CART

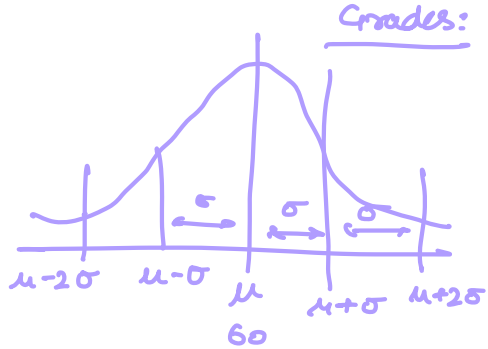
↳ Classification & Regression Tree.

IG Gini

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

Grades:



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

$n = 14$

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

$\bar{x} = 39.8$

Standard Deviation S(Hours) $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = 9.32$

Coefficient of Variation : $CV = \frac{S}{\bar{x}} * 100$

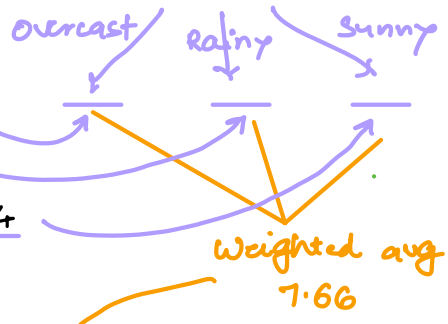
Stopping Condition = 23%.

→ ht of tree
→ no. of examples
→ CV

Attribute: Outlook

9.32

Outlook: D1, D2 .... D14



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

D3, D7, D 2, D13

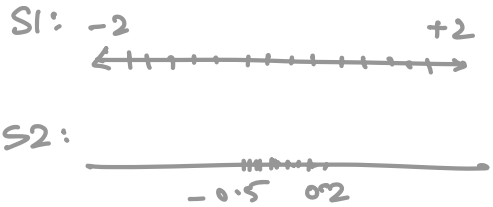
D1, D2, D8, 9, D11

D4, D5, 1 6, D10, D14

$$S(Hours, Outlook) = P(Sunny)*S(Sunny) + P(Overcast)*S(Overcast) + P(Rainy)*S(Rainy)$$
$$= (4/14)*3.49 + (5/14)*7.78 + (5/14)*10.87$$
$$= 7.66$$

Standard Deviation Reduction (Similar to IG)

$$SDR(Hours, Outlook) = S(Hours) - S(Hours, Outlook)$$
$$= 9.32 - 7.66 = 1.66$$



		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 largest SD is chosen.

Outlook

Sunny

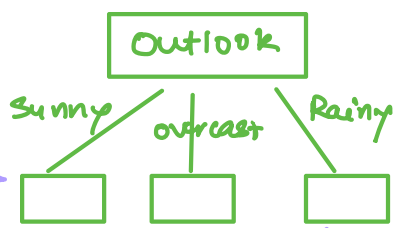
Outlook	Temp	Humidity	Windy	Hours Played	
Sunny	Mild	High	FALSE	45	D4
Sunny	Cool	Normal	FALSE	52	D5
Sunny	Cool	Normal	TRUE	23	D6
Sunny	Mild	Normal	FALSE	46	D10
Sunny	Mild	High	TRUE	30	D14

Overcast

Overcast	Hot	High	FALSE	46	D3
Overcast	Cool	Normal	TRUE	43	D7
Overcast	Mild	High	TRUE	52	D12
Overcast	Hot	Normal	FALSE	44	D13

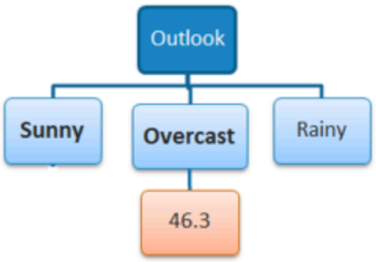
Rainy

Rainy	Hot	High	FALSE	25	D1
Rainy	Hot	High	TRUE	30	D2
Rainy	Mild	High	FALSE	35	D8
Rainy	Cool	Normal	FALSE	38	D9
Rainy	Mild	Normal	TRUE	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	22%	5
	Sunny	10.87	39.2	28%	5



