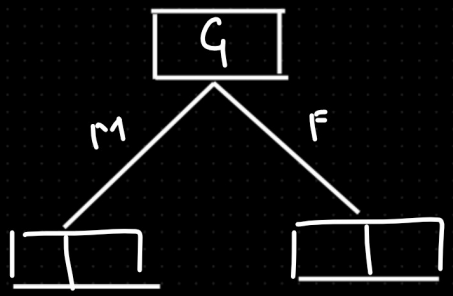
1. Decision tree regressor
2. HyperParameter tuning
3. Cross validation
4. Task | few interview question
5. Pruning (Pre, Post)

Categorical

Gender	Heart Disease
M	Y
F	N
M	Y
F	Y
M	N

Gender (categorical)

Dependent - (categorical)



Numerical

Weight	Heart Disease
220	Y
180	Y
225	Y
150	N
155	N

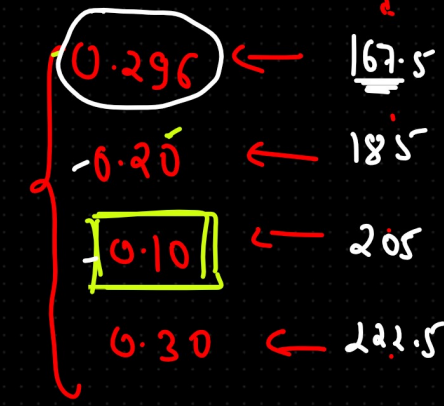
Categorical | Regression

Classification

1. Sort the value
2. take avg of adjustant value
3. Wrt- every avg value i need to find out

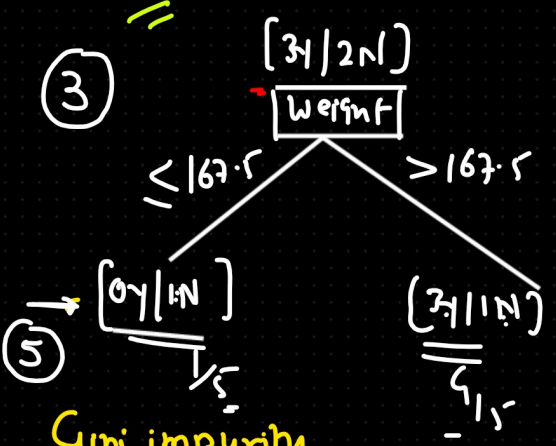
[Gini impurity | entropy- EG]

Weight	Heart-Disease
220	Y
180	Y
225	Y
150	N
155	N



Whereever we will get a low impurity will take that threshold only

Weight	HD
155	N
180	Y
90	N
220	Y
225	Y

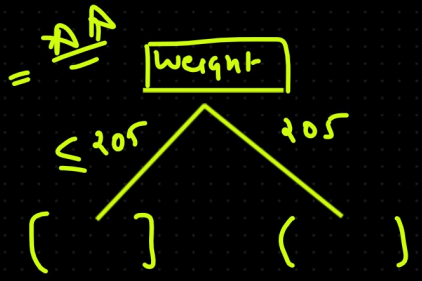


Gini impurity

$$1 - \sum_{i=1}^{n} p_i^2$$

Info gain always will be high that only feature we will select-

Gini impurity will only lowest one



Gini impurity [L] = 0

Gini imp [R] = $1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$

$= 1 - \left[\frac{9}{16} + \frac{1}{16}\right]$

$= 1 - \left[\frac{10}{16}\right]$

$= \frac{16-10}{16} = \frac{6}{16} = 0.37$

⇒ **Weight Gini impurity**

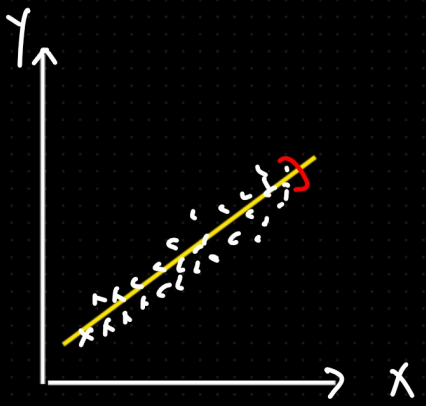
$= \frac{1}{5} \times 0 + \left(\frac{4}{5} \times 0.37\right)$

