12-NOV-2022

~~Decision Tree Regressor~~

Build Decision Trees with numeric values

Continuous → Weight	Heart Disease (Y/N)
220	Y
180	Y
225	Y
190	N
155	N

Steps

① Sort the data in ascending order

weight	Heart disease (Y/N)
155	N
180	Y
190	N
220	Y
225	Y

② find the correct threshold.

Q. How can we find threshold?

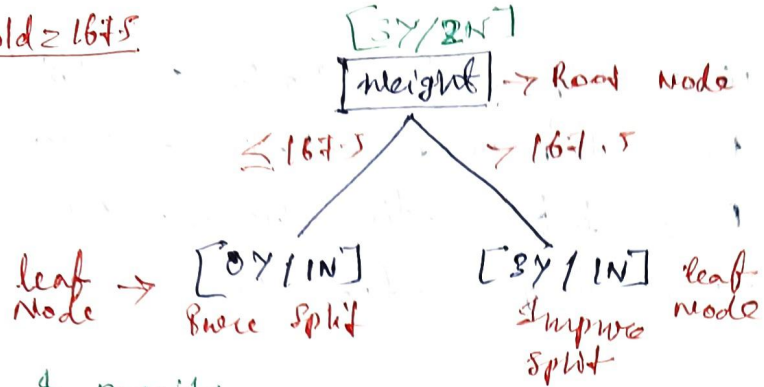
Ans ② find the average of the adjacent value

Average	weight	Heart disease (Y/N)
	155	N
167.5		
	180	Y
185		
	190	N
205		
	220	Y
222.5		
	225	N

③ find the ~~brini~~ impurity or ~~T.E.~~ entropy for every average value and finally calculate Information Gain.

{ Note: We can take other threshold as well. }

Threshold = 167.5



Gini Impurity

$$G.I = 1 - \sum_{i=1}^{n} P_i^2$$

$G.I$ [left Node] = 0

$G.I$ [Right Node] = $1 - (P_Y^2 + P_N^2)$

$= 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)$

$= 1 - \left(\frac{9+1}{16}\right)$

$= \frac{6}{16} = \frac{3}{8}$

$= 0.375$

Information Gain

Information Gain = $G.I[Root] - \sum_{v \in values} \frac{|Sv|}{|S|} H.I[child]$ → Weighted Impurity

$G.I$ [Root Node] = $1 - (P_Y^2 + P_N^2)$

$= 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)$

$= 1 - \left(\frac{9}{25} + \frac{4}{25}\right)$

$= \frac{12}{25} = 0.48$

$I.G = 0.48 - \left\{\frac{1}{5} \times 0 + \frac{4}{5} \times 0.375\right\} = 0.18$ → with 167.5 as threshold

~~Threshold~~

Let's assume,

for threshold 167.5, Information Gain = 0.18 (calculated)
for threshold 185, Information Gain = 0.20 (assumed)
for threshold 205, Information Gain = 0.60 (assumed)
for threshold 222.5, Information Gain = 0.3 (assumed)

~~We~~ We will choose threshold as 222.5 because it is having more information gain.

