Decision Tree (DT)

Agenda

1. Build DT with numeric values
2. DT regressor
3. Pre pruning and Post pruning.

1. Build DT with numeric values.

I/P weight (numeric)	O/P heart disease
220	Y
180	Y
226	Y
190	N
155	N

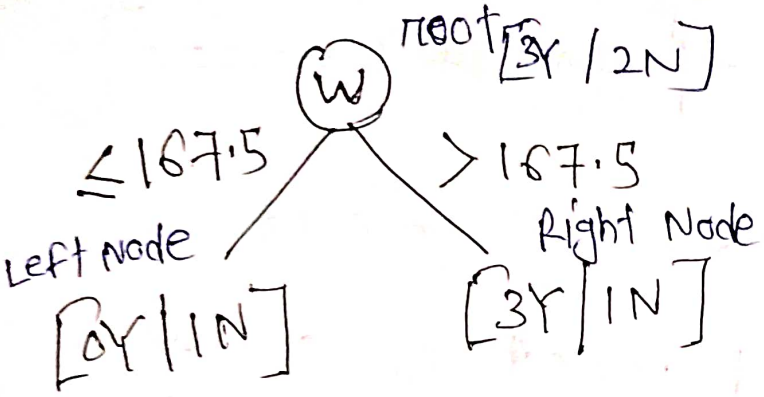
After ↓ sort

weight	heart disease
155	N
180	Y
190	N
220	Y
226	Y

Avg of adjacent value → 167.5, 185, 205, 223

weight	heart disease
155	N
180	Y
190	N
220	Y
226	Y

1. Sort the values.
2. Average of the adjacent values.
3. W.r.t every avg. value we need to find out gini impurity/ entropy then Info. Gain



[like this threshold (167.5) we have to segregate each average values]

$$\text{Gini impurity} = 1 - \sum_{i=1}^{n} p^2$$

$$(\text{left node}) = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right]$$

Leaf Node $= 1 - 1 = 0$

$$\text{Gini impurity (GI)} \\ (\text{Right Node}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$$

$$= 1 - \frac{10}{16} = 1 - \frac{5}{8} = \frac{3}{8}$$

$$= 0.375$$

$$\text{Info. Gain} = GI[\text{ROOT}] - \sum \frac{|S_v|}{|S|} \times GI[\text{child}]$$

$$GI(\text{ROOT}) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right]$$

3Y / 2N

$$= 1 - \left(\frac{13}{25}\right) = \frac{12}{25} = 0.48$$

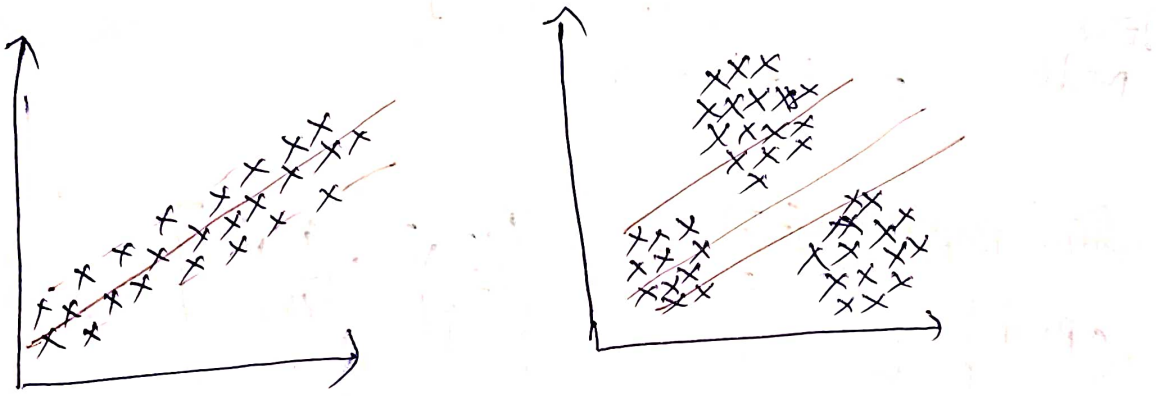
Info. Gain = $0.48 - \left[\frac{1}{5} \times 0 + \frac{4}{5} \times 0.375\right]$