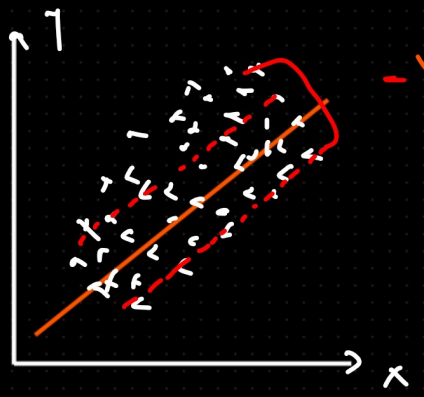
$= 0 + 0.296 = \boxed{0.296}$

Decision-tree-regressor

Height	Weight (dependent)
165	65
160	50
180	90
170	85
175	70

Gender	Weight (Dependent)
M	66
F	50
M	90
F	85
M	70



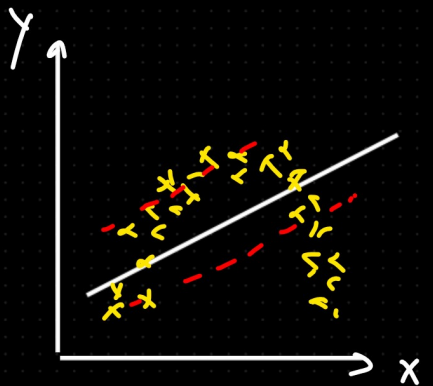


- $Y = mx + C$

⇓

Linear reg

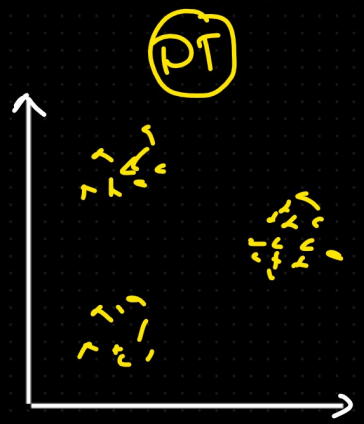
Svm



Complex → DT

⇓

Ensemble



Height	Weight
165	65
160	50
180	90
170	85
175	70

CART

{ Binary
Gini impurity }

ID3

Id
or multiclass

{ Binary classification
IG => entropy }

C4.5

4
40
400

Classification

1. sort
2. average
3. Gini impurity | IG

Height	weight
160	50
165	65
170	85
175	70
180	90

121 ← 162.5
116 ← 167.5
105 ← 172.5
150 ← 177.5

(50, 65, 85, 70, 90)

Var | Height

≤ 162.5 | > 162.5

⇒ [50] | [65, 85, 70, 90]

Var L | Var R

1/5 | 4/5

5

Regression

1. sort
2. average
3. Variance | Reduction in var

(MSE)

Var root = [50, 65, 85, 70, 90]

$$Var = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y})^2$$

$\sqrt{Var} = ?$ Std dev (RMSE)

