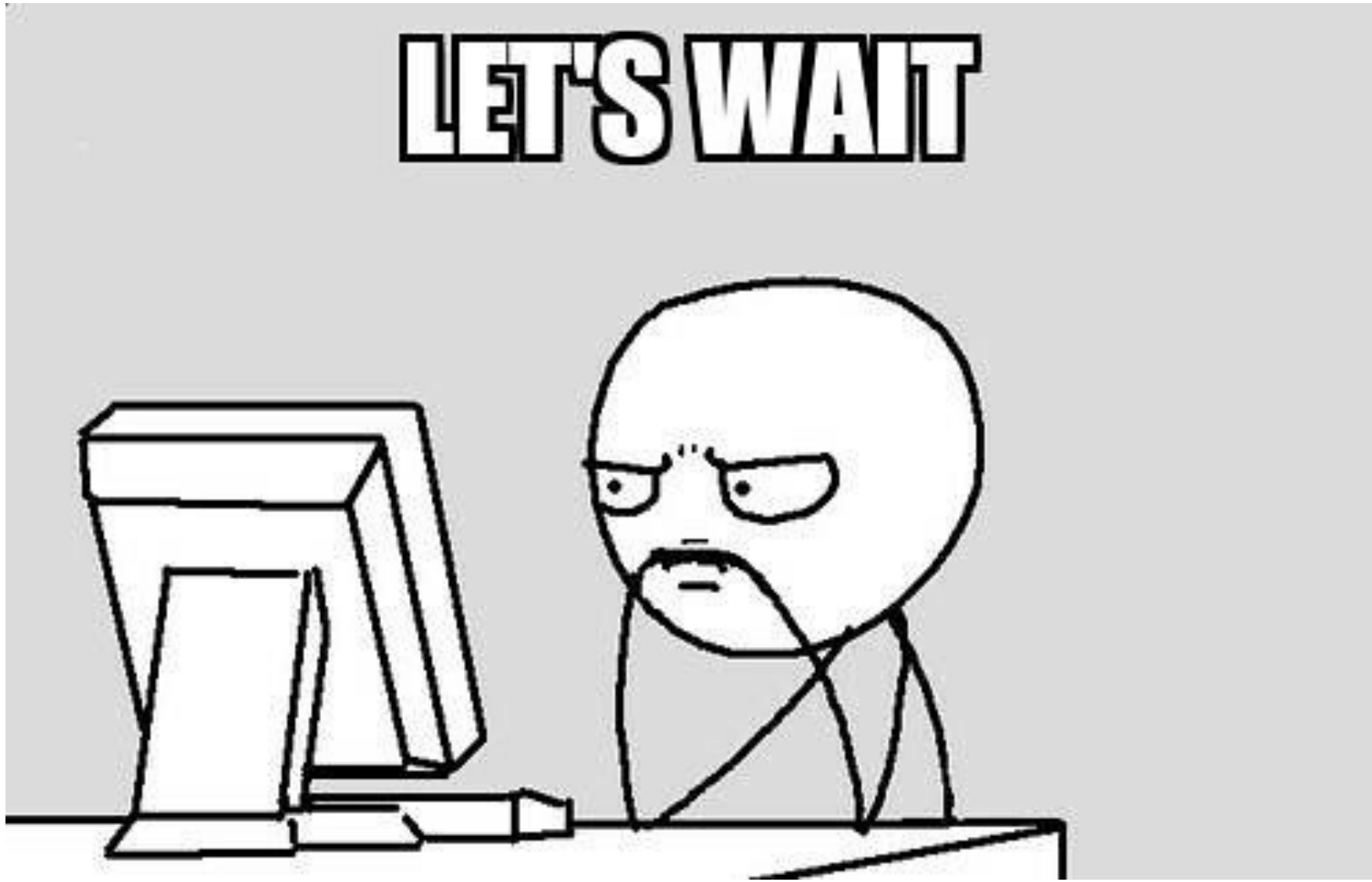

LET'S WAIT

#LifeKoKaroLift

Data Science Certification Program

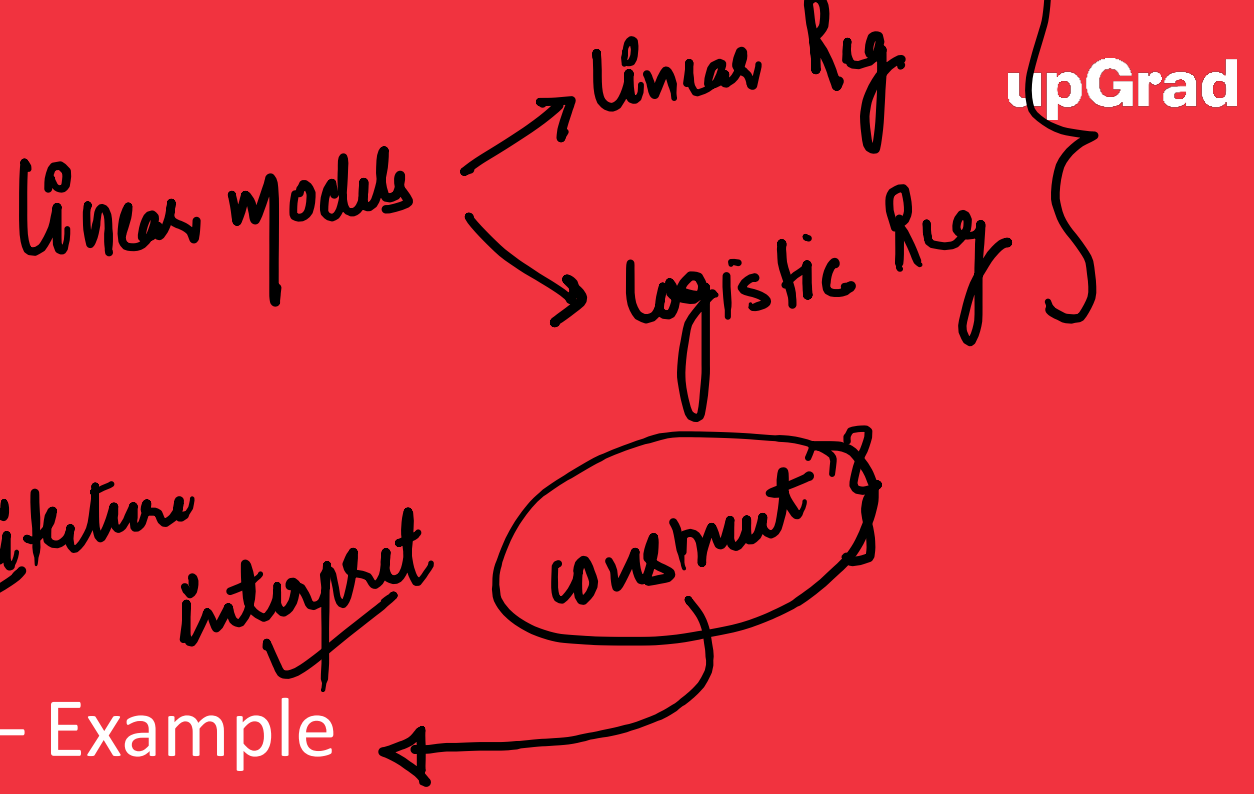
upGrad

Course : Machine Learning

Lecture On : Tree Models

Instructor : Shivam Garg

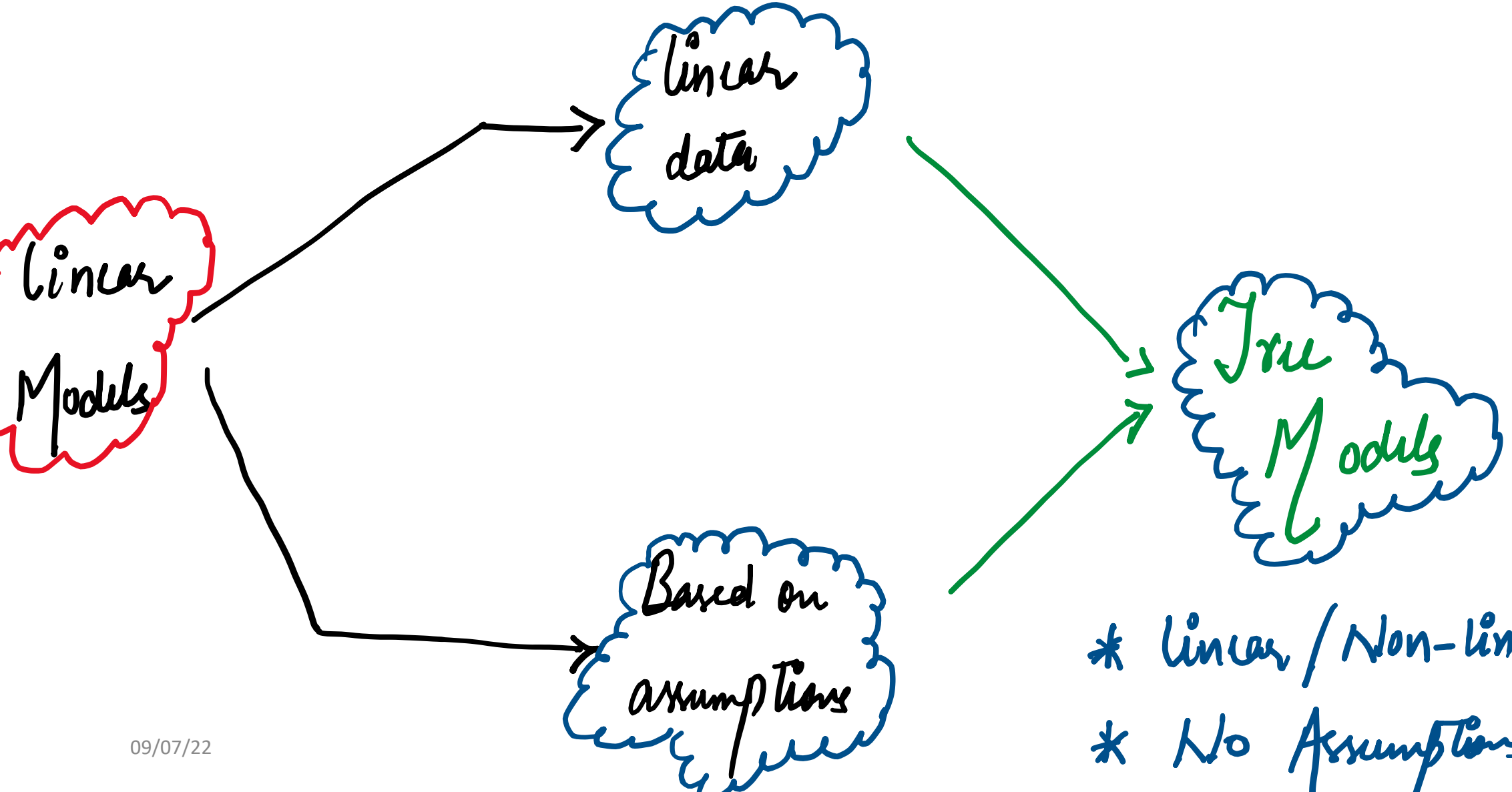
AGENDA



- Why Tree Models?
- Decision Trees
- Gini Index Calculation – Example
- Doubt Session

Why Tree Models?:

upGrad

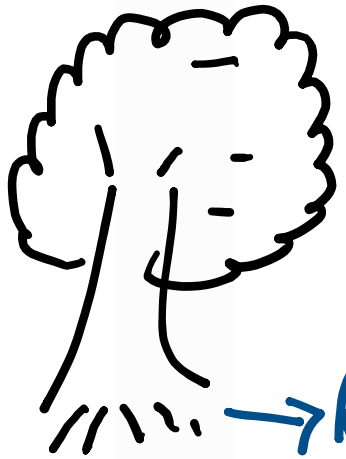


Linear Model	Tree Models	Neural Networks
Linear data	Linear/ Non-linear data	Non-linear
* High interpretable	* partially interpretable	* No interpretability
* less computation	* More computation	* High computation

