Decision Tree with Continuous Attributes (Information Gain)

Discrete Y value

x: Temperature	40	48	60	72	80	90
y: PlayTennis	No	No	Yes	Yes	Yes	No

→ Continuous values

- Sort the examples in increasing order of continuous attribute (temperature here)
- Identify when there is a change in class labels.

Temperature	40	48	60	72	80	90
PlayTennis	No	No	Yes	Yes	Yes	No

- For every change, calculate the average of boundary values.

Temperature	40	48	60	72	80	90
PlayTennis	No	No	Yes	Yes	Yes	No

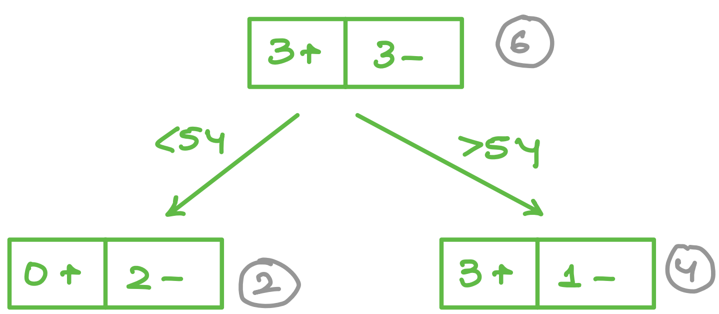
$\frac{48+60}{2} = 54$ $\qquad$ $\frac{80+90}{2} = 85$

- When you have more than one threshold value, which one to select?

Calculate MI through 54 and 85, and pick the one which gives maximum MI.

Threshold: 54

Temperature	40	48	60	72	80	90
PlayTennis	No	No	Yes	Yes	Yes	No



$$H(y) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1 \quad \Big\} \text{ Before split (BS)}$$

$$H(y \mid x_j < 54) = 0$$

$$H(y \mid x_j > 54) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \quad \Big\} \text{ After split (AS)}$$

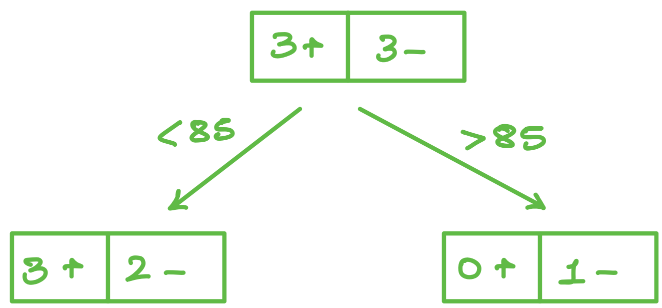
$$H(y \mid x_j = temp\ 54) = \frac{2}{6} * 0 + \frac{4}{6} * 0.8113 \quad \Big\} \text{ weighted avg.}$$

$$MI(y \mid x_j = temp\ 54) = \overbrace{H(y)}^{B.S.} - \overbrace{H(y \mid x_j = temp\ 54)}^{A.S.}$$

$$= 1 - \left(\frac{2}{6} * 0 + \frac{4}{6} * 0.8113\right) = 0.4591$$

Threshold: 85

Temperature	40	48	60	72	80	90
PlayTennis	No	No	Yes	Yes	Yes	No



$$H(y) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$$

$$H(y \mid x_j < 85) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

$$H(y \mid x_j > 85) = 0$$

$$H(y \mid x_j = temp\ 85) = \frac{5}{6} * 0.971 + \frac{1}{6} * 0$$

$$MI(y \mid x_j = temp\ 85) = H(y) - H(y \mid x_j = temp\ 85)$$

$$= 1 - \left(\frac{5}{6} * 0.971 + \frac{1}{6} * 0\right) = 0.1908$$

$MI(y|x_j = temp54) = 0.4591$

$MI(y|x_j = temp85) = 0.1908$

IG maximized
54 is better boundary.

Temperature	40	48	60	72	80	90
PlayTennis	No	No	Yes	Yes	Yes	No

Overfitting

<54

40	48
No	No

>54

60	72	80	90
Yes	Yes	Yes	No

Continuous values with Gini Index

Annual Income	Label
60	No
70	No
75	No
85	Yes
90	Yes
95	Yes
100	No
120	No
125	No
220	No

- Arrange continuous valued attribute in increasing order.
- Select split point
  - Change from one label to another label.