$Var(L) = 0$

$Var(R) = [65, 85, 70, 90]$

$$Avg = \frac{65+85+70+90}{4}$$

$$= 77.5$$

$$Var(R) = \frac{(77.5-65)^2+(77.5-85)^2+(77.5-70)^2+(77.5-90)^2}{4}$$

$Var(R) = 106.25$

$$Mean = \frac{50+65+85+70+90}{5}$$

$$= 72$$

$$Var(Root) = \frac{(72-50)^2+(72-65)^2+(72-85)^2+(72-70)^2+(72-90)^2}{5}$$

$$= \frac{(22)^2+(7)^2+(13)^2+(2)^2+(18)^2}{5}$$

$Var(Root) = 206$

$Var(Root) = 206$

$Var(R) = 106.25$

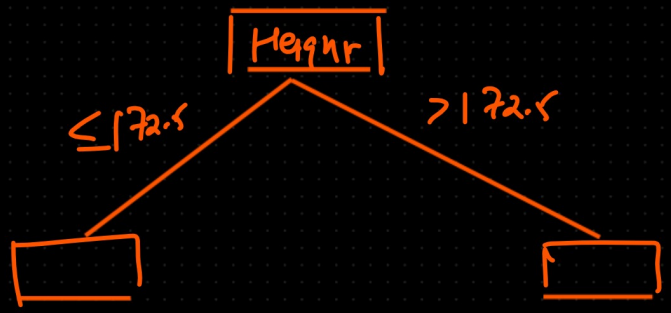
$Var(L) = 0$

$$\text{Reduction in var} = Var(root) - \sum_{i=1}^{n} w_i \times var(Child)$$

$$= 206 - \left[\tfrac{1}{5} \times 0 + \tfrac{4}{5} \times 106.25\right]$$

$$= 206 - [0 + 85] = 121$$

Lowest Reduction in var



Classification

Weight	Gender	O/NODb
50	m	O
60	f	O
70	m	NO
60	f	O
50	m	NO

D.P

= Weight | Gender | Height

D.P