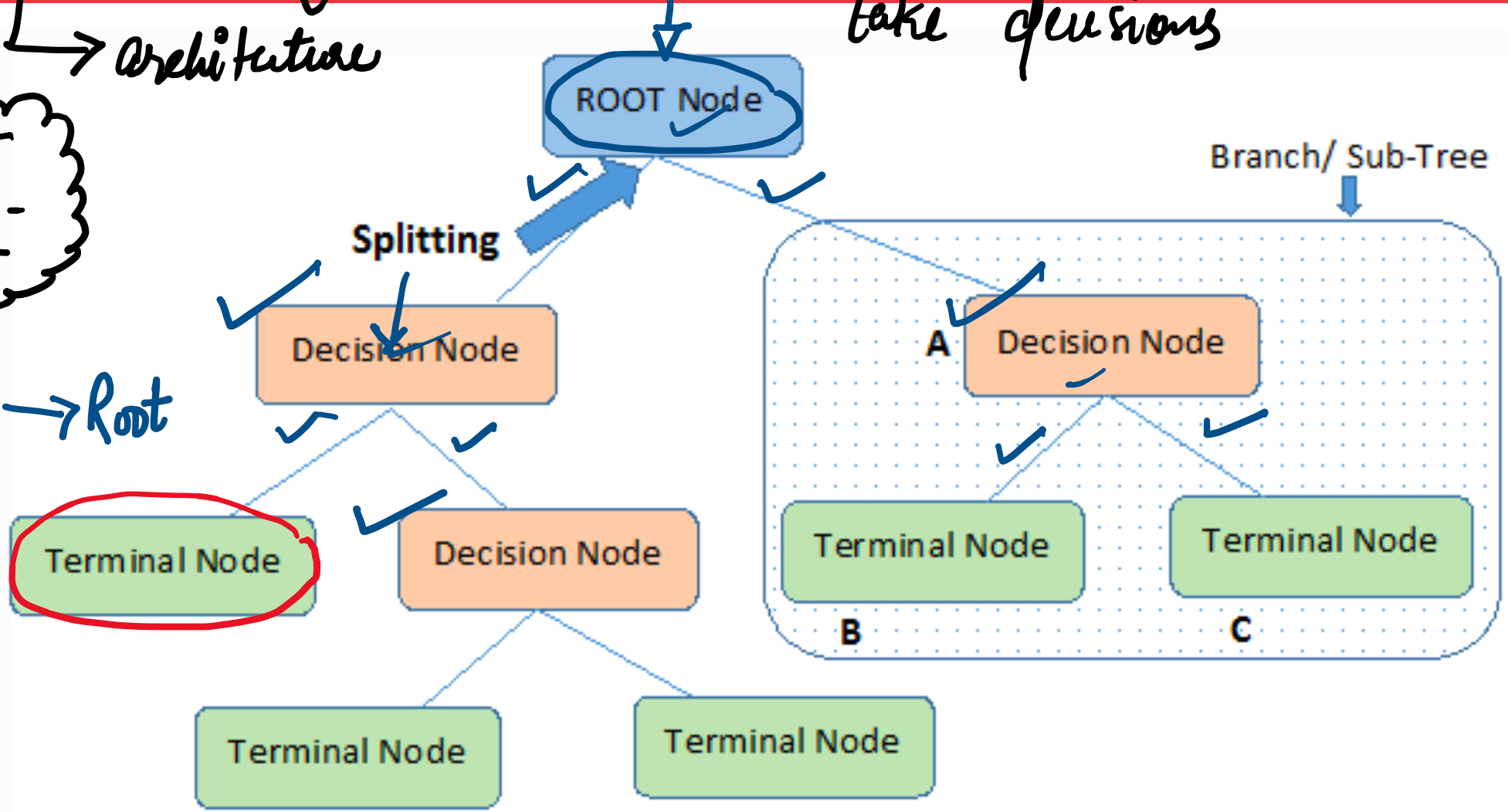
Decision Tree:

} Tree like structure which is used to take decisions

upGrad

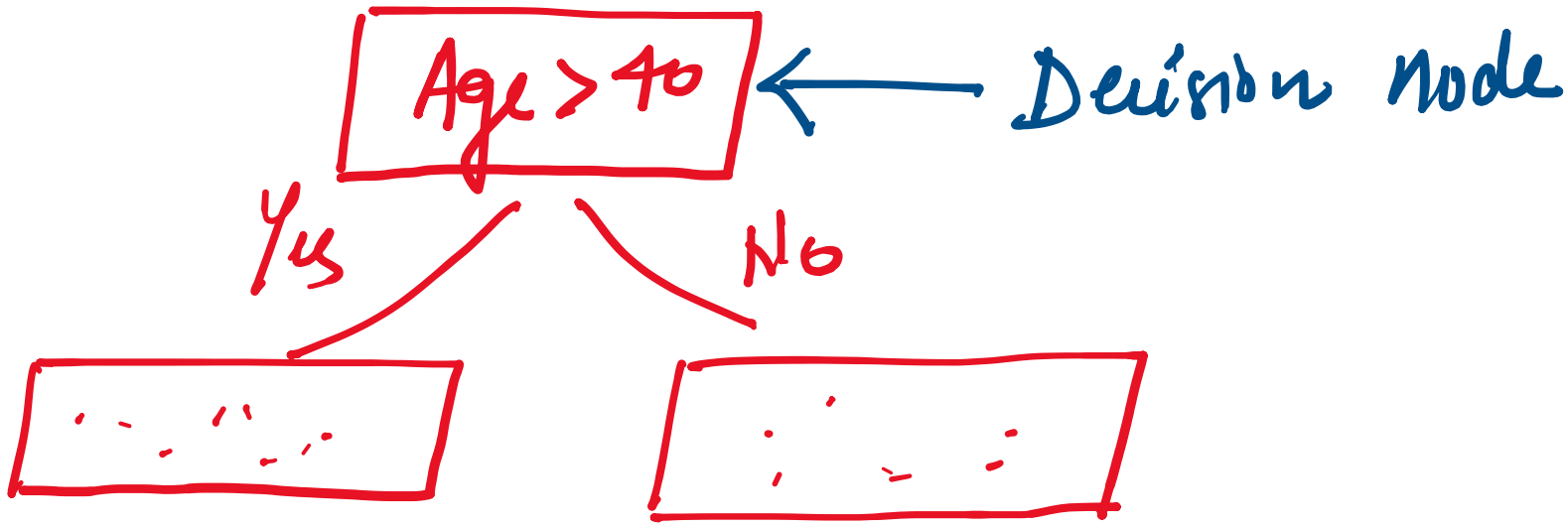


→ Root



Note:- A is parent node of B and C.

