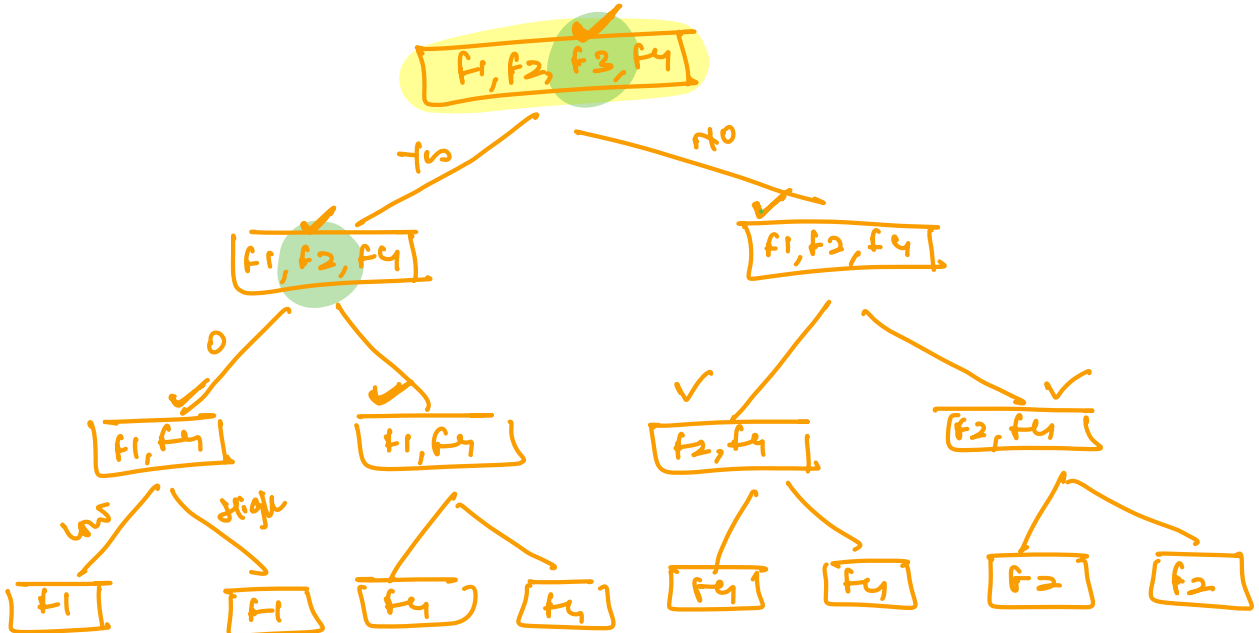
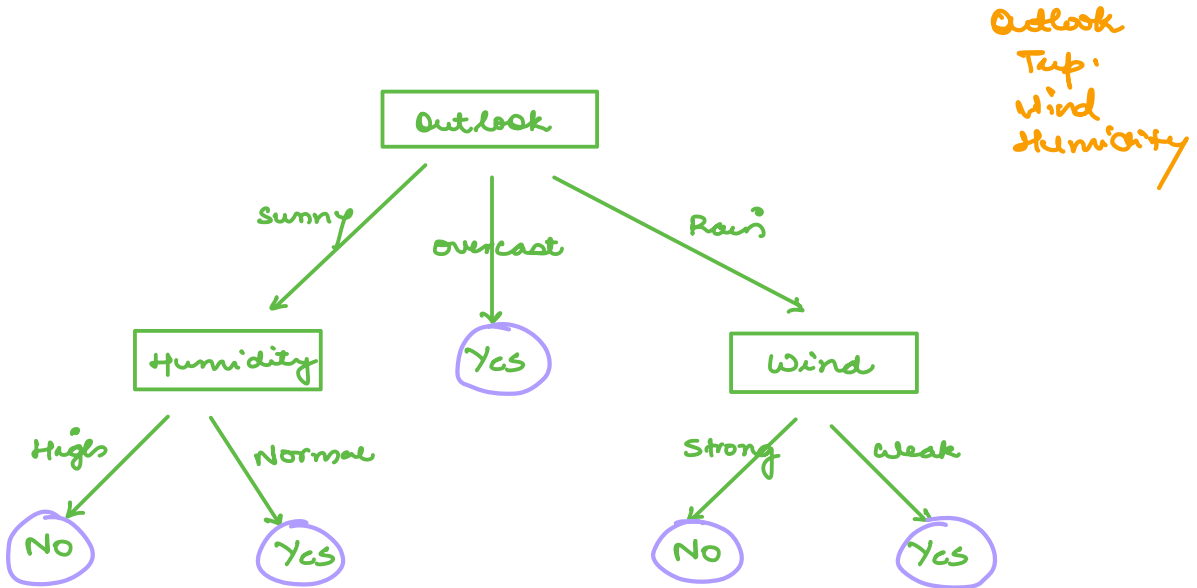
