects the result much.

Although, gini is easier to compute than entropy. since entropy has long term calculation. That's why CART algorithm uses gini as the default algorithm.

- CAID: Chi-Square Automatic Interaction Detection

- It finds out the statistical significance between the differences between sub-nodes and parent node.

- we measure it by the sum of squares of standardized differences between observed and expected frequencies of the target variable.

- It works with categorical target variable.

Questions:

- How to decide a threshold?

we look after the custom mechanism that is being designed, internally in all these algorithms.

ID3, C45, CART, CAID.

Internally, we find out that sometimes we take average and based on that we try to divide or make different bins / blocks.

These are some ways to create threshold.

- $$\boxed{\text{Information Gain} \propto \frac{1}{\text{Gini Impurity}}}$$