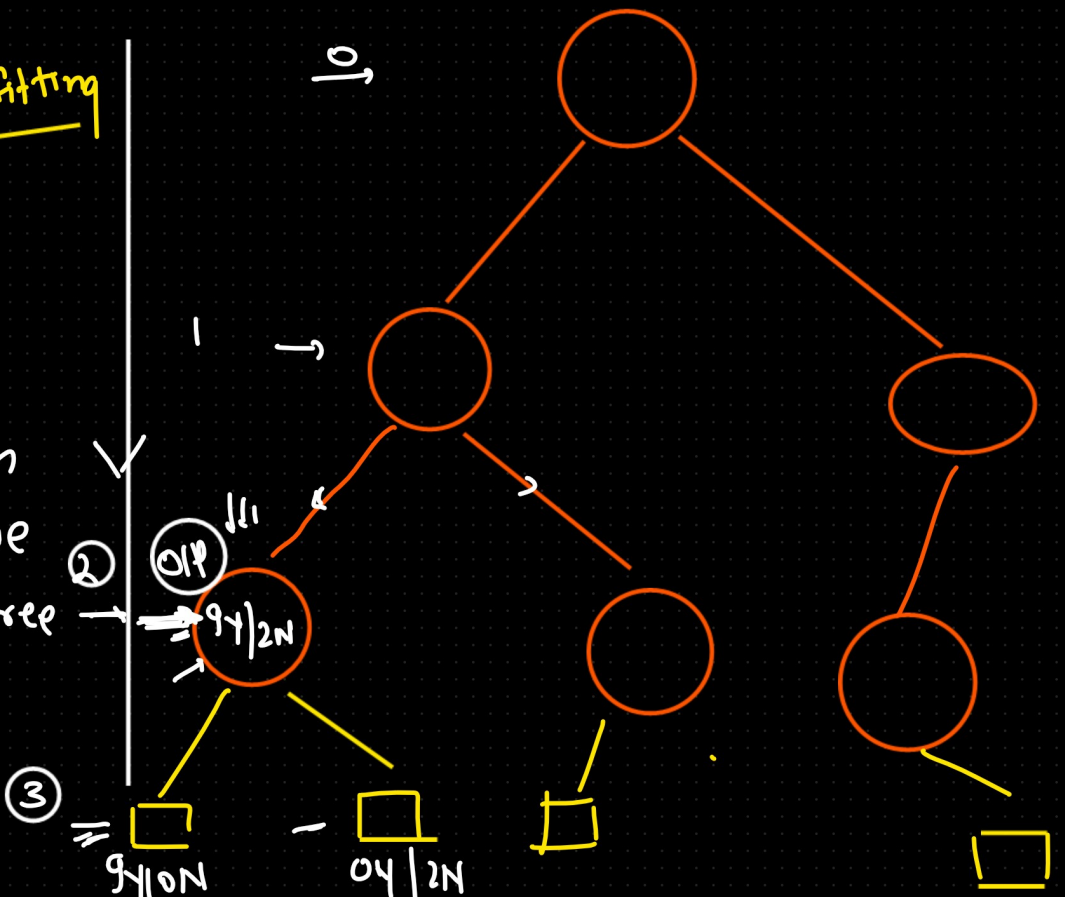
Pruning
- Pre Pruning
- Post Pruning

Why?

Overfitting

Cutting

Pre Pruning ⇒ While i am building Decision tree at that time only we decide the depth of the tree (Structure)



1. max_depth = 2
2. min_sample_split
3. min_sample_leaf
4. max_feature

} hyperparameter

Overfitting ⇒ training accuracy will be high
test accuracy will low

Underfitting ⇒ training acc is low
test acc is also low

IDZE

training accuracy and test accuracy should be near to each other there won't be much Diff

Hyperparameter ⇒ $Y = mx + c$

Linear reg

Parametric

Non Parametric model

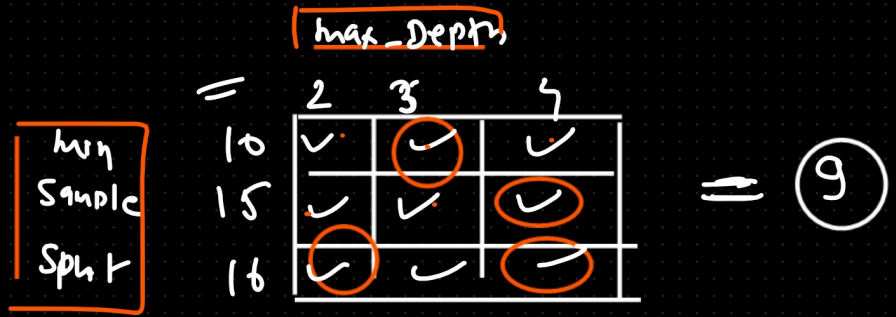
mat. Depth
min_sample

Cross validation

1. Grid Search CV
2. Random Search CV

Max_Depth = [2, 3, 4]

Min_Sample_Split = [10, 15, 16]