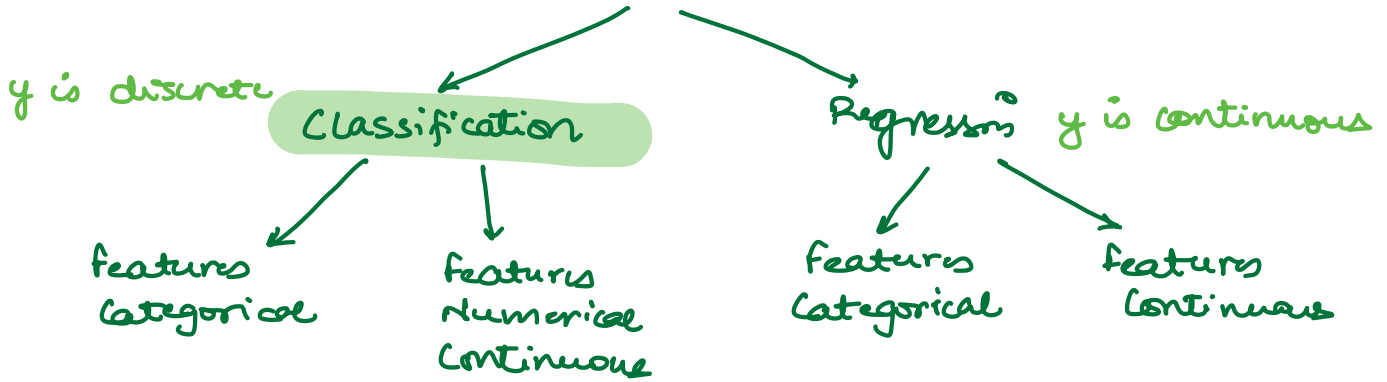
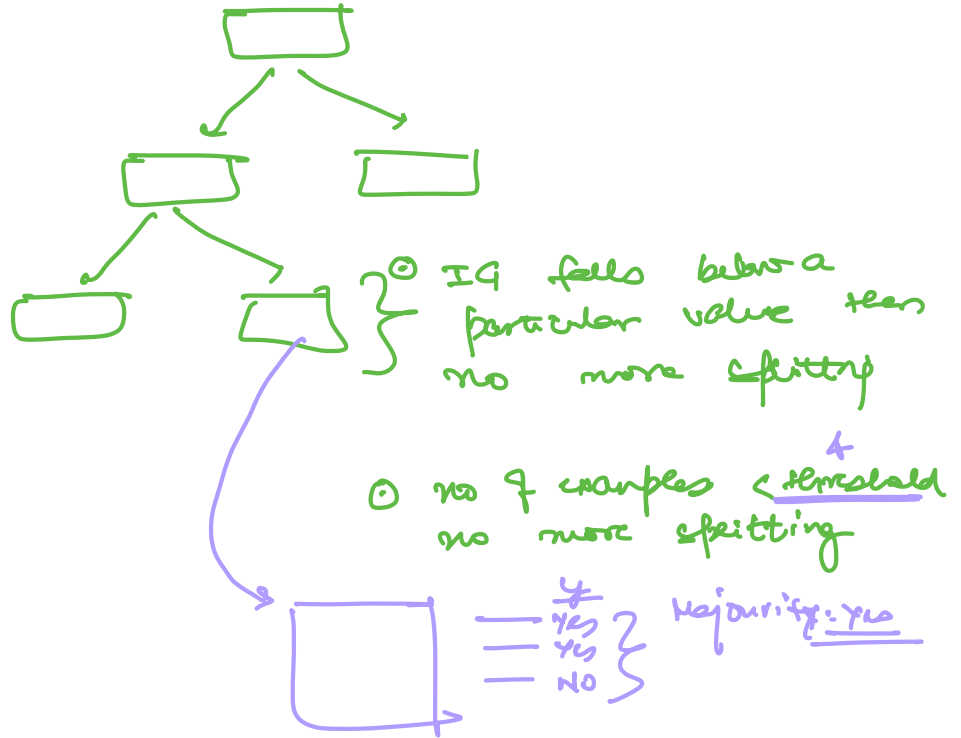