Outlook - Sunny

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

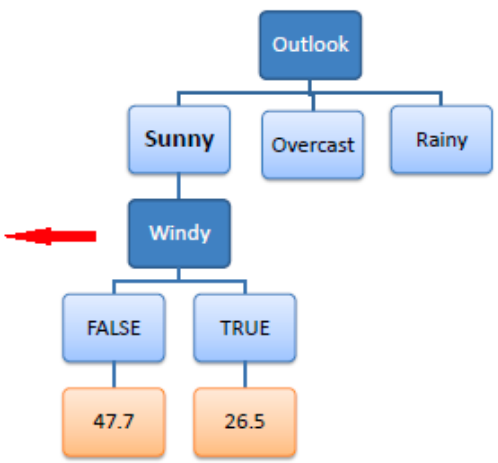
SDR = 10.87-((2/5)*7.5 + (3/5)*12.5) = **0.370**

		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

<=3 stop



Outlook - Rainy

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

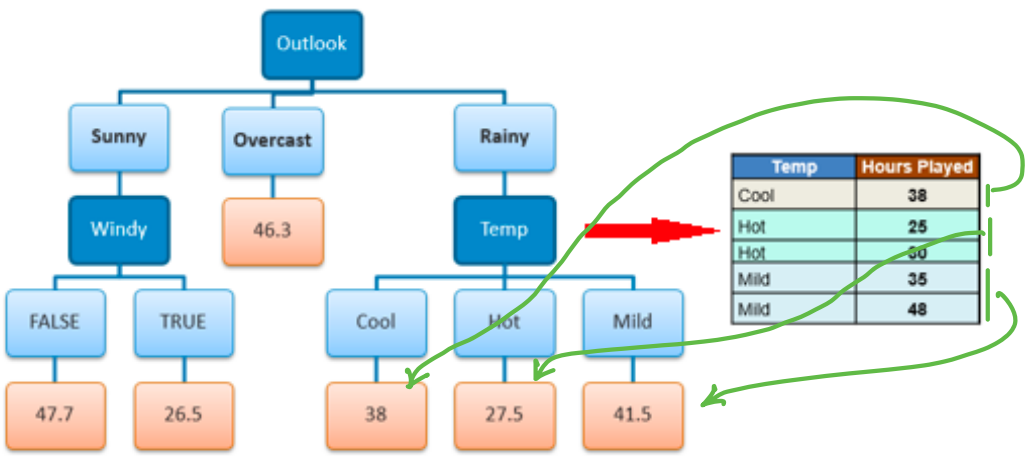
SDR = 7.78 - ((1/5)*0+(2/5)*2.5 + (2/5)*6.5) = **4.18**

		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

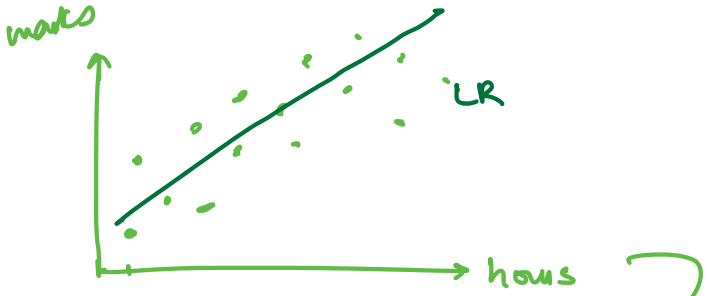
SDR = 7.78 - ((3/5)*5.6 + (2/5)*9.0) = **0.82**