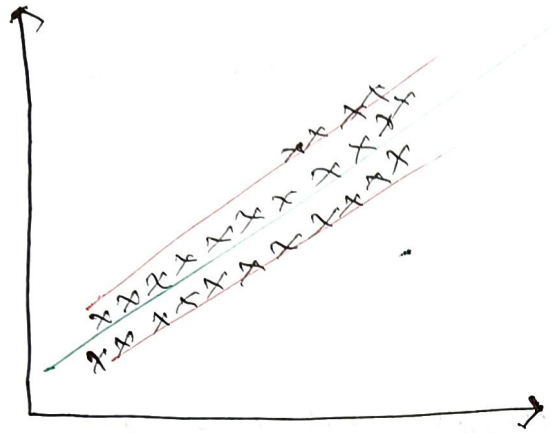
>

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

Decision - Trees - Regressor

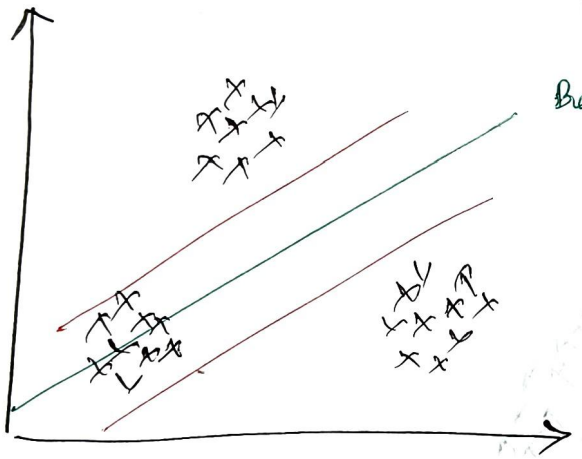
Q. why do we need DTR?

Ans



Best fit line

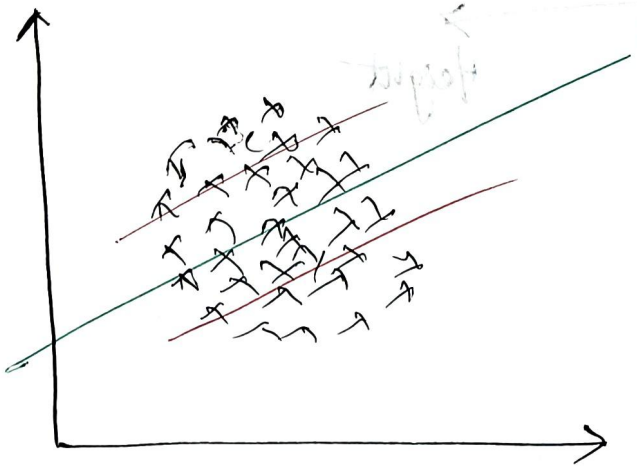
Good R-Square w.r.t linear Regression, SVM