# DECISION TREES

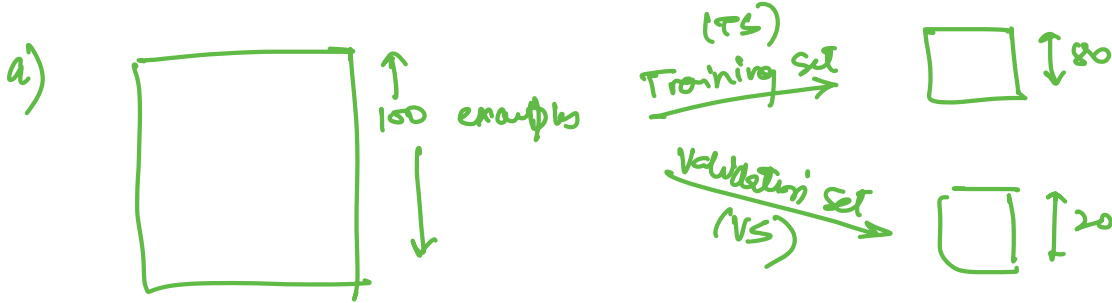


Overfitting: → DT has adapted itself too much to the training data.  
It is not generalizable for test data.

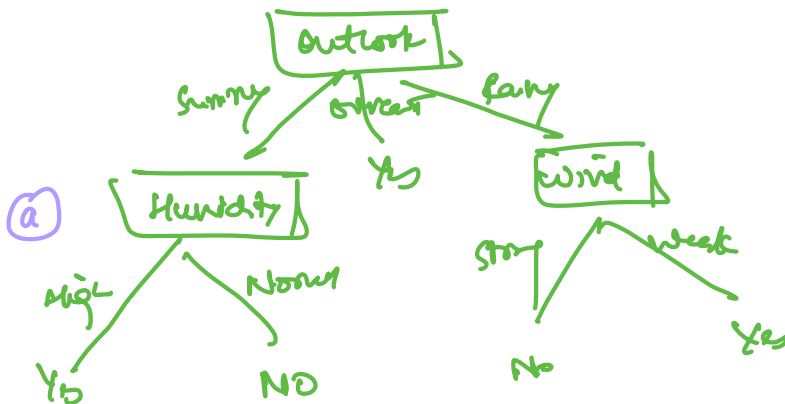
- Pre pruning (while create DT)



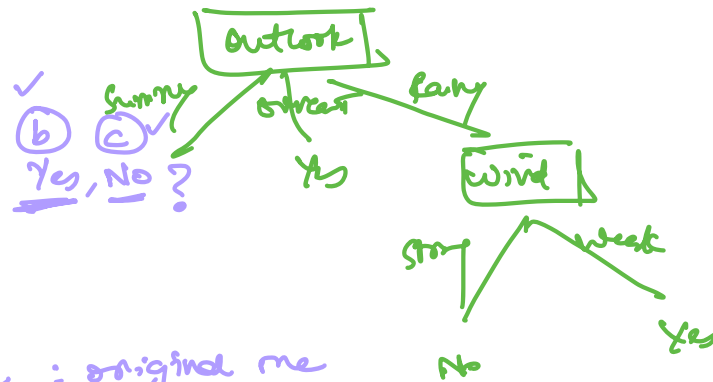
- Post Pruning



b) TS compute entire tree



c) Pick each node in BU manner, stupidly check if I prune this node will it improve my performance  $\rightarrow$  accuracy.



Validations:

$a > b$  and  $a > c$  : original me (no pruning)

$b > a$  and  $b > c$  : Yes

$c > a$  and  $c > b$  : No

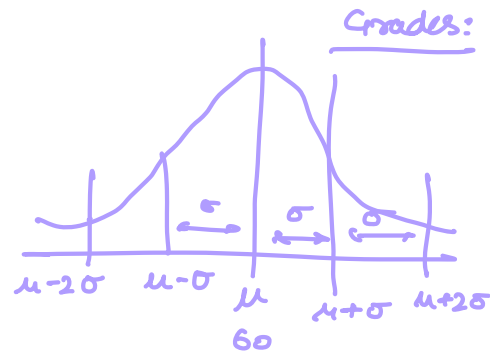
# CART

↳ Classification & Regression Tree.



Day	Outlook	Temp.	Humidity	Windy	Hours Played
D1	Rainy	Hot	High	False	25
D2	Rainy	Hot	High	True	30
D3	Overcast	Hot	High	False	46
D4	Sunny	Mild	High	False	45
D5	Sunny	Cool	Normal	False	52
D6	Sunny	Cool	Normal	True	23
D7	Overcast	Cool	Normal	True	43
D8	Rainy	Mild	High	False	35
D9	Rainy	Cool	Normal	False	38
D10	Sunny	Mild	Normal	False	46
D11	Rainy	Mild	Normal	True	48
D12	Overcast	Mild	High	True	52
D13	Overcast	Hot	Normal	False	44
D14	Sunny	Mild	High	True	30

Output is continuous value, Regression



Before Splitting

Day	Hours Played
D1	25
D2	30
D3	46
D4	45
D5	52
D6	23
D7	43
D8	35
D9	38
D10	46
D11	48
D12	52
D13	44
D14	30

$n = 14$

$$\text{Average} = \frac{25+30+46+45+52+23+43+35+38+46+48+52+44+30}{14}$$

$\bar{x} = 39.8$

Standard Deviation =  $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$  = 9.32 <sup>BS</sup>