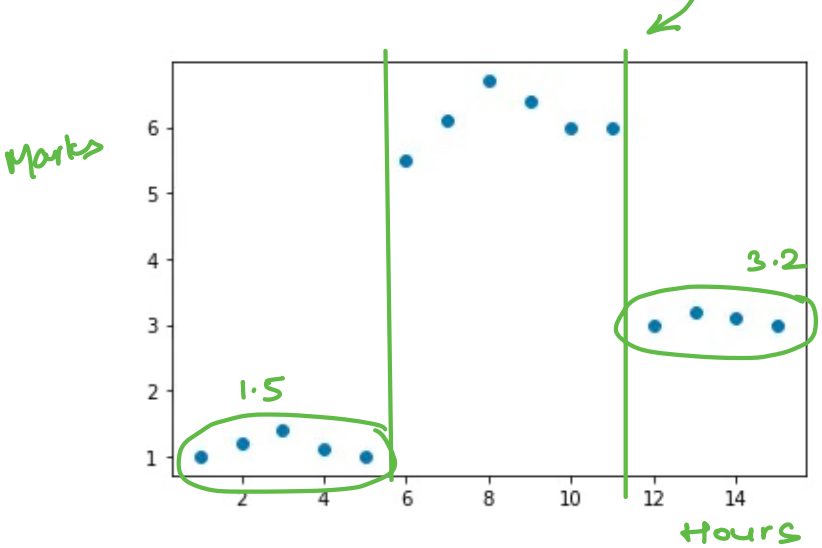


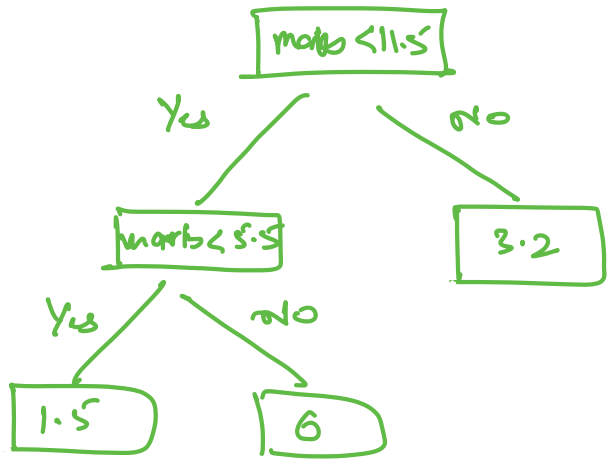
Reference : https://saedsayad.com/decision_tree_reg.htm
https://www.youtube.com/watch?v=1i_V-2spSKs
https://www.youtube.com/watch?v=cxz53CU0y_4

DT Regression → features discrete values

→ features have continuous values.