= 9

How many time

9 times

max_Depth, min_Sample_split, Leaf

more complex

= Grid search ⇒ In Grid search we take every combination

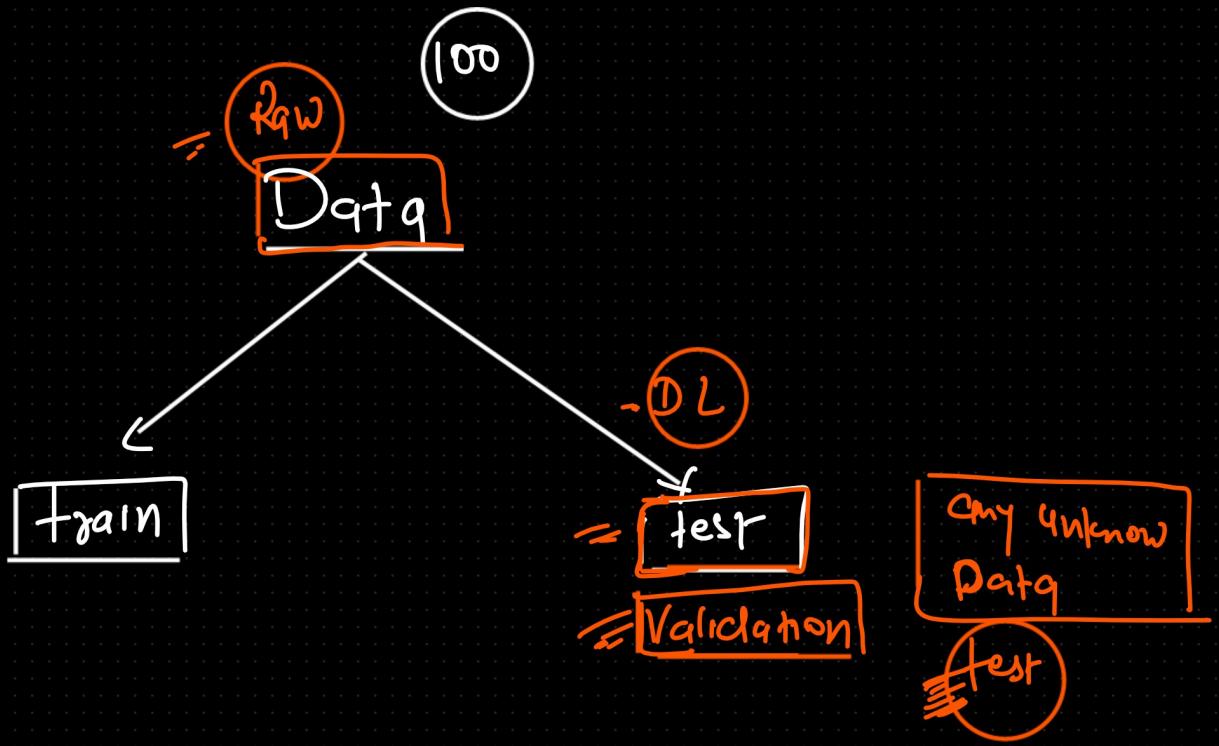
Random Search ⇒ we take random combination

less complex

Cross validation

Diff = Random split (train_test_split)

1. 87 accuracy
2. 88 accuracy
3. 85 accuracy
4. 80 accuracy



= Cross-Validation (testing)

CV

K-Fold CV

10 Data Point

= k = 5 fold

$\frac{10}{5} = 2$ ⇒ test Data

10 − 2 = 8 ⇒ training data

$\frac{10}{10}$

10 10

I J

testing training

- K = 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | = 90
- K = 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ⇒ 95

train Test train

- K = 3 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ⇒ 91

training test training

- K = 4 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ⇒ 90
- K = 5 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ⇒ 85

Assumption

Data

5|6 Grid Search CV

max_Depth = 3

min-sample_Split = 3

K-fold CV K = 5

⇒ 3 X 3 X 5

⇒ 45

Instead of Random Split ⇒ Cross validation

↓

Cross testing ⇒ iterate on each Data Point

disadvantage:

overfitting
if there any complex pattern of the data then it won't perform
it is little sensitive missing value
for the large dataset it won't be good

how to use decision tree for the feature selection?

if my data in imbalanced then should I use the DT and it is robust to DT?

first handle the imbalance data and then go for the building up this DT.