Coefficient of Variation =  $CV = \frac{s}{\bar{x}} * 100$   
 = 23%  
 Stopping Condition

Pruning:

- ht of tree
- no. of examples
- CV

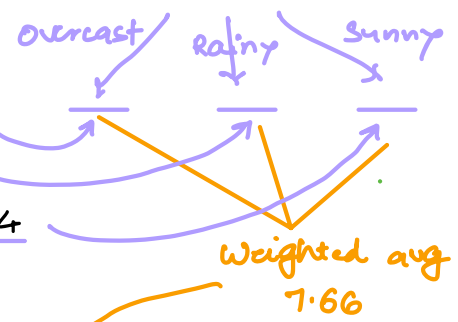
Attribute: Outlook

9.32 9.32

Outlook:  $D_1, D_2 \dots D_{14}$

		Hours Played (StDev)	Count
Outlook	Overcast	3.49	4
	Rainy	7.78	5
	Sunny	10.87	5
			14

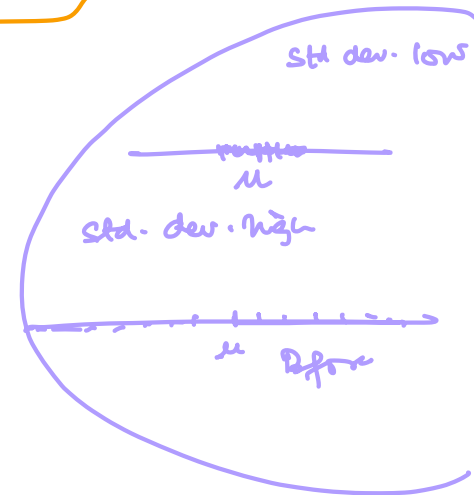
→  $D_3, D_7, D_2, D_{13}$   
 →  $D_1, D_2, D_8, 9, D_{11}$   
 →  $D_4, D_5, 16, D_{10}, D_{14}$



$$S(\text{Hours, Outlook}) = P(\text{Sunny}) * S(\text{Sunny}) + P(\text{Overcast}) * S(\text{Overcast}) + P(\text{Rainy}) * S(\text{Rainy})$$

$$= (4/14) * 3.49 + (5/14) * 7.78 + (5/14) * 10.87$$

AS = 7.66



Standard Deviation Reduction

(Similar to IG)

$$SDR(\text{Hours, Outlook}) = S(\text{Hours}) - S(\text{Hours, Outlook})$$

$$= 9.32 - 7.66 = 1.66$$

$$SDR = \frac{BS - AS}{\frac{1}{100}}$$

= High

		Hours Played (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
		SDR=1.66

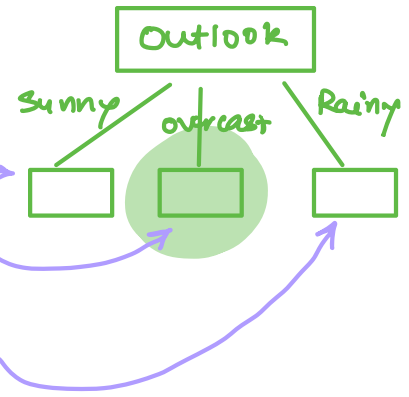
		Hours Played (StDev)
Temp.	Cool	10.51
	Hot	8.95
	Mild	7.65
		SDR= 0.48

		Hours Played (StDev)
Humidity	High	9.36
	Normal	8.37
		SDR=0.28

		Hours Played (StDev)
Windy	False	7.87
	True	10.59
		SDR=0.29

Attribute with longest SD is chosen.

Outlook	Temp	Humidity	Windy	Hours Played
Sunny	Mild	High	FALSE	45 → D4
	Sunny	Cool	FALSE	52 → D5
	Sunny	Cool	Normal	23 → D6
	Sunny	Mild	Normal	46 → D10
	Sunny	Mild	High	30 → D14
Overcast	Hot	High	FALSE	46 → D3
	Overcast	Cool	Normal	43 → D7
	Overcast	Mild	High	52 → D12
	Overcast	Hot	Normal	44 → D13
Rainy	Hot	High	FALSE	25 → D1
	Rainy	Hot	High	30 → D2
	Rainy	Mild	High	35 → D8
	Rainy	Cool	Normal	38 → D9
	Rainy	Mild	Normal	48 → D11



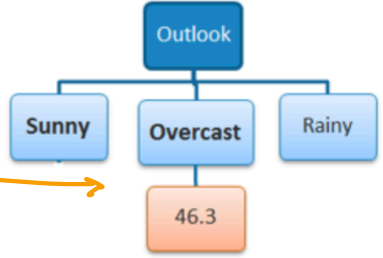
## Outlook - Overcast

		Hours Played (StDev)	Hours Played (AVG)	Hours Played (CV)	Count
Outlook	Overcast	3.49	46.3	8%	4
	Rainy	7.78	35.2	28%	5
	Sunny	10.87	39.2	28%	5