Best fit line

Worst R-Square w.r.t linear Regression, SVM

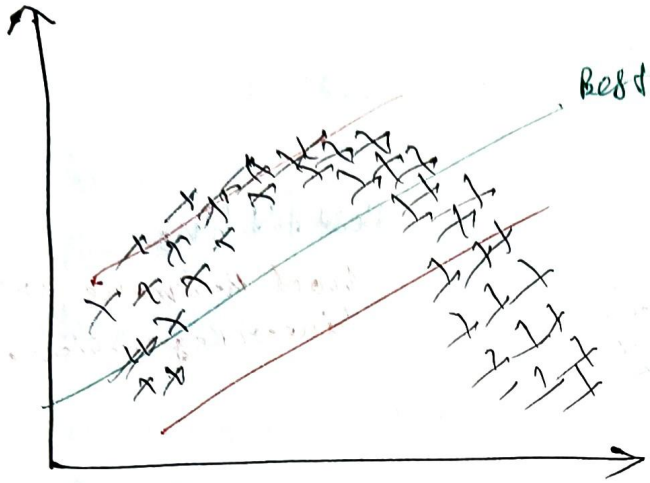
We Decision Tree



Best fit line

Worst R-Square w.r.t linear Regression, SVM

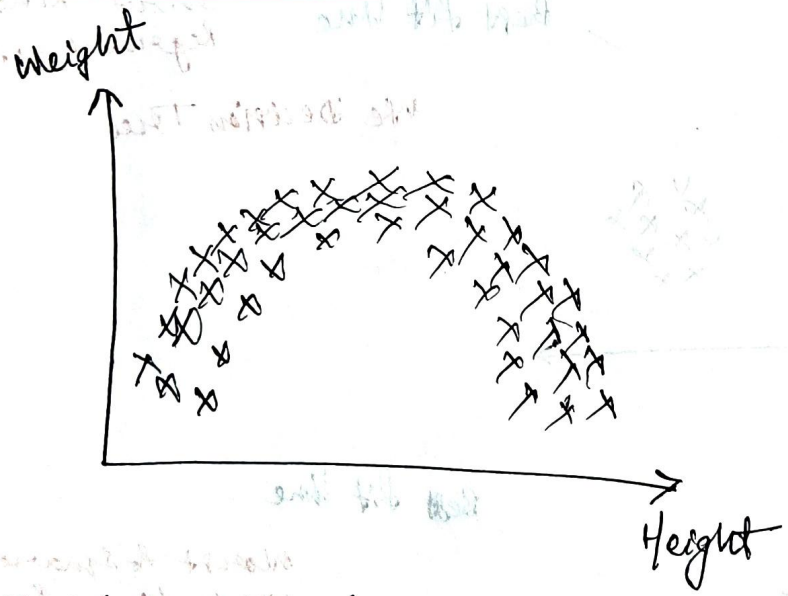
We Decision Tree



Worst R-Square with
Linear Regression, SVM
the Decision Tree

Note:- We can use SVM Kernel and can find data in higher dimension but that works more for classification mode. SVM Kernel will not work in properly in Regressor model.

Non-linear data



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

① Sort the values in ascending order and calculate adj. avg. value.

Average	Height	Weight
162.5	160	50
	165	65
167.5 ~~170~~	170	85
172.5 ~~175~~	175	70
177.5	180	90

classification problem steps in DT

① Entropy
② Gini Impurity
③ Information Gain

Regression problem steps in DT

① Mean
② MSE / MAE / RMSE
③ Reduction Variation

} we will select the ~~feature~~ threshold having less Mean, MSE, RMSE, MAE values.

Threshold = 162.5

Height [50, 65, 85, 170, 90], mean = 72

≤ 162.5 | > 162

calculate Means

[50] mean = 50

[65, 85, 170, 90] Mean = (65 + 85 + 70 + 90)/5 = 77.5

calculate MSE/MAE/RMSE

Variance = $MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$

deviation

avg the mean

Root Node

$$\text{Height[Variance or MSE]} = \frac{(72-50)^2 + (72-65)^2 + (72-85)^2 + (72-70)^2 + (72-90)^2}{5}$$

$$= 206$$

~~Height~~ [50]

$$\text{Variance[left]} = \frac{(50-50)^2}{1} = 0$$

[65, 85, 70, 90]

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

$$= 106.25$$

Note:
Variance and MSE are same.

Reduction in Variance

~~W.r.t~~ ~~W.left~~ {W.left = 1/5, Weight = 4/5}

$$\text{Reduction in Variance} = \text{Var(root)} - \sum_{i=1}^{n} w_i \, \text{var[child]}$$

w_i = weighted Average

$$\text{Reduction in Variance} = 206 - \left(\frac{1}{5} \times 0 + \frac{4}{5} \times 106.25 \right)$$

$$= 121$$

So, For threshold 162.5, Reduction in variance = 121.

→ choose threshold having largest Reduction in Variance value.

Let's assume,

for threshold 162.5, RV = 121 (calculated)
for threshold 167.5, RV = 142 (Assumed)
for threshold 172.5, RV = 106 (Assumed)
for threshold 177.5, RV = 131 (Assumed)