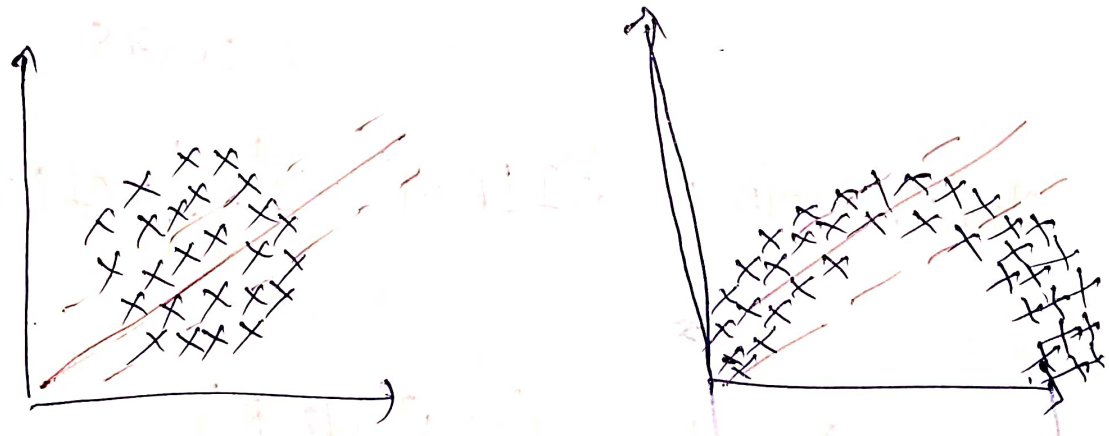
(167.5)

$= 0.48 - 0.30 = 0.18$

→ Find the threshold values and segregate data w.r.t each threshold.

Decision Tree Regressor





By using Linear Regression or SVM we will not get high accuracy (R^2) in regression.

So we will use Decision tree regressor (DTR)

I/P (numeric) height	o/p (numeric) weight
165	65
160	50
180	90
170	85
175	70

Sort values →

height I/P	weight O/P
160	50
162.5 < 165	65
167.5 < 170	85
172.5 < 175	70
177.5 < 180	90

(DT) Classification

→ entropy / Gini Impurity

→ Information Gain

(DT) VS Regression

→ Mean

→ MSE | MAE | RMSE

→ Reduction in variance

[height] [50, 65, 85, 70, 90]

mean = 72

≤ 162.5 | > 162.5

[50] [65, 85, 70, 90]

mean = 50

mean = $\frac{65+85+70+90}{4} = 77.5$

$$\text{MSE / variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{hight (variance)} = \frac{(72-50)^2 + (72-65)^2 + (72-85)^2 + (72-70)^2 + (72-90)^2}{5}$$

$$= 206$$

for left variance = 0

$$\text{Var[Right]} = \frac{(77.5-65)^2 + (77.5-85)^2 + (77.5-70)^2 + (77.5-90)^2}{4}$$

$$= 106.25$$

$$\text{Reduction in variance} = \text{Var[root]} - \sum_{i=1}^{n} w_i \times \text{Var[child]}$$

$$\text{Reduction in variance} = 206 - \left[\frac{1}{5} \times 0 + \frac{4}{5} \times 106.25\right]$$

$$= 206 - 85 = 121$$

→ we have to find the reduction in variance for each threshold

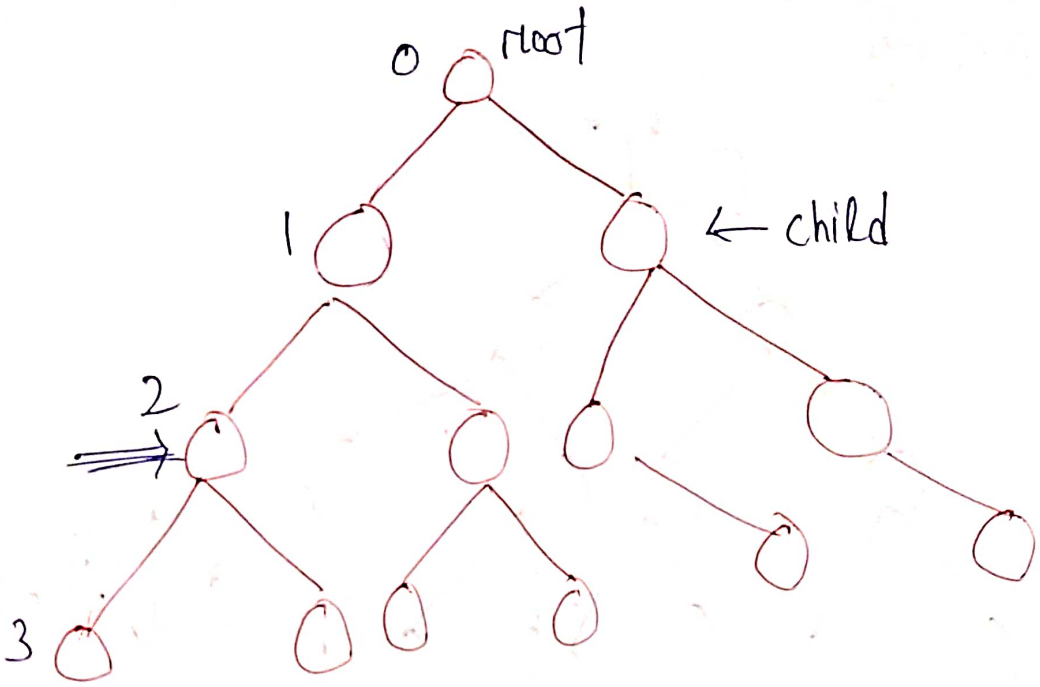
→ Then find the lowest variance.