$D3 - 46$   
 $D7 - 43$   
 $D12 - 52$   
 $D13 - 44$   


---

 $\frac{185}{4} = 46.25$



## Outlook - Sunny

D5  
 D6  
 D9  
 D10

Temp	Humidity	Windy	Hours Played
Mild	High	FALSE	45
Cool	Normal	FALSE	52
Cool	Normal	TRUE	23
Mild	Normal	FALSE	46
Mild	High	TRUE	30
			S = 10.87
			AVG = 39.2
			CV = 28%

		Hours Played (StDev)	Count
Temp	Cool	14.50	2
	Mild	7.32	3

$SDR = 10.87 - ((2/5) * 14.5 + (3/5) * 7.32) = 0.678$

		Hours Played (StDev)	Count
Humidity	High	7.50	2
	Normal	12.50	3

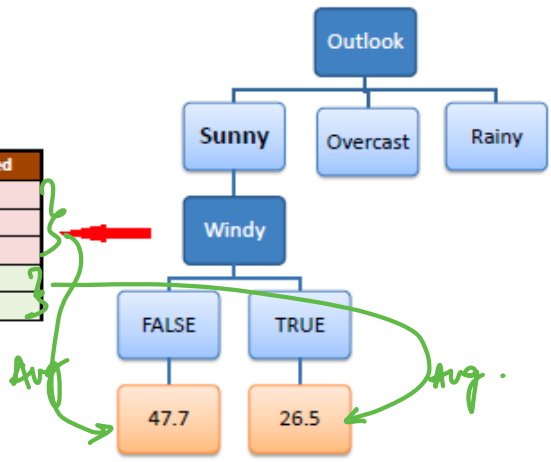
$SDR = 10.87 - ((2/5) * 7.5 + (3/5) * 12.5) = 0.370$

		Hours Played (StDev)	Count
Windy	False	3.09	3
	True	3.50	2

$SDR = 10.87 - ((3/5) * 3.09 + (2/5) * 3.5) = 7.62$

Temp	Humidity	Windy	Hours Played
Mild	High	FALSE	45
Cool	Normal	FALSE	52
Mild	Normal	FALSE	46
Cool	Normal	TRUE	23
Mild	High	TRUE	30

$\leq 3$  stop



# Outlook - Rainy

D1  
D2  
D8  
D9  
D11

Temp	Humidity	Windy	Hours Played
Hot	High	FALSE	25
Hot	High	TRUE	30
Mild	High	FALSE	35
Cool	Normal	FALSE	38
Mild	Normal	TRUE	48
			S = 7.78
			AVG = 35.2
			CV = 22%

		Hours Played (StDev)	Count
Temp	Cool	0	1
	Hot	2.5	2
	Mild	6.5	2

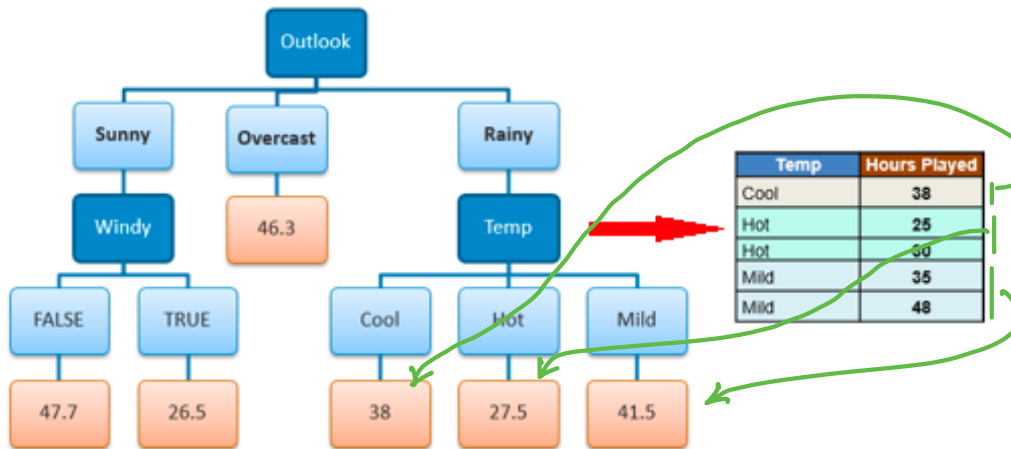
$$SDR = 7.78 - ((1/5)*0 + (2/5)*2.5 + (2/5)*6.5) = 4.18$$