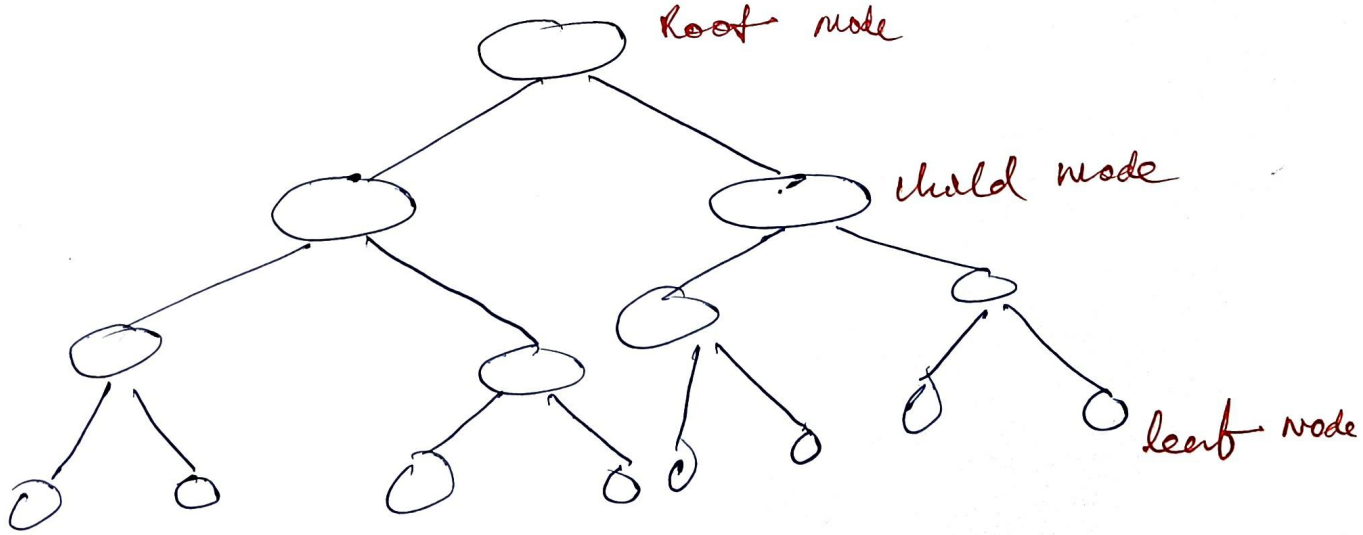
Hence, we will choose threshold as 162.5 because it has lowest Reduction in variance value.

→ In multiple features take the 1st features and build the trees using least RV value.

→ Gini impurity will always perform binary classification.

{ Blackbox Model → We can not visualize the mathematical calculation
whitebox Model → We can visualize the mathematical calculation behind the algorithm. }