Annual Income	Label	Split Point
60	No	
70	No	
75	No	<80 >=80
85	Yes	
90	Yes	
95	Yes	<97.5 >=97.5
100	No	
120	No	
125	No	
220	No	

avg

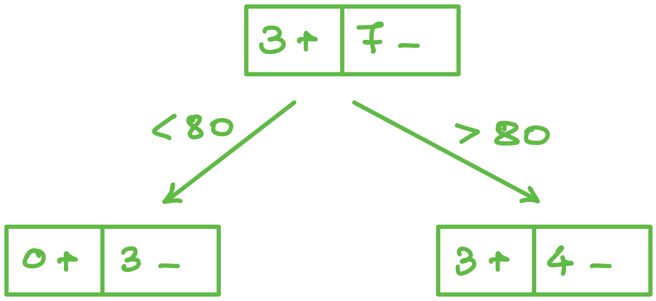
avg

mean

mean

– Take split point : 80

Annual Income	Label
60	No
70	No
75	No
85	Yes
90	Yes
95	Yes
100	No
120	No
125	No
220	No



$$Gini(y \mid x_j < 80) = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2 \right] = 0$$

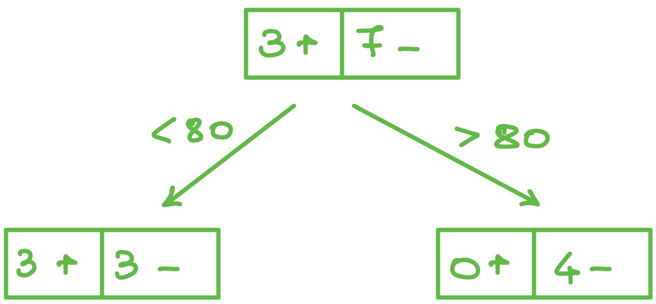
$$Gini(y \mid x_j > 80) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2 \right] = 0.4897$$

$$Gini(y \mid x_j \; 80) = \frac{3}{10} * 0 + \frac{7}{10} * 0.4897$$

$$= 0.3427$$

Take split Point : 97.5

Annual Income	Label
60	No
70	No
75	No
85	Yes
90	Yes
95	Yes
100	No
120	No
125	No
220	No



$$Gini(y \mid x_j < 97.5) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2 \right] = 0.5$$

$$Gini(y \mid x_j > 97.5) = 1 - \left[\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2 \right] = 0$$

$$Gini(y \mid x_j \; 97.5) = \frac{6}{10} * 0.5 + \frac{4}{10} * 0 = 0.3$$

$Gini(y|x_j\ 80) = 0.3427$

$Gini(y|x_j\ 97.5) = 0.3$

97.5 has smaller value. 97.5 is better splitting point.

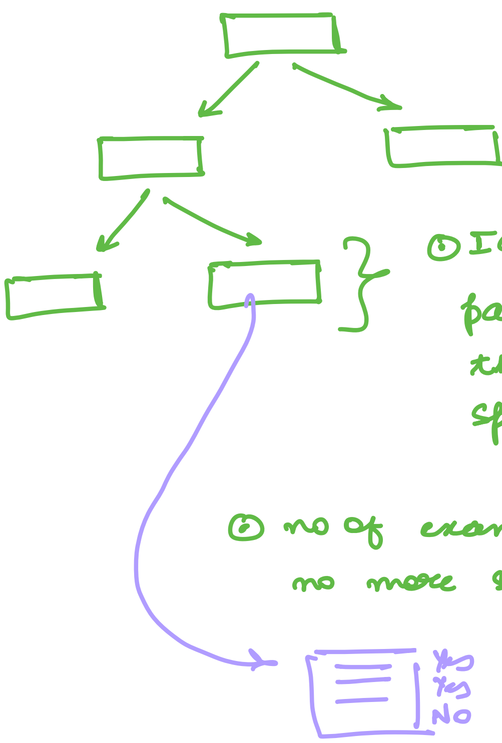
DECISION TREES

- Classification (y is discrete)
  - features Categorical
  - features Numerical (continuous)
- Regression (y is continuous)
  - features Categorical
  - features Continuous