		Hours Played (StDev)	Count
Humidity	High	4.1	3
	Normal	5.0	2

$$SDR = 7.78 - ((3/5)*4.1 + (2/5)*5.0) = 3.32$$

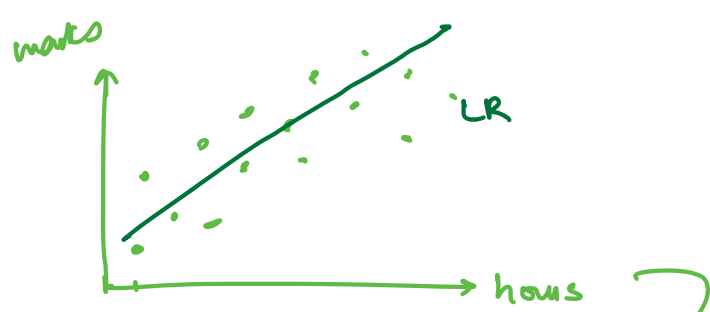
		Hours Played (StDev)	Count
Windy	False	5.6	3
	True	9.0	2

$$SDR = 7.78 - ((3/5)*5.6 + (2/5)*9.0) = 0.82$$

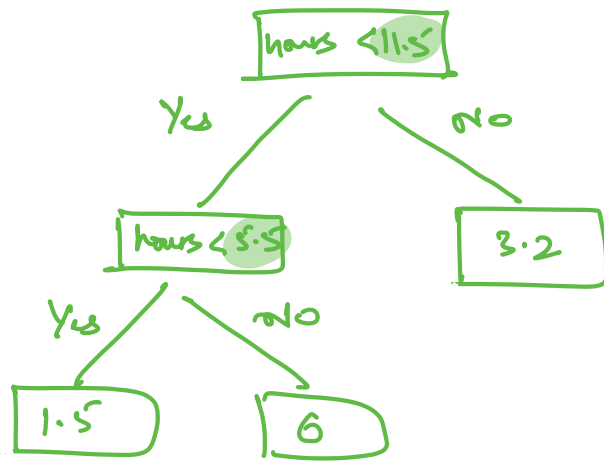
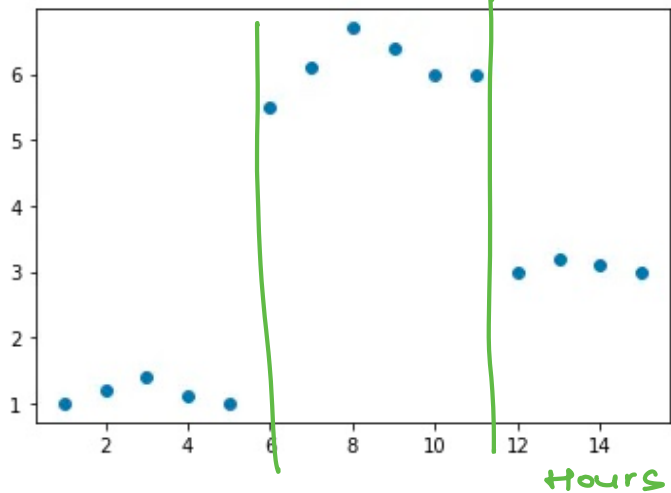


DT Regression → features discrete values *just discussed*

→ features have continuous values ✓



Marks



hours

marks

X	Y
1	1
2	1.2
3	1.4
4	1.1
5	1
6	5.5
7	6.1
8	6.7
9	6.4
10	6
11	6
12	3
13	3.2
14	3.1
15	3

1.5

2.5

3.5

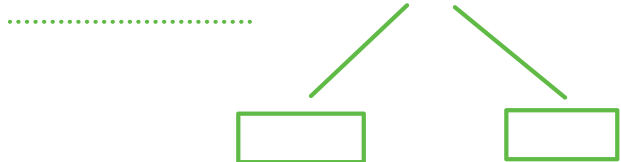
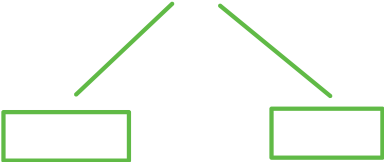
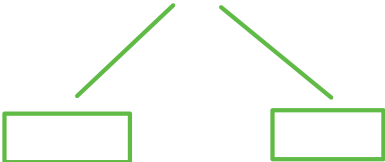
14 diff values