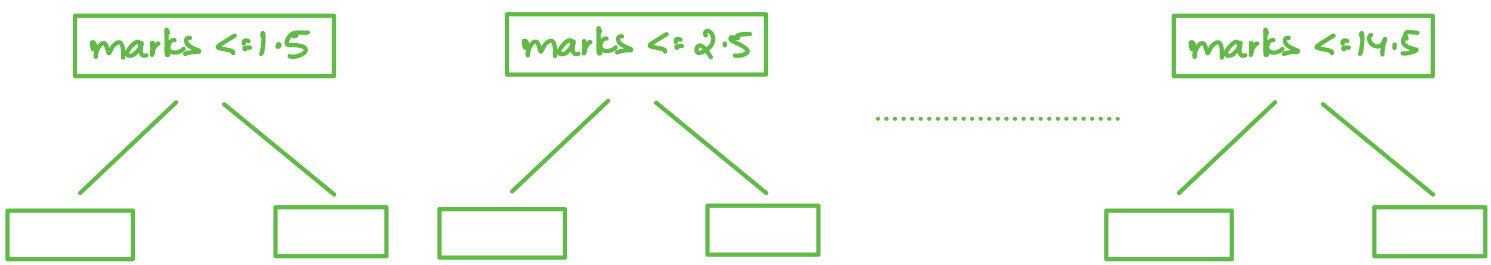




hours → X, marks → Y

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
14.5
14 diff values

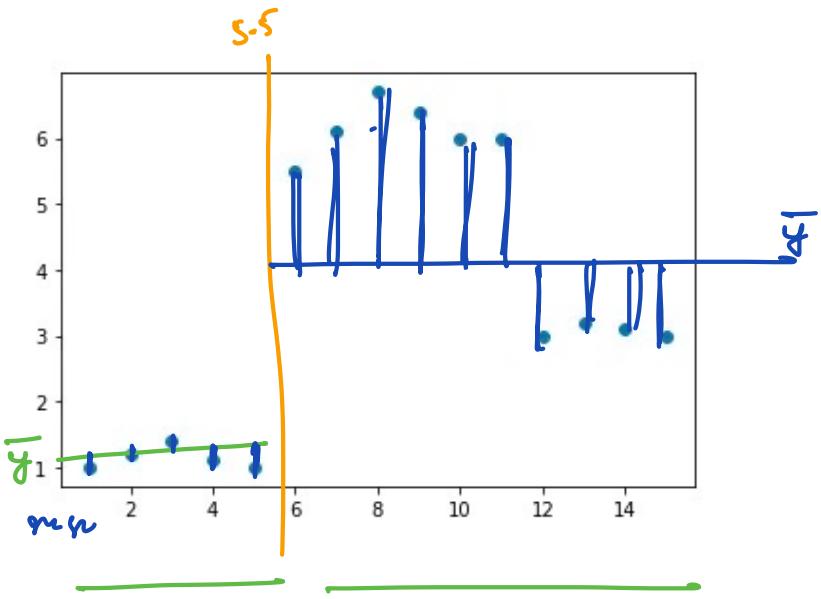


Before Splitting

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

$\bar{y}$	$(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2/n$
	7.005		
	5.987		
	5.048		
	6.486		
	7.005		
3.647	3.435	70.299	4.686
	6.019		
	9.323		
	7.581		
	5.538		
	5.538		
	0.418		
	0.2		
	0.299		
	0.418		