Overfitting:

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

- Pre-Pruning
(while creating DT)



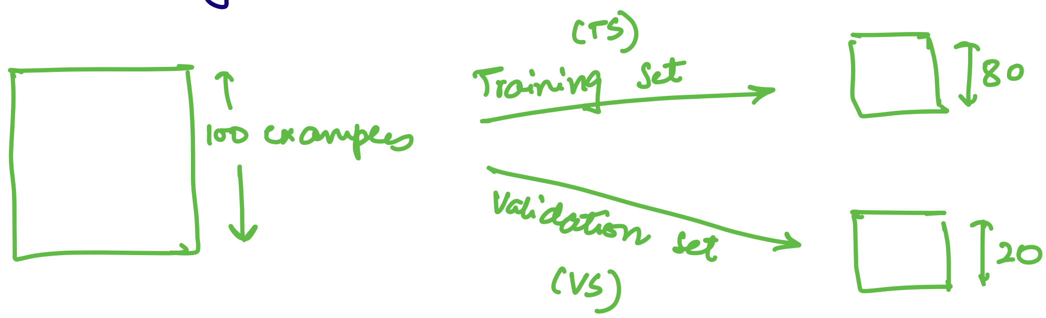
① IG falls below a particular value than no more splitting.

② no of examples < threshold no more splitting

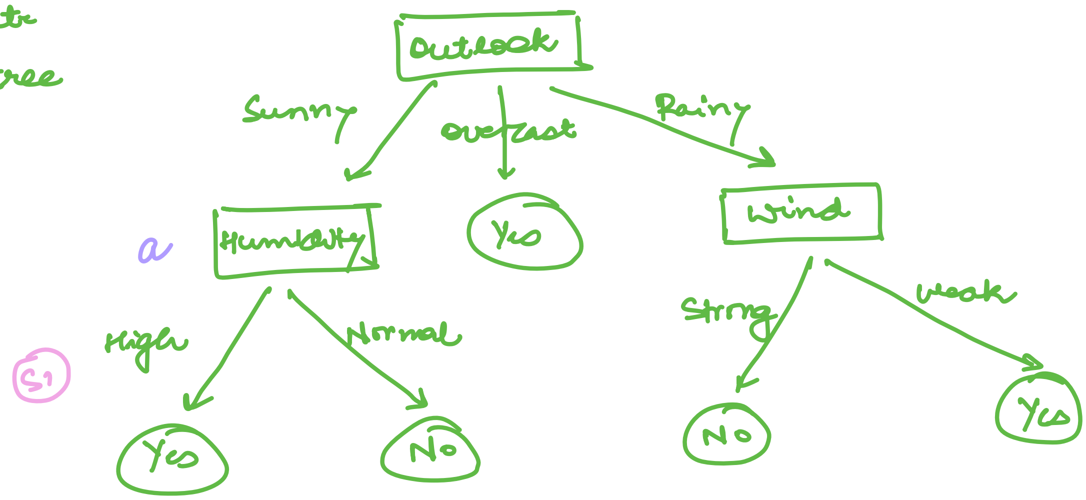
majority: Yes

– Post Pruning

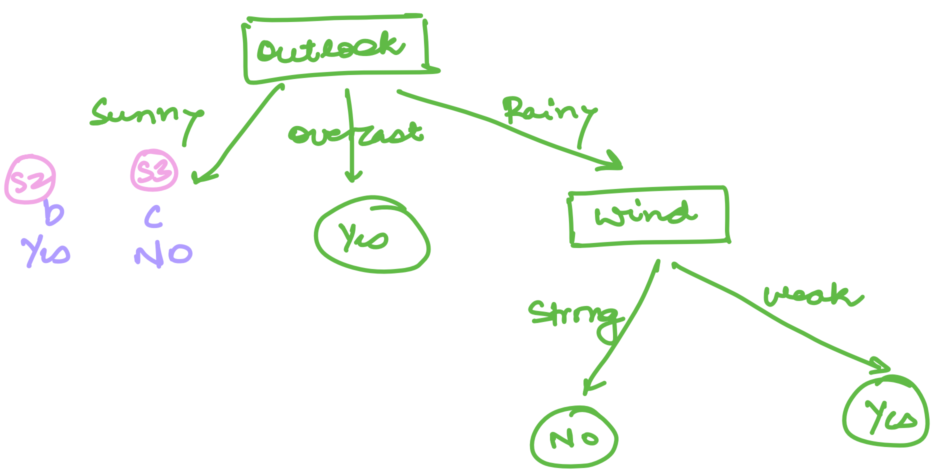
a)



b) TS compute entire tree



ⓒ Pick each node in BU manner, $\boxed{\text{Humidity}}$ check if I prune this node will it improve my $\underbrace{\text{performance}}_{\text{accuracy}}$