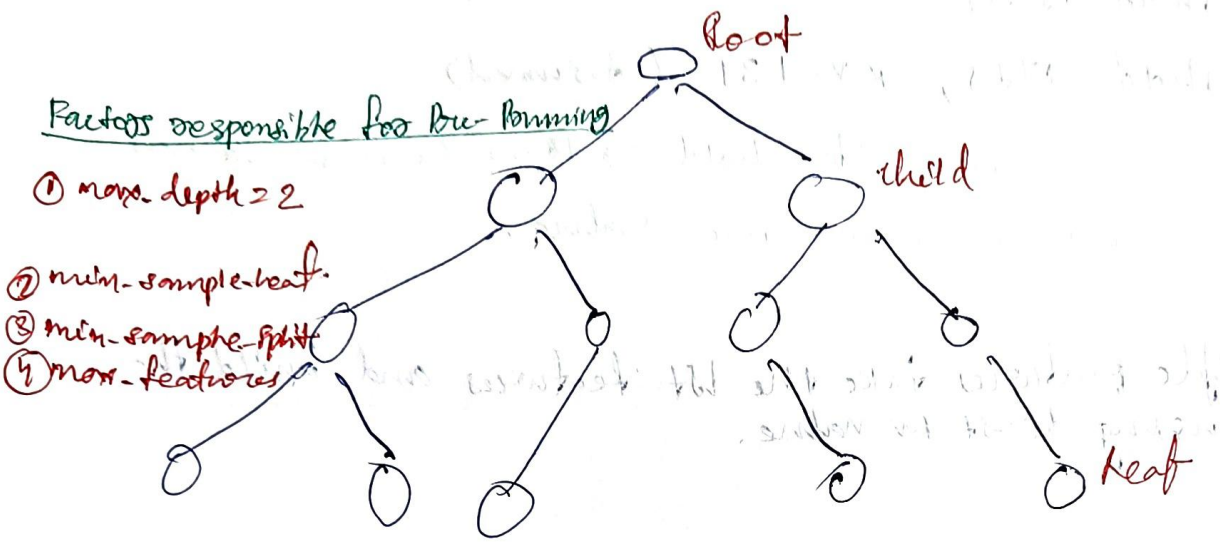
Pre-Pruning and Post Pruning



→ Pruning means cutting.
→ We do pruning to avoid overfitting in the model.

Pre-Pruning

→ While creating decision tree, decide whether to prune cut the trees or not.



Factors responsible for Pre-Pruning

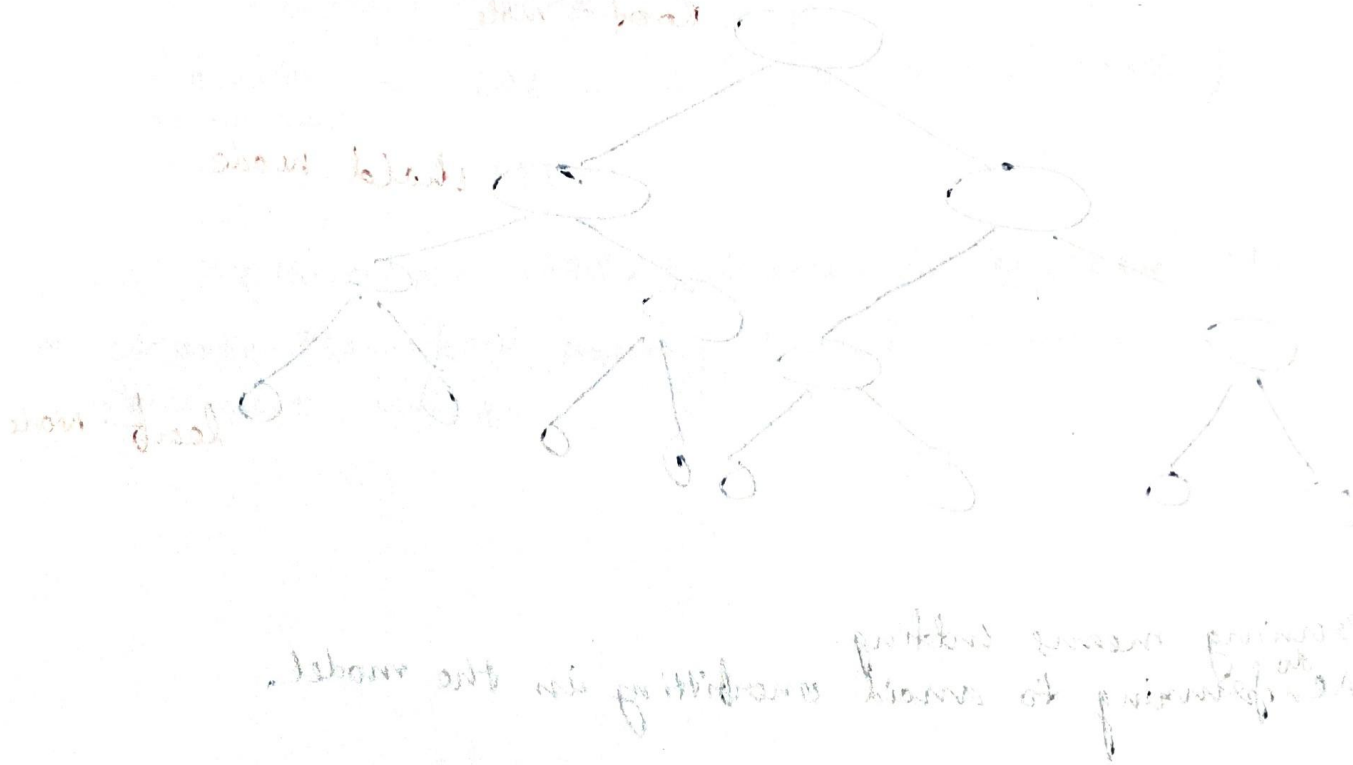
① max_depth = 2
② min_sample_leaf.
③ min_samples_split
④ max_features