14.5

marks  $\leq 1.5$

marks  $\leq 2.5$

marks  $\leq 14.5$



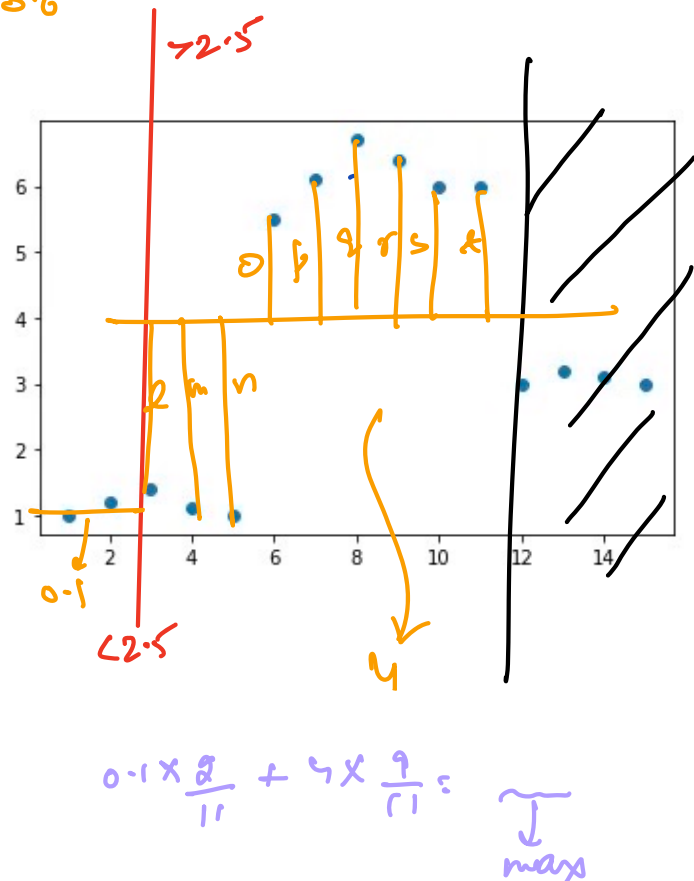
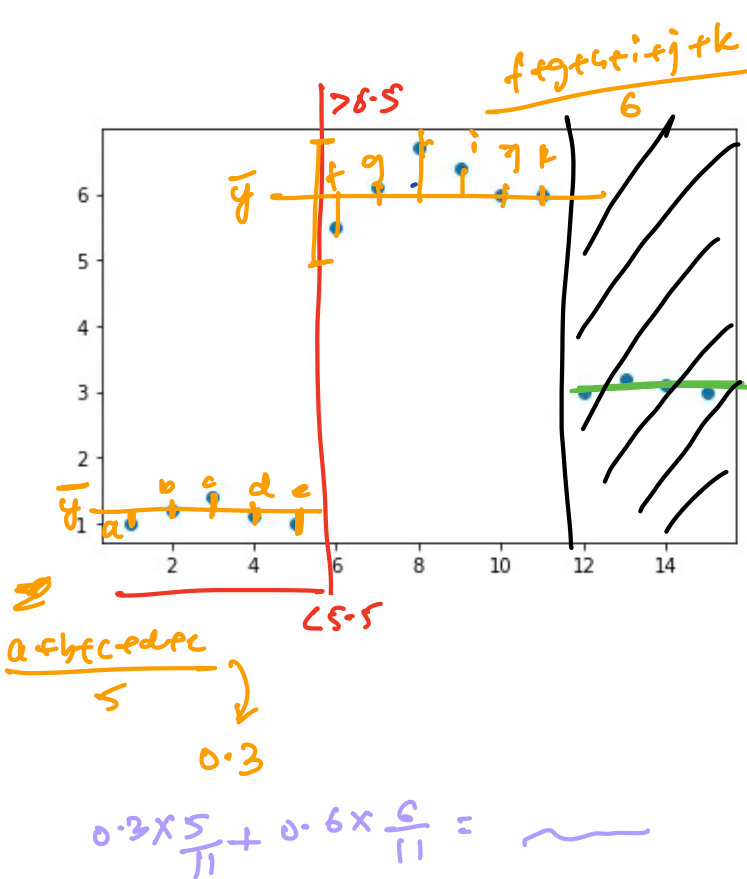


# Before Splitting

X	Y	$\bar{y}$	$(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2/n$
1	1	3.647	7.005	70.299	4.686
2	1.2		5.987		
3	1.4		5.048		
4	1.1		6.486		
5	1		7.005		
6	5.5		3.435		
7	6.1		6.019		
8	6.7		9.323		
9	6.4		7.581		
10	6		5.538		
11	6		5.538		
12	3		0.418		
13	3.2		0.2		
14	3.1		0.299		
15	3		0.418		