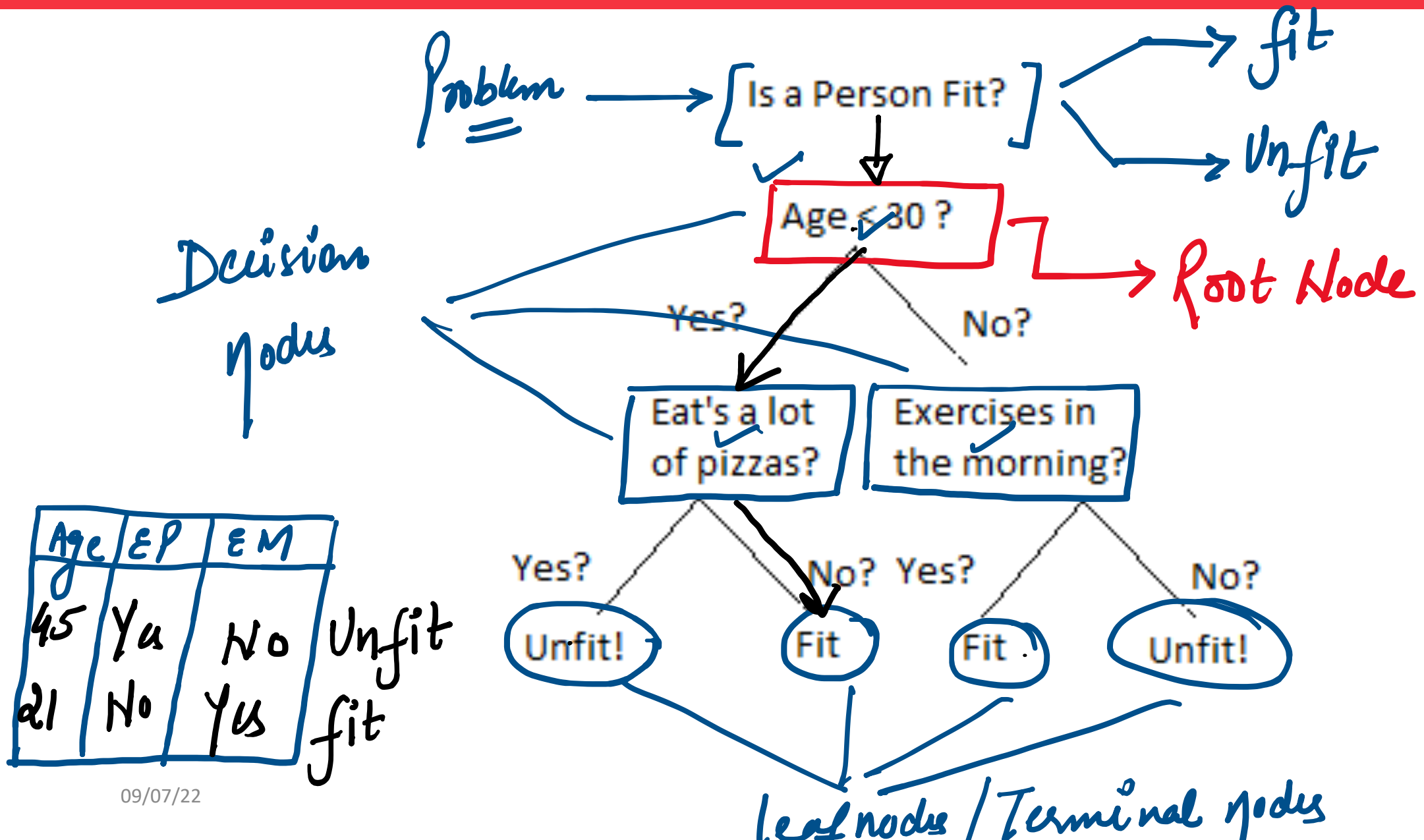
→ Decision Node :- This is the node where we split the data based on some condition on a feature



→ Root Node :- It is very first decision node.

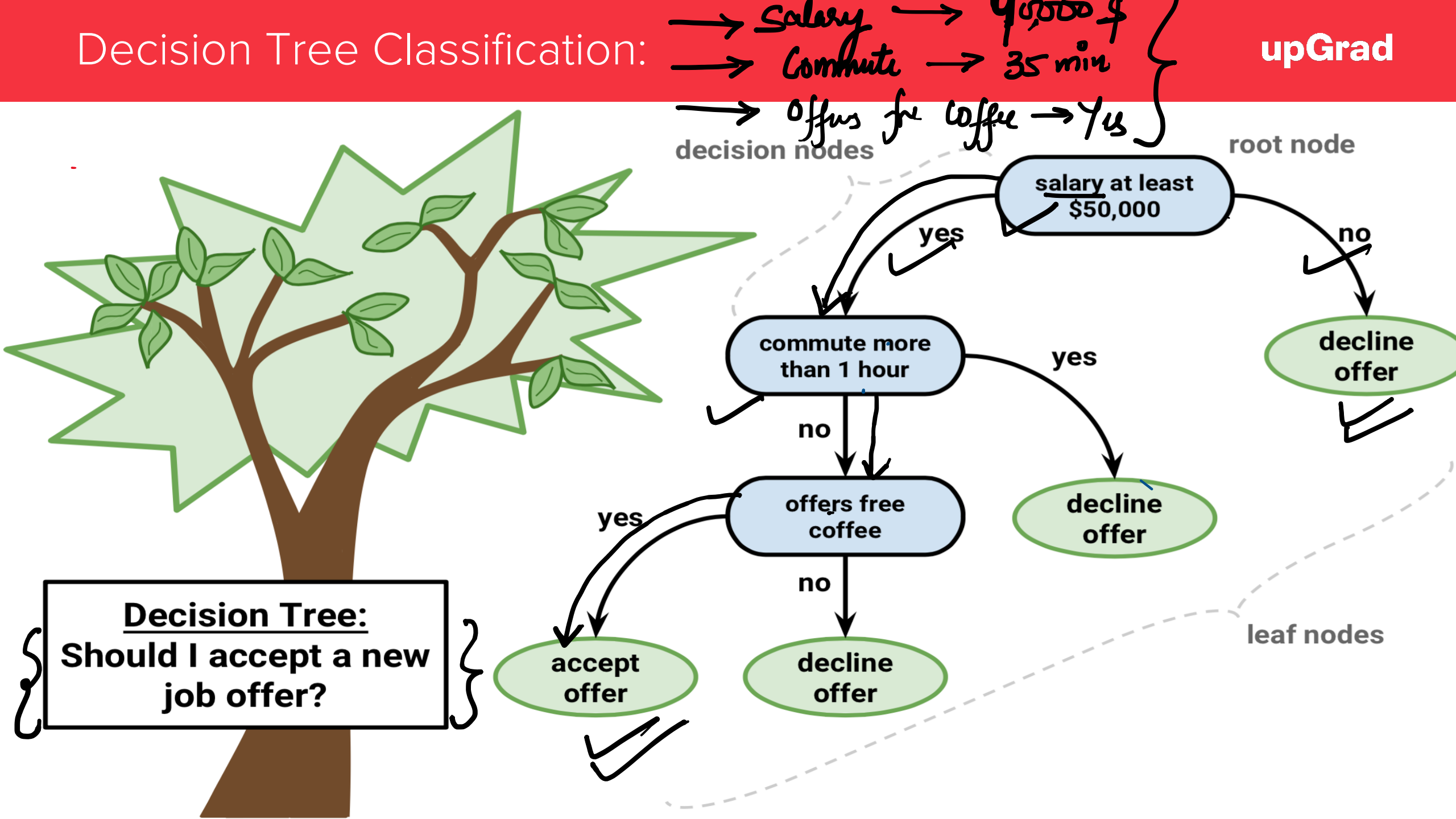
→ Leaf Node or Terminal node :- from where you terminates from the tree and these nodes contain the final output from the tree.