Validation:

a > b and a > c : original one (no pruning)

b > a and b > c : Yes

c > a and c > b : NO

CART

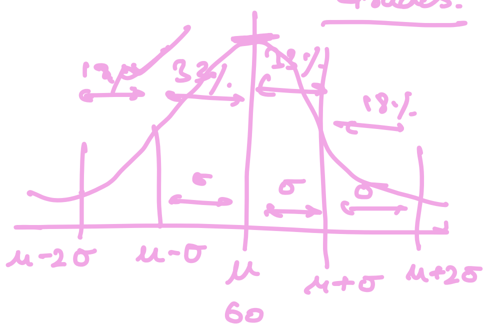
↳ Classification & Regression Tree.

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

$\bar{x} = 39.8$

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

$\underbrace{\text{Coefficient of Variation}}_{\text{Stopping Condition}} = CV = \dfrac{S}{\bar{x}} * 100$

$= 23\%$

Pruning:

→ ht of tree

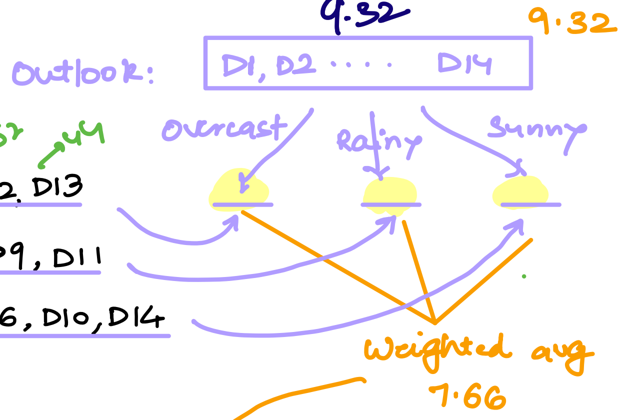
→ no. of examples

→ CV

Attribute: Outlook

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

46 43 52 44
→ D3, D7, D12, D13
→ D1, D2, D8, D9, D11
→ D4, D5, D6, 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

AS = 7.66

(Similar to IG)

Standard Deviation Reduction

SDR(Hours , Outlook) = **S**(Hours) – **S**(Hours, Outlook)

= 9.32 – 7.66 = 1.66

SDR = BS – AS

~~low~~

= High

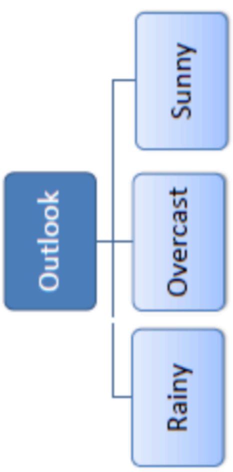
		Hours Played (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
	SDR=1.66	

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

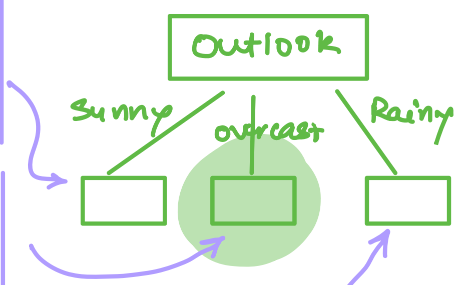
		Hours Played (StDev)
Humidity	High	9.36
	Normal	8.37
	SDR=0.28	

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

Attribute with largest SD is chosen.