variance

error before splitting



error = 4.686

marks  $\leq$  5.5

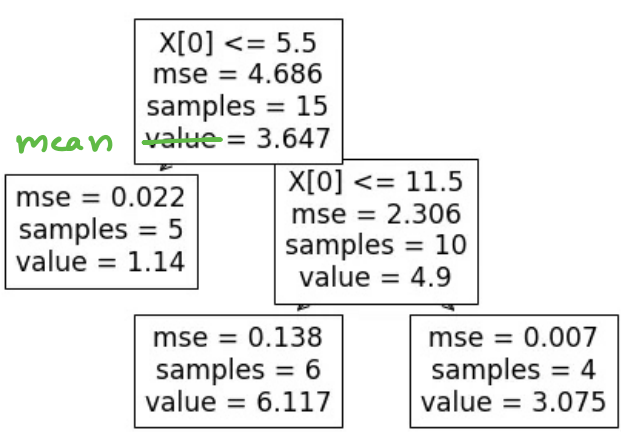
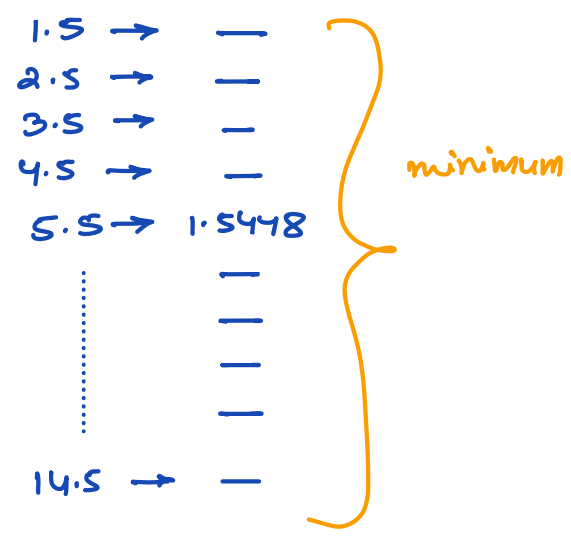
5  
mse  
0.0224

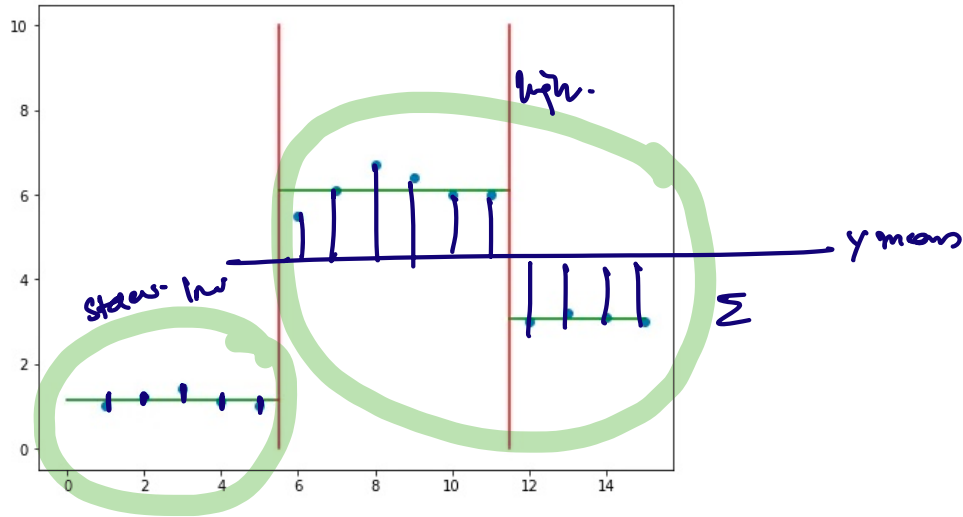
10  
mse  
2.306

X	Y	$\bar{y}$	$(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2/n$
1	1		0.0196		
2	1.2		0.0036		
3	1.4	1.14	0.0676	0.112	0.0224
4	1.1		0.0016		
5	1		0.0196		
6	5.5		0.36		
7	6.1		1.44		
8	6.7		3.24		
9	6.4	4.9	2.25	23.06	2.306
10	6		1.21		
11	6		1.21		
12	3		3.61		
13	3.2		2.89		
14	3.1		3.24		
15	3		3.61		

5.5

Weighted mean =  $\frac{5}{15} * 0.0224 + \frac{10}{15} * 2.306 = 1.5448$





References :

[https://www.youtube.com/watch?v=\\_wZ1Lo7bhGg](https://www.youtube.com/watch?v=_wZ1Lo7bhGg)

<https://www.youtube.com/watch?v=sLXtCwxg5kl>

<https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047>

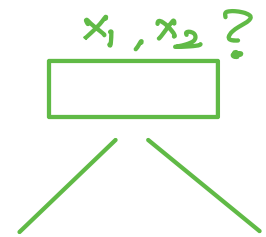
→ stop ? no. of samples

→ multiple features .

Multiple continuous features ?

$x_1$	$x_2$	$y$
2	10	
3	12	
4	13	
5	14	
6	15	
7	17	

↓ ↓  
continuous values ?



	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$
$x_1 \rightarrow$	2.5	3.5	4.5	5.5	6.5
$x_2 \rightarrow$	11	12.5	13.5	14.5	16
	$e_6$	$e_7$	$e_8$	$e_9$	$e_{10}$

} take minimum of all errors

$$x_1 \leq 4.5$$

$x_1, x_2?$

$x_1, x_2?$

2	10
3	12
4	13

	$c_1$	$c_2$
$c_3$	2.5	3.5
$c_4$	11	12.5