Pre Pruning / Post Pruning

Both pruning methods are used to cut or prun the tree to avoid overfitting.

Pre pruning

Prepruning is done while we are creating decision tree.



if max_depth = 2

we will create tree upto 2.

if minimum_sample_leaf = 10

we cut the tree in it should not obey these criterion.

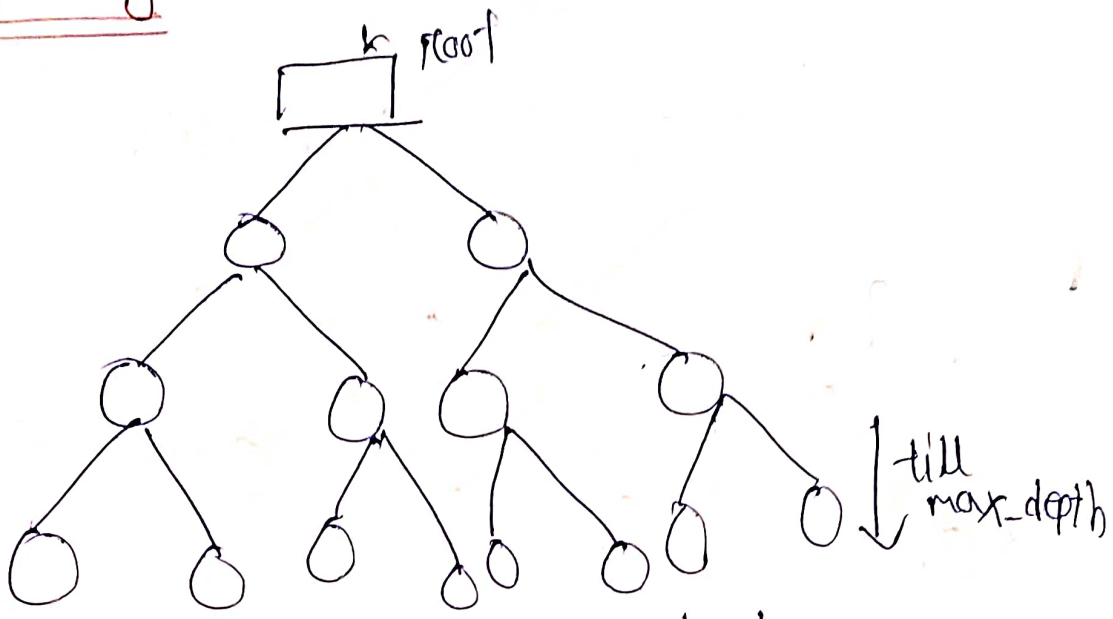
minimum_sample_split = 5

if that node has less sample then. minimum_sample_split then we will not split that node.

1. max-depth
2. minimum_sample_split
3. minimum_sample_leaf
4. max_feature

} hyperparameter for prepruning

Post-Pruning



→ Build Decision tree till last

- ccp_alpha (cost complexcity pruning)

→ cut the decision tree using ccp_alpha.

ccp_alpha: This is a threshold compared with entropy / gini impurity

ccp_alpha ↓ (low) Decision tree depth ↓ (low)

ccp_alpha ↑ DT depth ↑