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



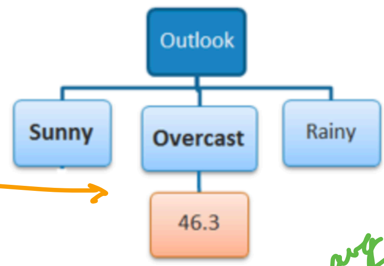
Outlook - Overcast

D3 — 46
D7 — 43
D12 — 52
D13 — 44

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

		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

≤ 4 stop



Outlook - Sunny

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

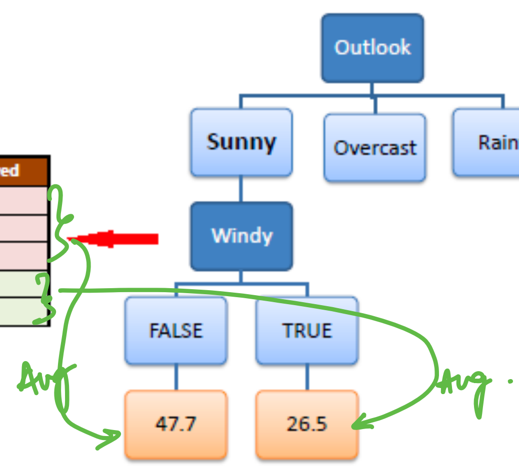
SDR = 10.87-((2/5)*7.5 + (3/5)*12.5) = 0.370

		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

<=4 stop



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

Outlook - Rainy

	Temp	Humidity	Windy	Hours Played
D1	Hot	High	FALSE	25
D2	Hot	High	TRUE	30
D8	Mild	High	FALSE	35
D9	Cool	Normal	FALSE	38
D11	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

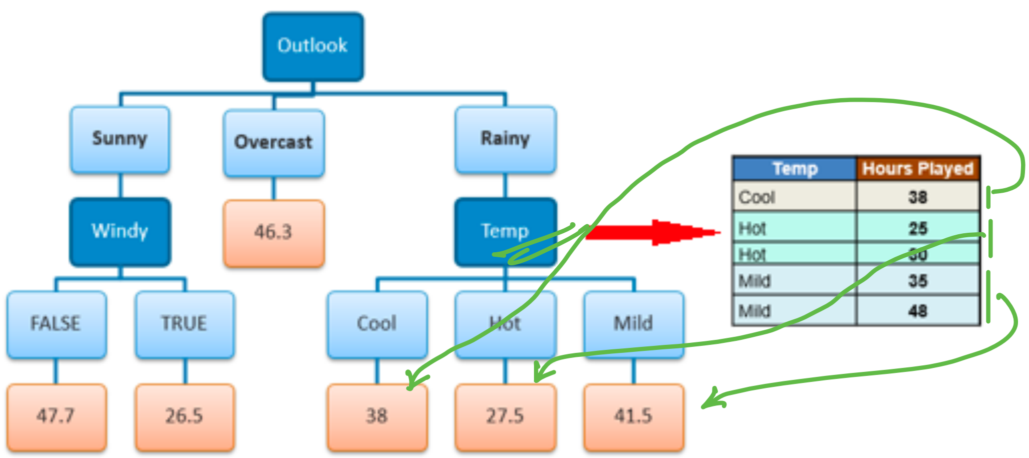
SDR = 7.78 - ((1/5)*0+(2/5)*2.5 + (2/5)*6.5) = **4.18**

		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**

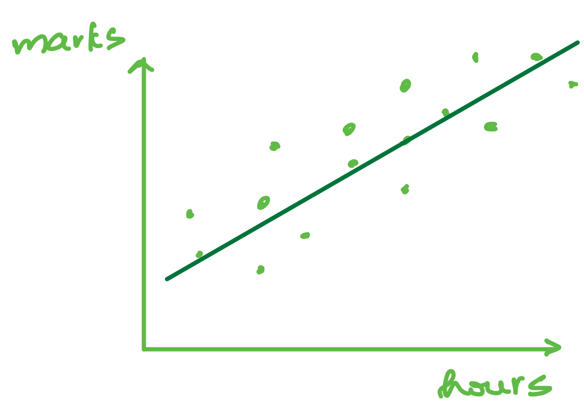
		Hours Played (StDev)	Count
Windy	False	5.6	3
	True	9.0	2