Interpretation of Decision Tree:



Age	EP	EM	
45	Yes	No	Unfit
21	No	Yes	fit

Decision Tree Classification:



Decision Tree Algorithm:

upGrad

1. Recursive binary splitting/partitioning the data into smaller subsets

2. Selecting the best rule from a variable/ attribute for the split

3. Applying the split based on the rules obtained from the attributes

4. Repeating the process for the subsets obtained

5. Continuing the process until the stopping criterion is reached

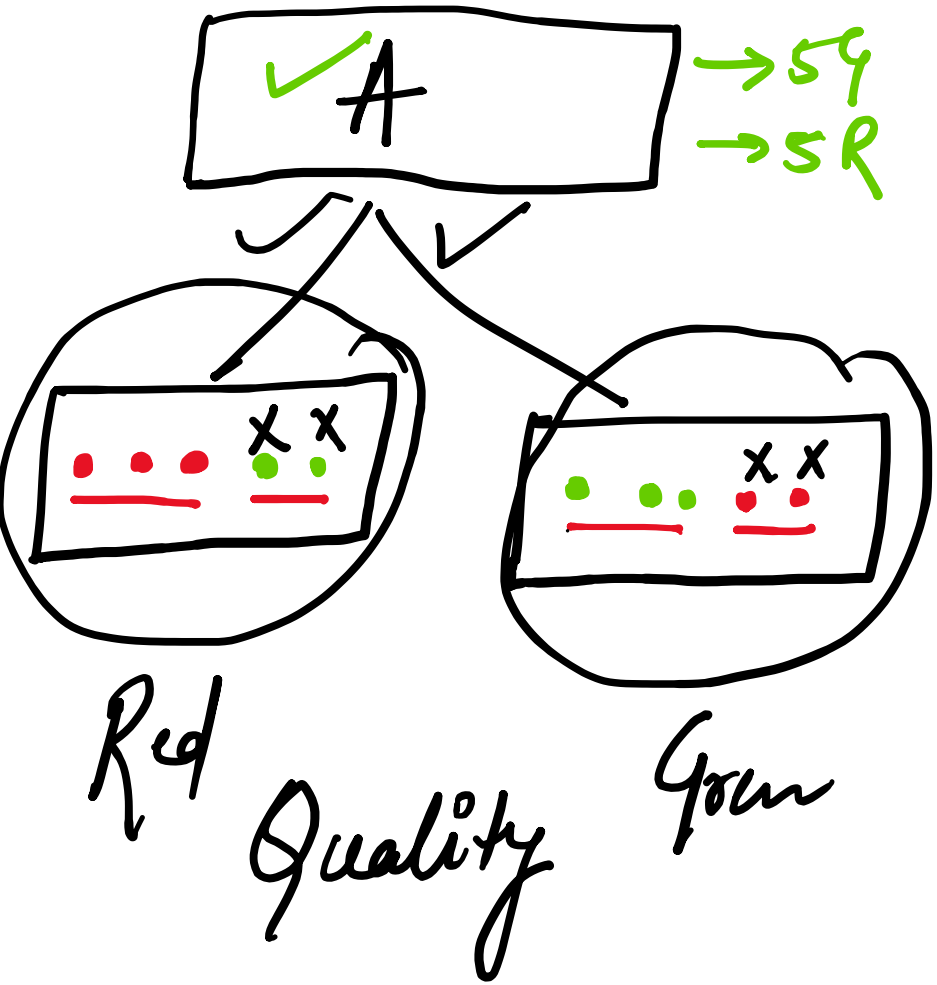
6. Assigning the majority class/average value as the prediction

Decision Tree algorithm

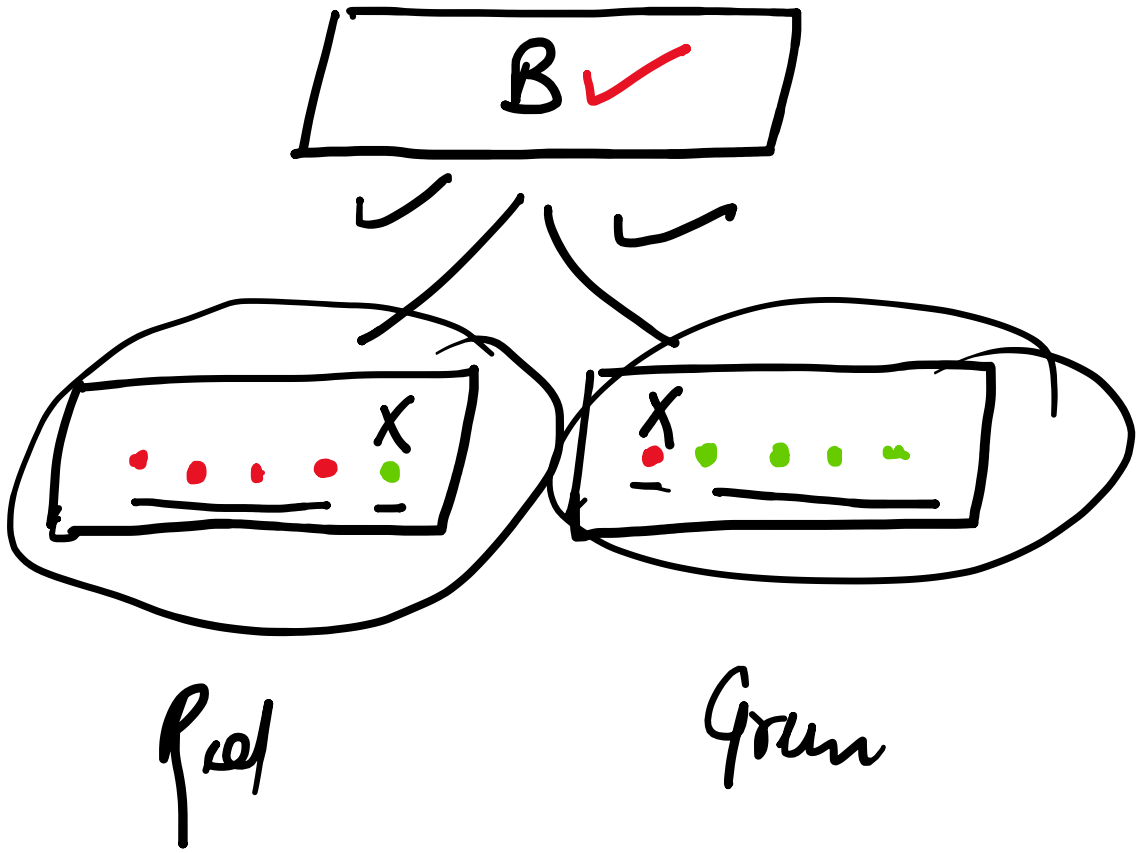
* Recursive selection of features / variables at every decision node using splitting condition and selection criteria