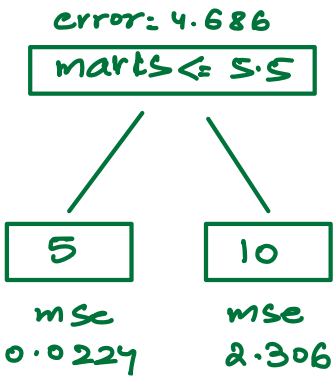
error before splitting



$\bar{y}_1$

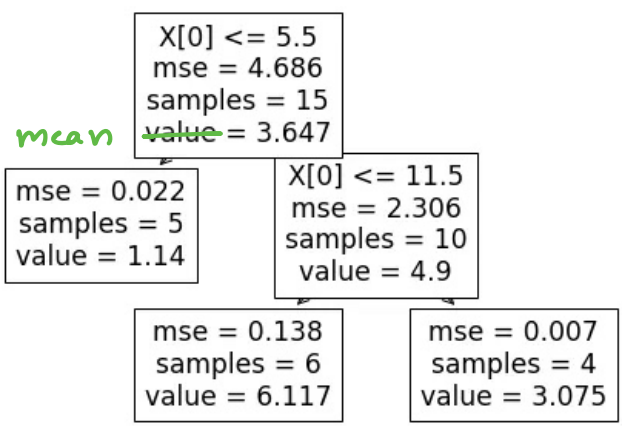
$\bar{y}_2$

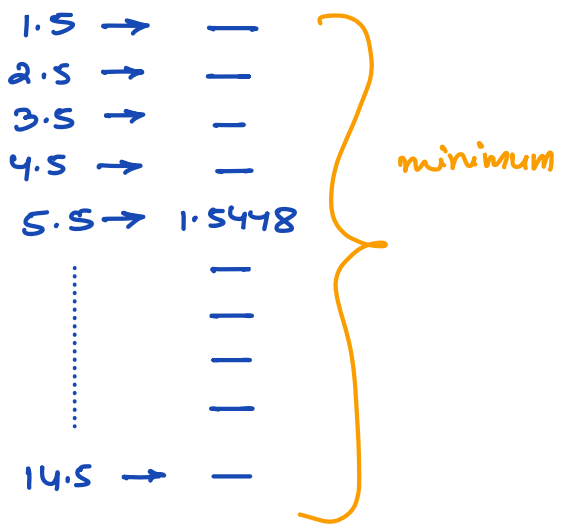
$$\frac{\Sigma(y-\bar{y})^2}{n}$$

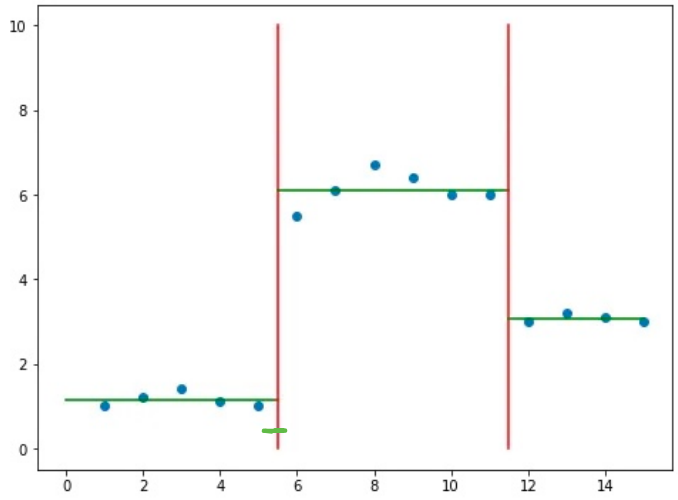


X	Y	$\bar{y}$	$(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2/n$
1	1	1.14	0.0196	0.112	0.0224
2	1.2		0.0036		
3	1.4		0.0676		
4	1.1		0.0016		
5	1		0.0196		
6	5.5	4.9	0.36	23.06	2.306
7	6.1		1.44		
8	6.7		3.24		
9	6.4		2.25		
10	6		1.21		
11	6		1.21		
12	3		3.61		
13	3.2		2.89		
14	3.1		3.24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

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=sLXtCwxg5kI
https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047

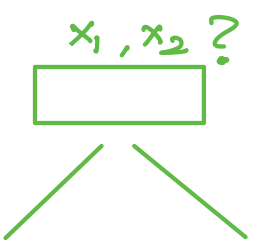
→ stop ? no. of samples

→ multiple features.

Multiple features ? Continuous

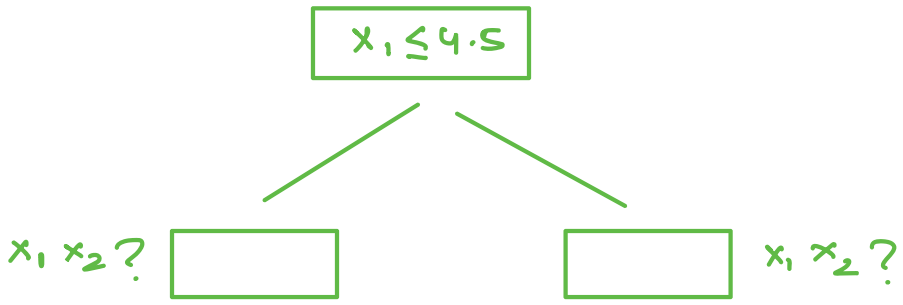
X_1	X_2	y
2	10	
3	12	
4	13	
5	14	
6	15	
7	17	

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 ?