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

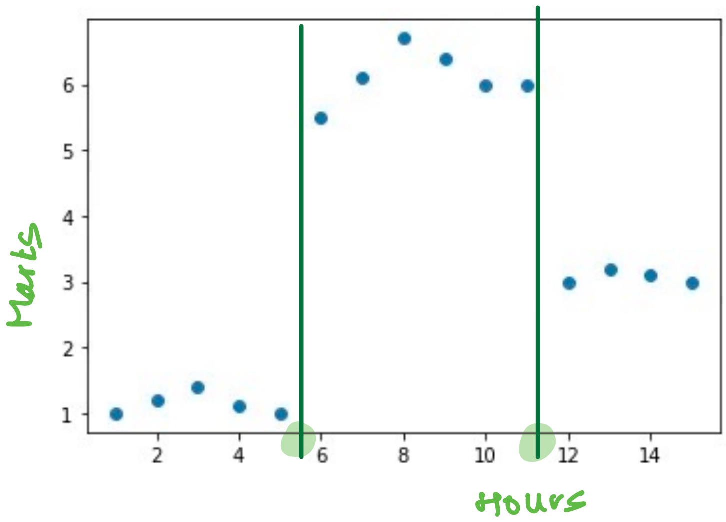


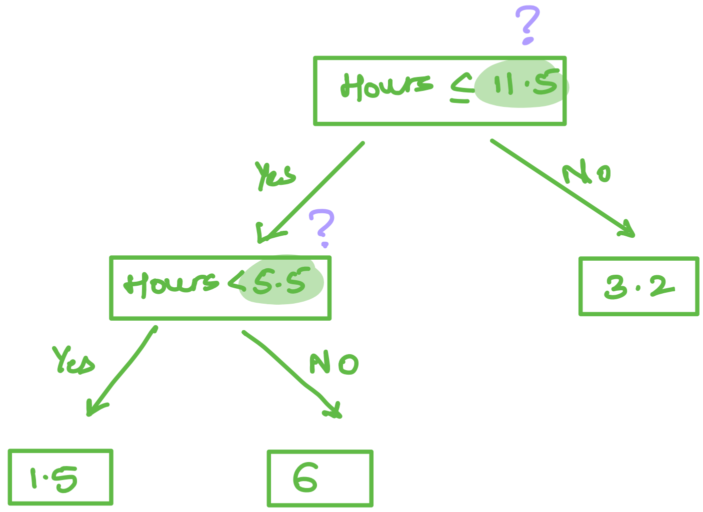
DT Regression → features discrete value → just discussed