  - Gini

10 4-Mischungen



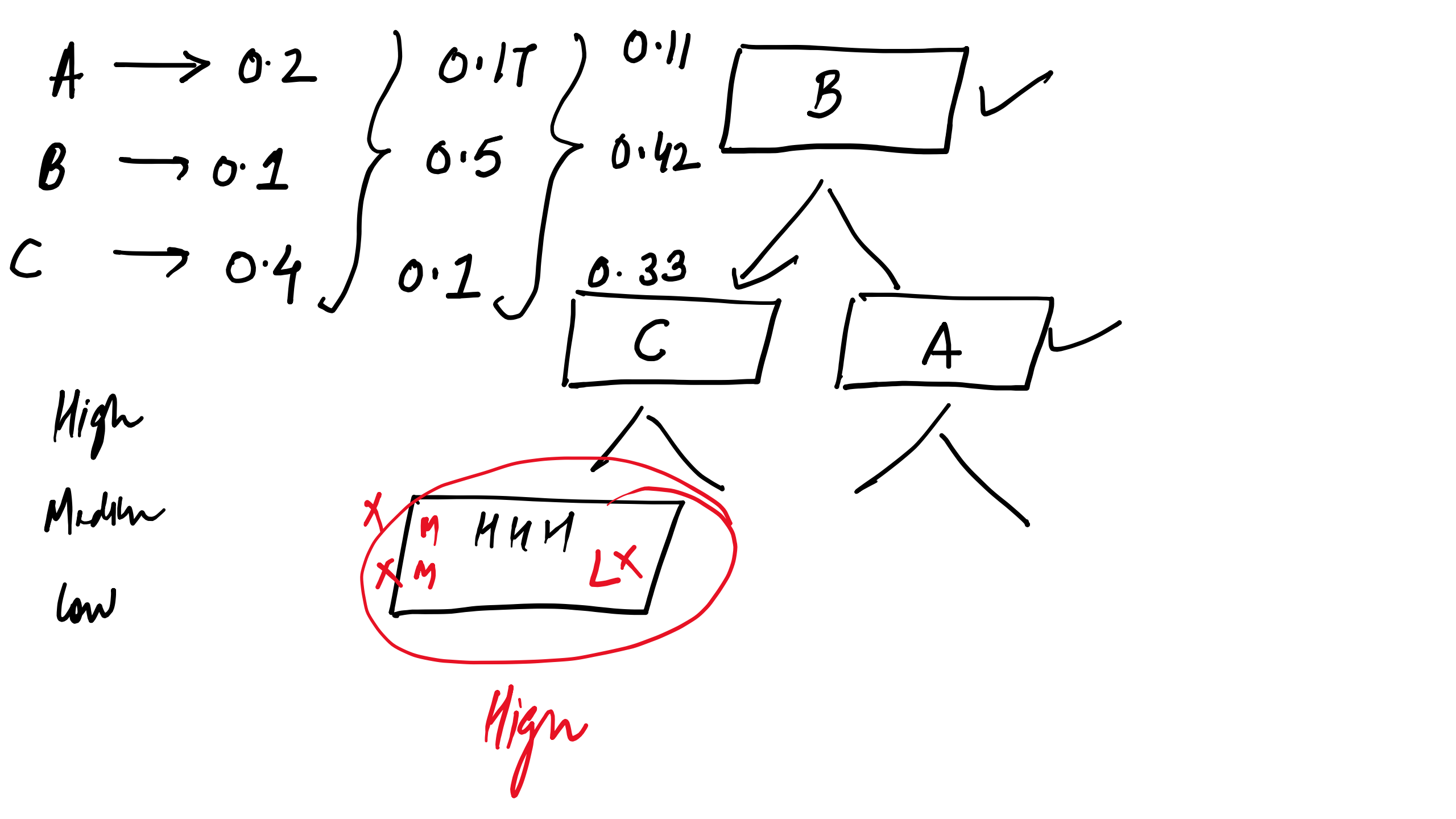
10 2-Mischungen



Gini → Statistical method which is used to judge the quality of split.

Gini ↑ ⟶ Quality ↓

We will be selecting the feature which is giving lower Gini


A → 0.2
B → 0.1
C → 0.4
0.1T
0.5
0.1
0.11
0.42
B
0.33
C
A
High
Medium
low
H H H
L X
High

Splitting Criteria:

