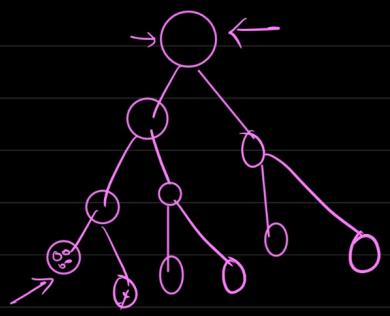
* Decision tree Post pruning and Preprunning

⟹ DT is a greedy algorithm.

⇓

DT will be fully grown until each leaf node is a pure node.

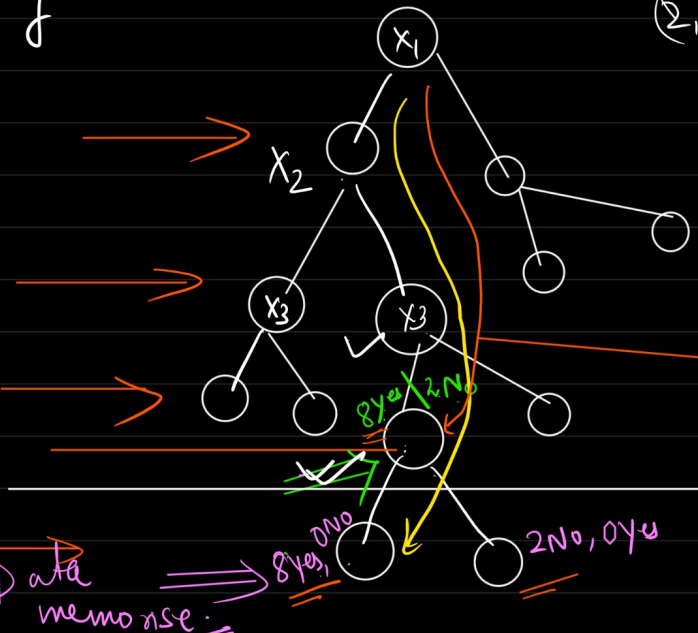


⇓

Overfitting → training acc ↑
testing acc ↓

* Generalised model → low bias, low variance
training acc ↑ testing acc ↑

X_1 X_2 X_3 y (2, 3, 5)

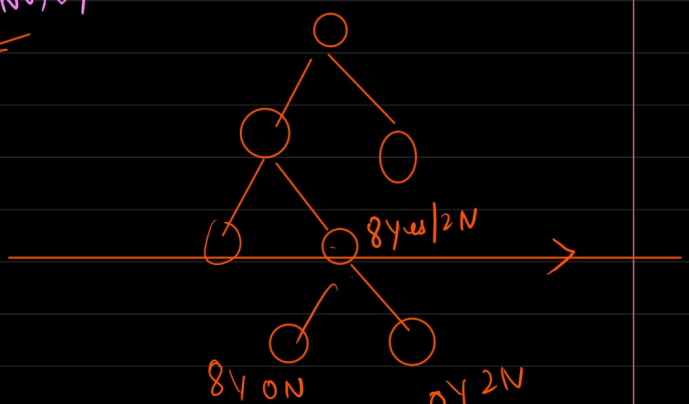


max-depth earlier = 4
Now = 3

→ Yes count is more, prediction would be Yes.

Data memorise.

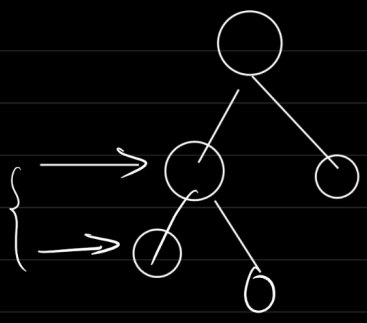
To reduce overfitting
① Post pruning.
② Pre-pruning
③ Ensembles.



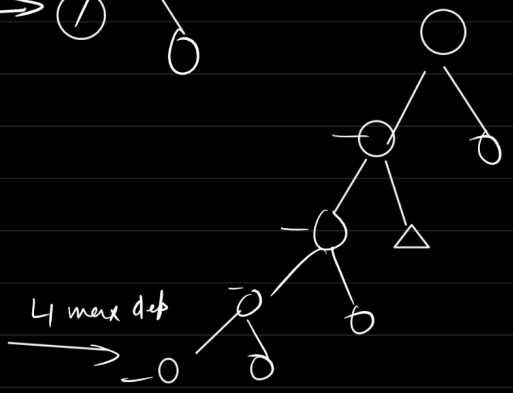
* Params to create D.T

→ Criteria ← Gini
Entropy.

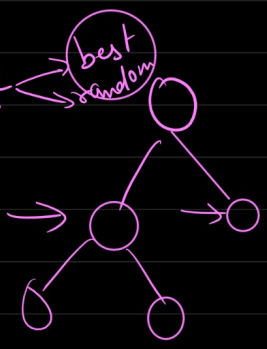
→ max_depth — maximum depth of tree.

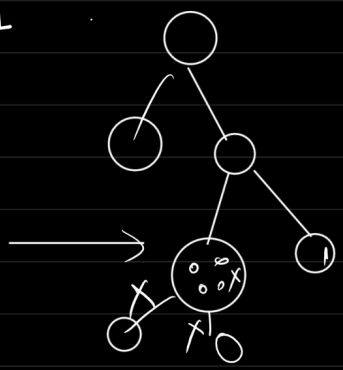


→ min_sample_leaf. — The minimum no of samples needed to be considered as a leaf node.
→ default is 1



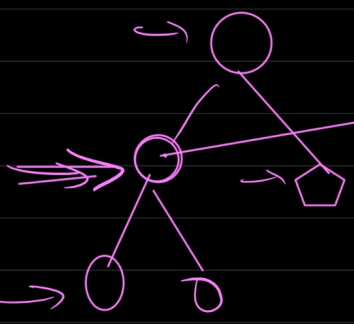
→ splitter ← best / random





5

→ min_samples_split — The minimum no of sample a node must contain to consider splitting.



→ 10 datapoint

default - 2

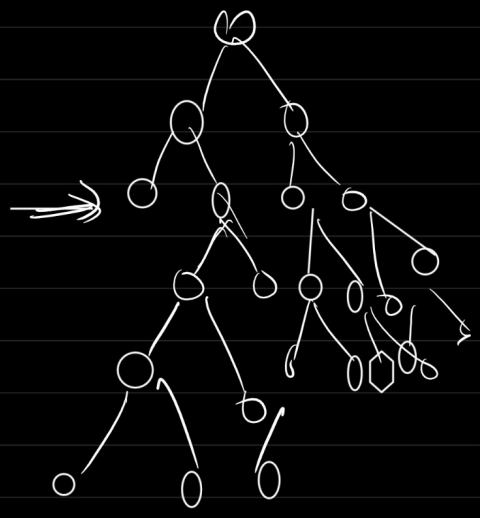
→ max features → The number of features to consider for the best split.

① Post pruning (cutting of dT after it is created)

→ Construct the entire Decision till leaf node

→ Prune the decision tree

→ Should be only used for Smaller dataset



② Pre-pruning

→ Limit the growth of DT

* Hyper parameter Tuning to select the best parameter.