DT Regression → features have continuous values



→ Linear Regression

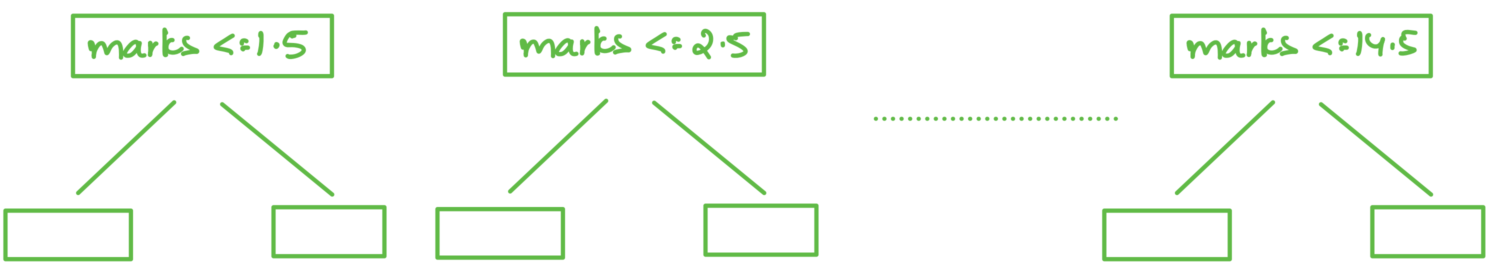




hours → X, marks → Y

1.5, 2.5, 3.5, … 14 diff values, … 14.5

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



Before Splitting

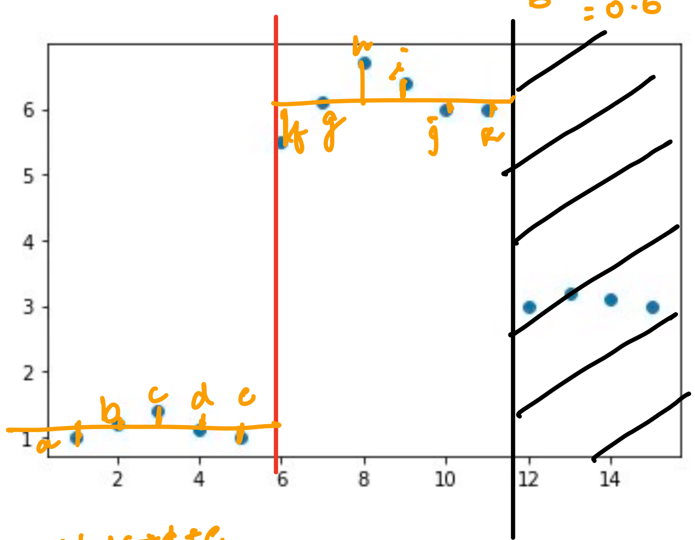
variance

X	Y	$\bar{y}$	$(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2$	$\Sigma(y-\bar{y})^2/n$
1	1		7.005		
2	1.2		5.987		
3	1.4		5.048		
4	1.1		6.486		
5	1		7.005		
6	5.5	3.647	3.435	70.299	4.686
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		

error before splitting?

5.5

$\frac{f+g+h+i+j+k}{6} = 0.6$

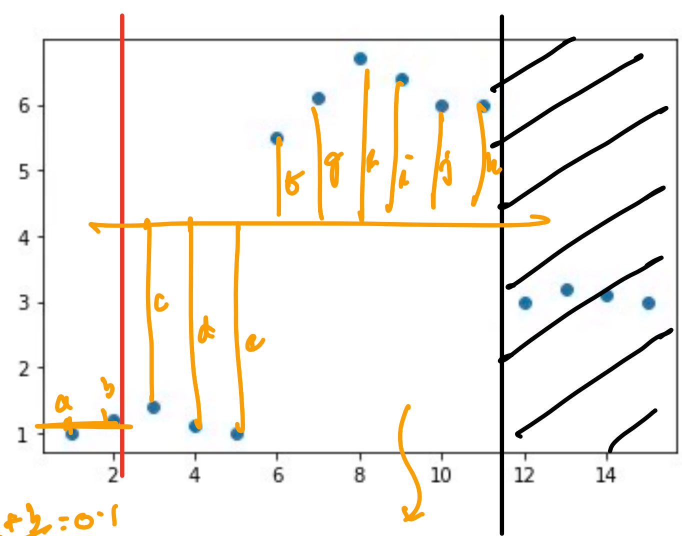


$\frac{a+b+c+d+e}{5}$

0.3

$\frac{5}{11}\times 0.3 + \frac{6}{11}\times 0.6$

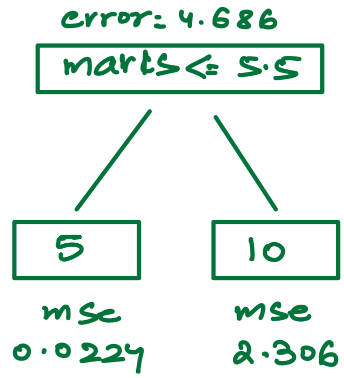
2.5



$\frac{a+b}{2} = 0.1$

4

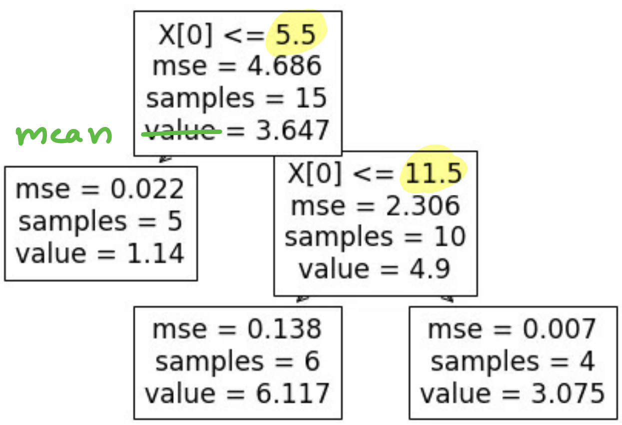
$\frac{2}{11}\times 0.1 + \frac{9}{11}\times 4$