upGrad

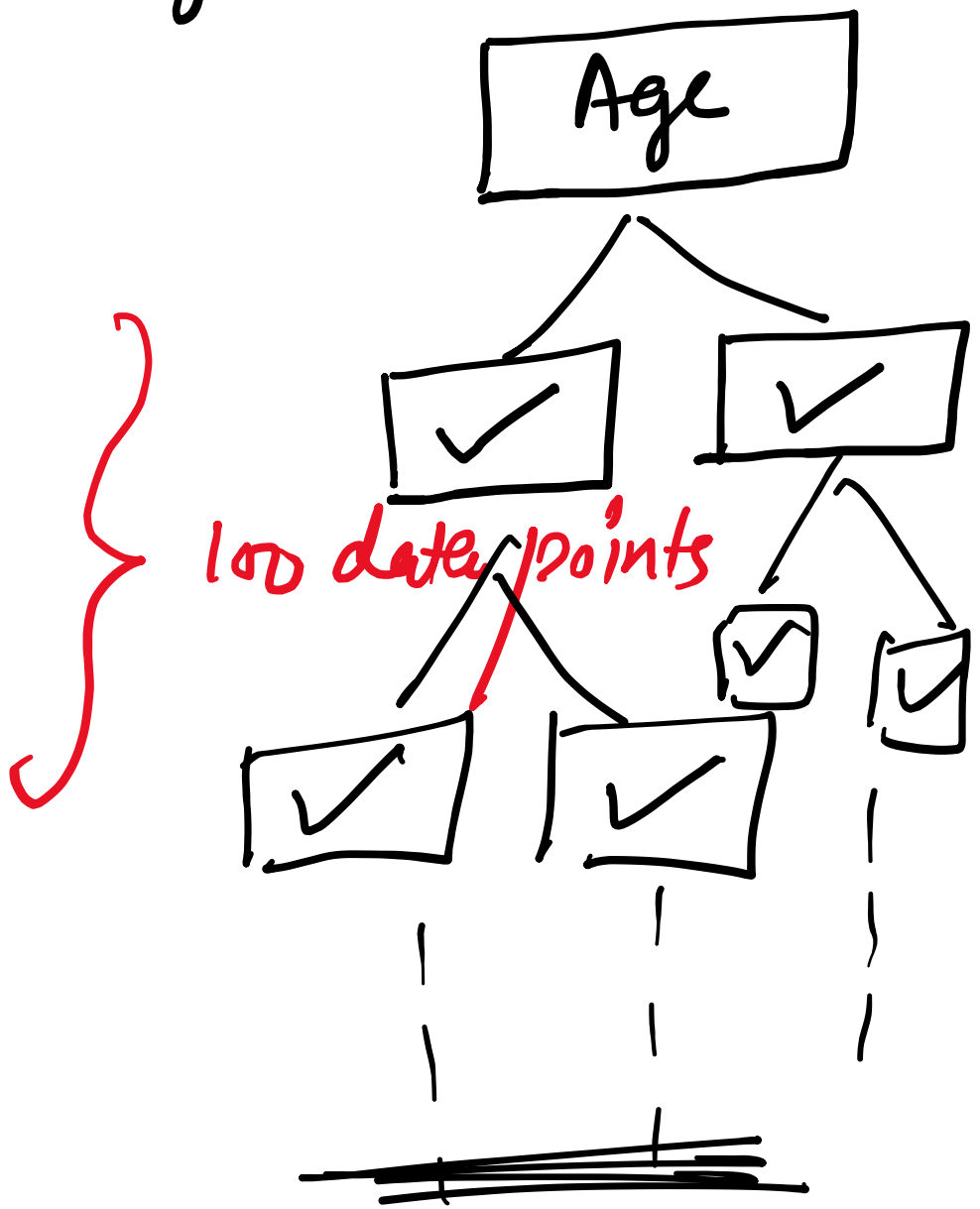
Recall that the Gini index is calculated as follows:

- $G = \sum_{i=1}^{k} p_i(1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2$

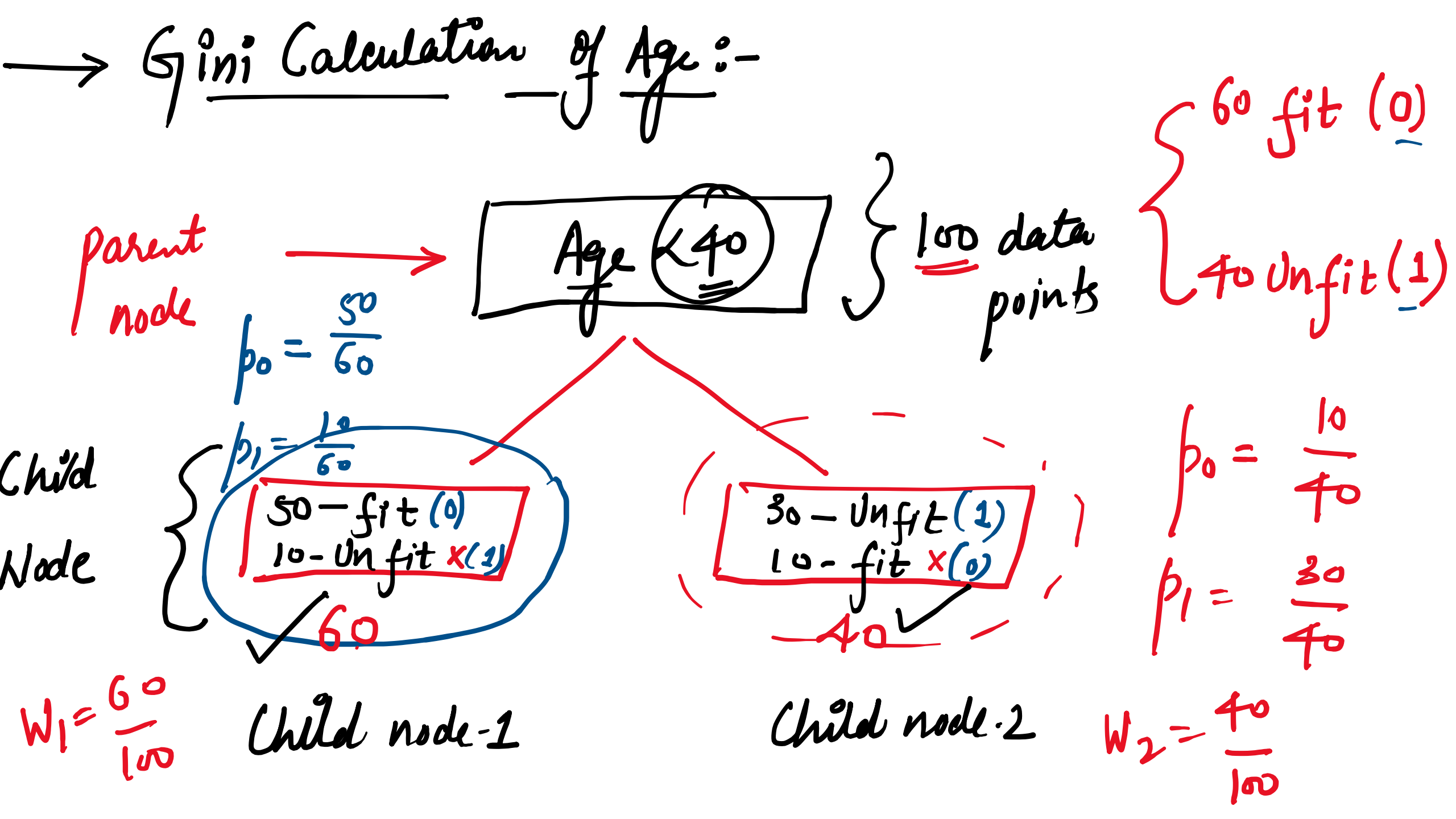
where p_i is the probability of finding a point with the label i, and k is the number of classes.