→ Before creating DT, we decide factors for pre-pruning.

Hyperparameter for Pre-Pruning

① max_depth=2
② min_samples_leaf
③ min_samples_split
④ ~~max_factors~~ max_features

→ When Ratio will be high, we can perform pruning ~~at~~ till that point.



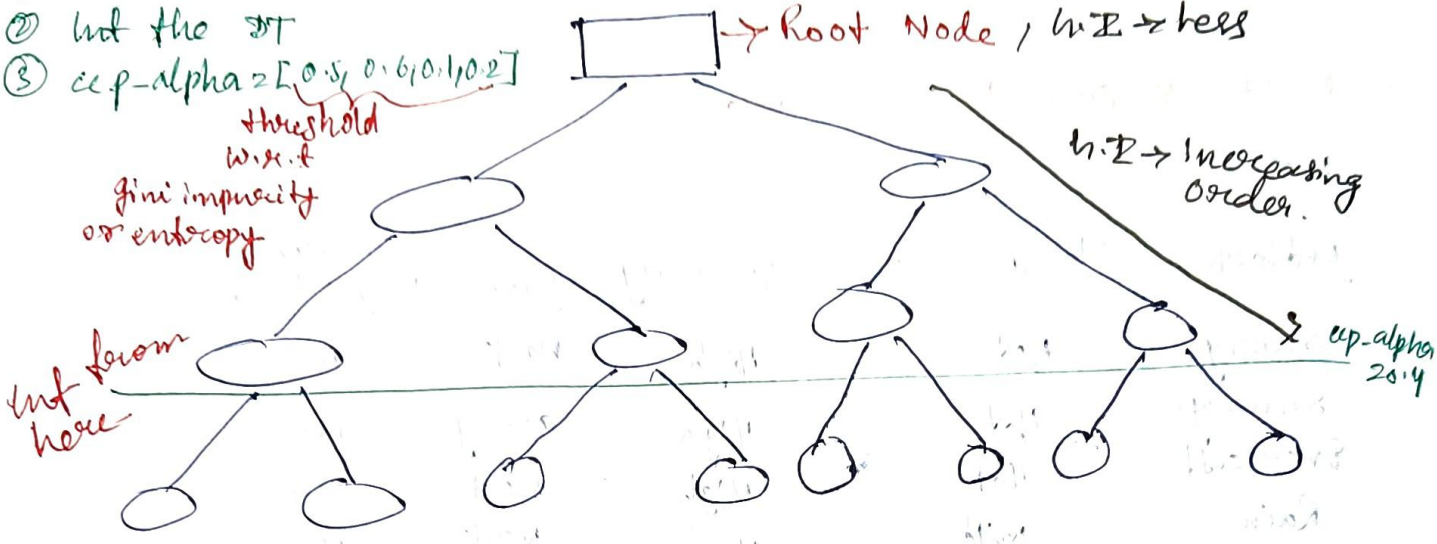
Post Pruning

→ First build the decision and then cut the trees.

Steps:

① Build the DT till last
② Cut the DT
③ ccp_alpha = [0.5, 0.6, 0.1, 0.2]

threshold w.r.t gini impurity or entropy



→ ccp_alpha → cost complexity Pruning

→ If ccp_alpha value doesn't match with h.I, then we will not go further.

→ ccp_alpha value & height of the DT

↓

Directly Proportional