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		

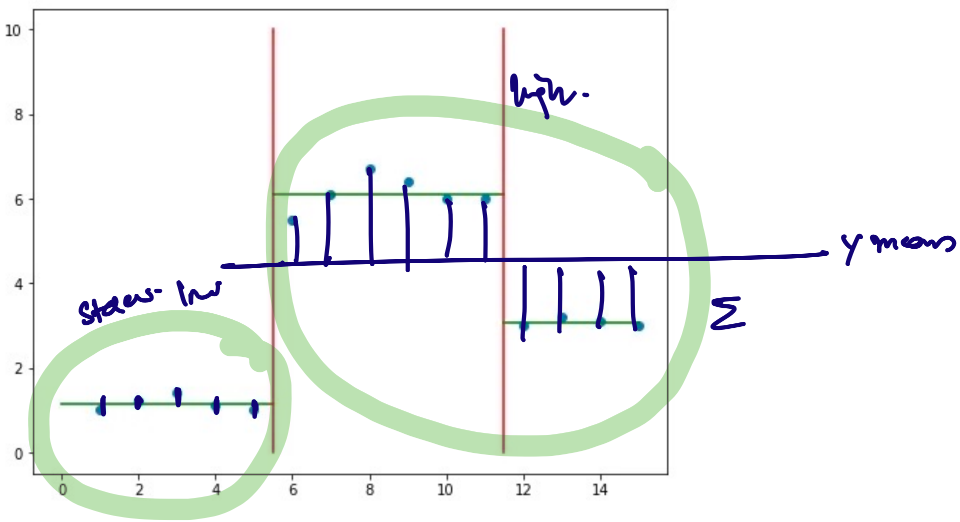
5.5

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



1.5 → —
2.5 → —
3.5 → —
4.5 → —
5.5 → 1.5448
—
—
—
—
14.5 → —

minimum



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

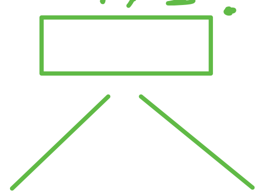
→ stop ? no. of samples

→ multiple features.

Multiple continuous features ?

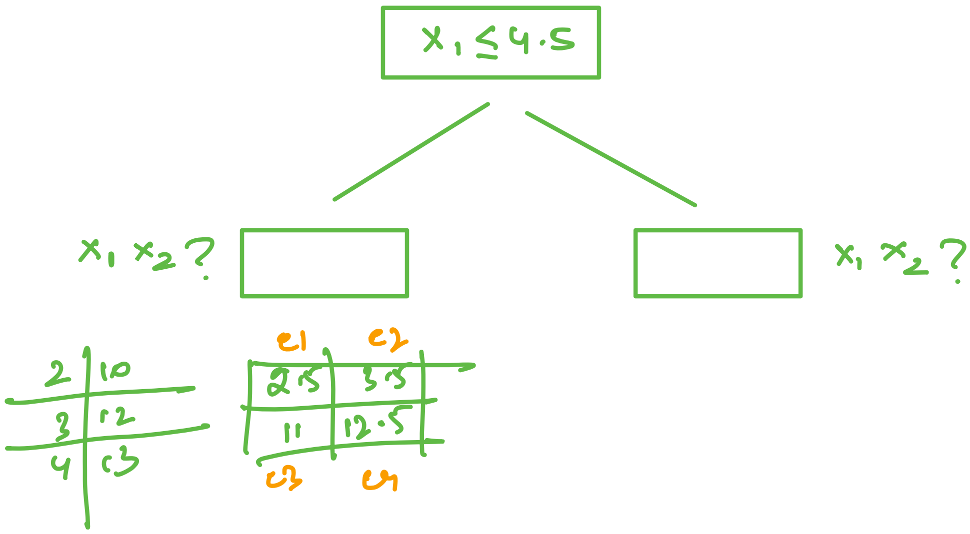
x_1	x_2	y
2	10	
3	12	
4	13	
5	14	
6	15	
7	17	

↓ continuous values ?



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

take minimum of all errors



x1 ≤ 4.5
x1 x2 ?
x1 x2 ?
2 10
3 12
4 13
e1 e2
2.5 3.5
11 12.5
e3 e4