→ predict whether a person fit or unfit

Age	Cholestrol	Target
20	70	Fit
55	150	Unfit
⋮	⋮	⋮



→ Gini Calculation of Age :-



→ Gini calculation for child node 1 :-

$$G_{C1} = \sum p_i (1 - p_i) = p_0 (1 - p_0) + p_1 (1 - p_1)$$

$$= \frac{5}{6}\left(1 - \frac{5}{6}\right) + \frac{1}{6}\left(1 - \frac{1}{6}\right)$$

$$= \frac{5}{6} \times \frac{1}{\cancel{6}_{3}} \times \cancel{2} = \frac{5}{18}$$

$$\boxed{G_{C1} = 0.277}$$ ✓

→ Gini Calculation for child node 2 :-

$$G_{C_2} = p_0(1-p_0) + p_1(1-p_1)$$

$$= \frac{1}{4}\left(1-\frac{1}{4}\right) + \frac{3}{4}\left(1-\frac{3}{4}\right)$$

$$= \frac{1}{4} \times \frac{3}{4_2} \times 2 = \frac{3}{8}$$

$$= 0.375 \checkmark$$

→ Gini for split:-

$$G_{age} = W_1 G_{C_1} + W_2 G_{C_2}$$

$$= \frac{60}{100} \times 0.277 + \frac{40}{100} \times 0.375$$

→ $\boxed{G_{age} = 0.3162}$

→ Same process to be repeated for Cholestrol

$$G_{Cholestrol} = 0.42$$

$$G_{age} = 0.3162$$

Age (28)

20 → C_1

24 → C_2

28 → C_3

35 → C_4

40 → C_5

Cholesterol (70)

40

45

70 → Min C

75

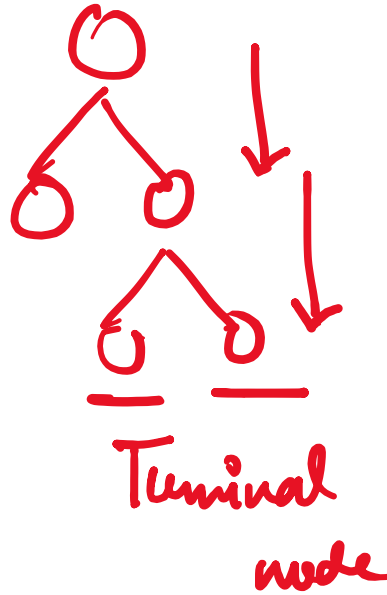
85

110

125

→ Hyperparameter :- } hidden trial

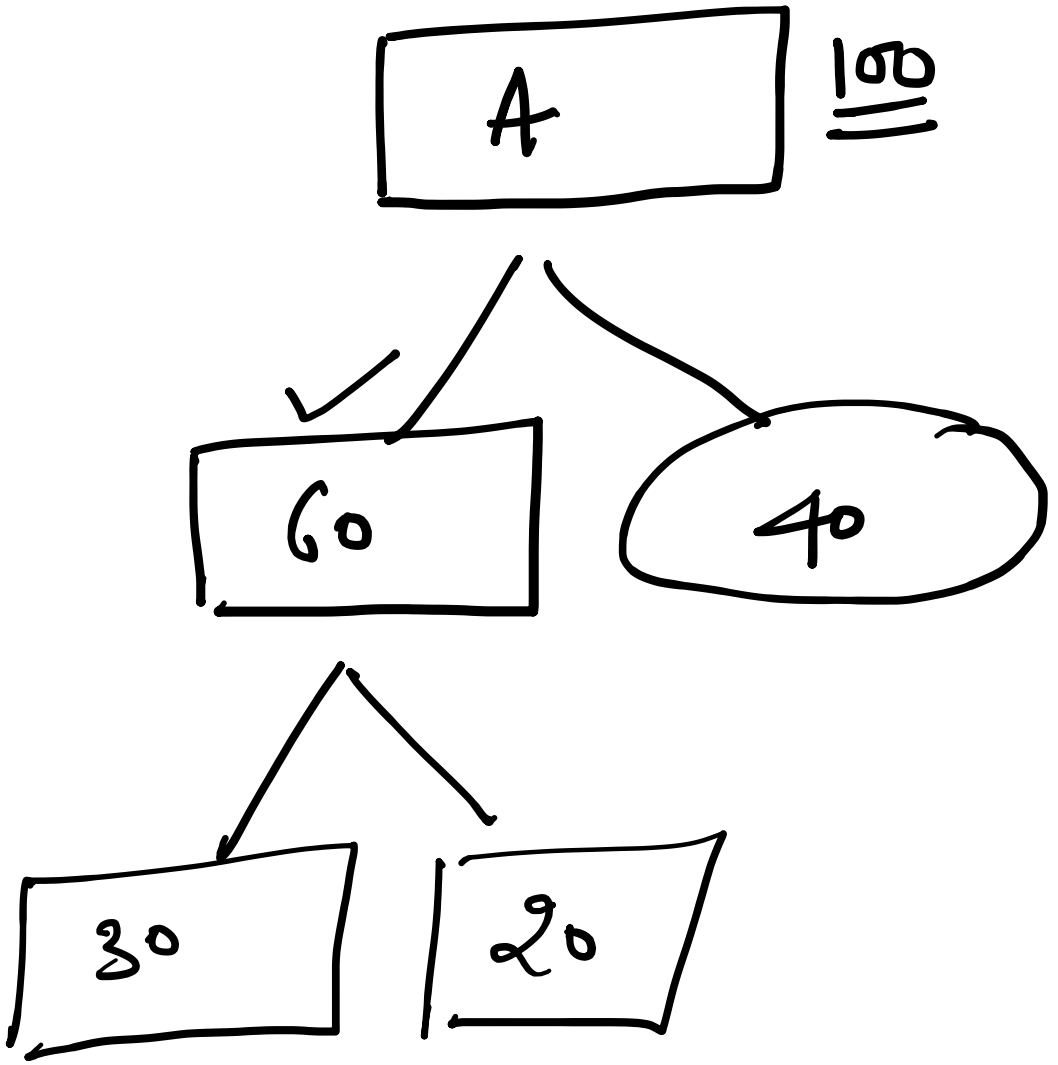
1- Max-depth →



2- Min sample split → minimum no. of samples required to make the split

max-depth – 2 ✓

min-sample split → 50 ✓



Doubts !!!

Random Forest Classification:

upGrad