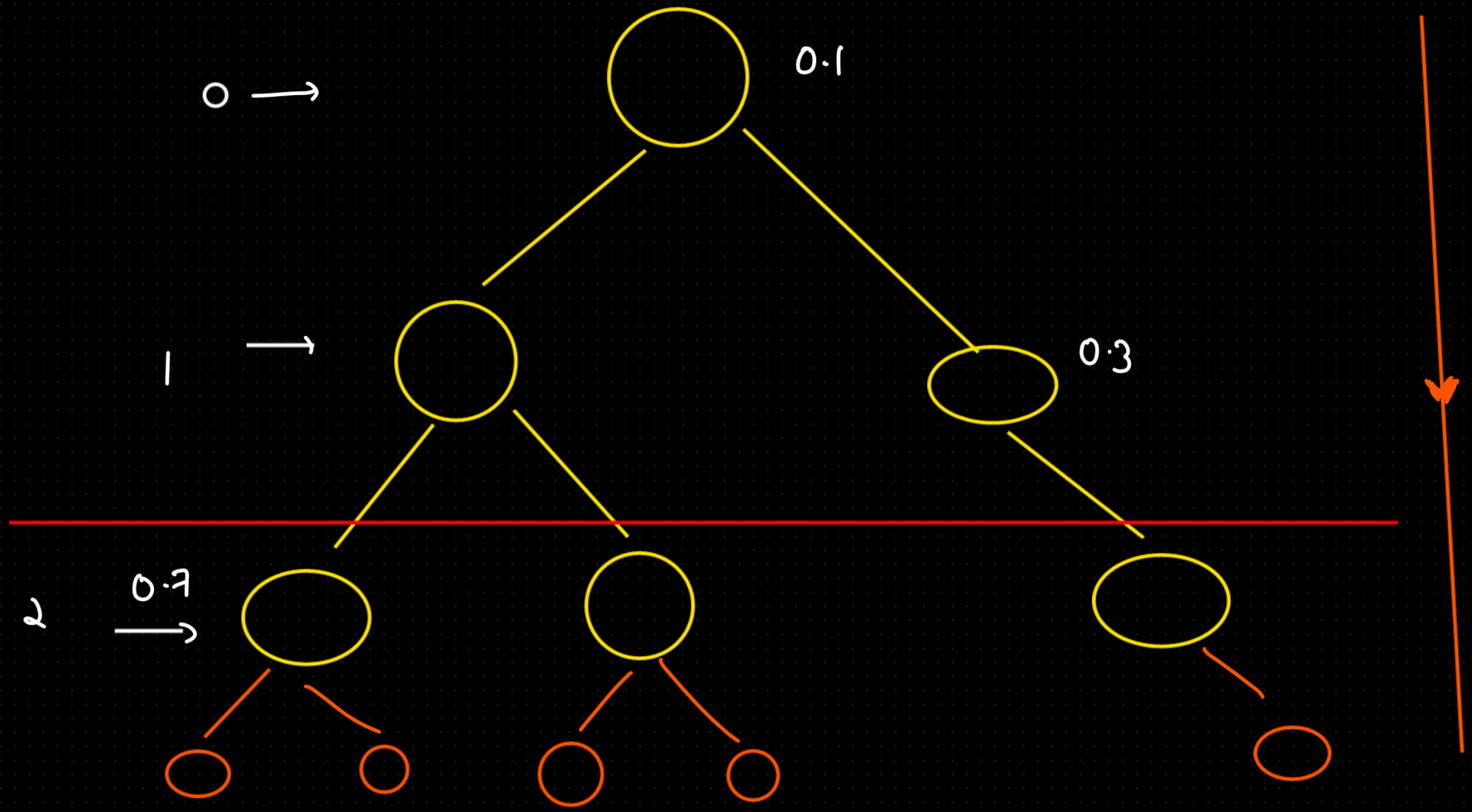
10==>> 8 yes and 2 no

if you are giving such data to decision tree it might lead to wrong predicitn

q10) can you give me those senecio there I should not use DT?

can you tell them disadvantage

Post Pruning ⇒ first we building the entire tree till complete Depth.
and then we are going to Prune it.
⇓
cutting



How it is Possible => Build up our DT ✓

With respect to every Node we have impurity value

Setup the threshold → 0.5, 0.9, 0.8, 0.6

ccp_alpha = [0.5, 0.9, 0.8, 0.6]

PrePruning => Decide the Depth of the tree while building up.

(PostPruning) => cut the branches after building up the complete tree

CCP-alpha -[-----]