Decision Tree Example:

Entropy (ID3 Algorithm)

$\longleftarrow x \longrightarrow$ $\longleftarrow y \longrightarrow$

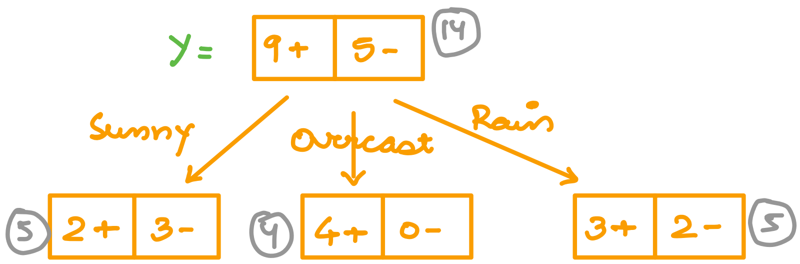
Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Find the attribute which is giving maximum information out of available attributes.

Root Node

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Outlook



$$H(Y) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Initial Entropy (Before split)

$$H(Y|x_j = \text{Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$H(Y|x_j = \text{Overcast}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$H(Y|x_j = \text{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

outlook

$$MI(Y, x_j) = \underbrace{H(Y)}_{\text{Before split}} - H(Y|x_j)$$

$$MI(Y, x_j = \text{outlook}) = H(Y) - \Big(P(x_j = \text{Sunny}) \cdot H(y \mid x_j = \text{Sunny}) +$$
$$P(x_j = \text{Overcast}) \cdot H(y \mid x_j = \text{overcast}) +$$
$$P(x_j = \text{Rain}) \cdot H(y \mid x_j = \text{Rain})\Big)$$

After split

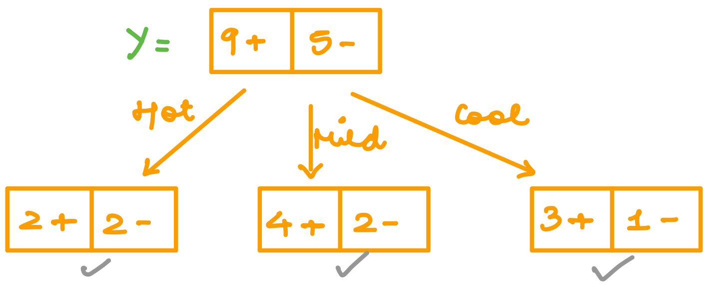
$$= 0.94 - \frac{5}{14}(0.971) - \frac{4}{14} \cdot 0 - \frac{5}{14}(0.971)$$

$$= 0.2464$$

Information Gain in outlook Attribute
(Mutual Information)

Attribute : Temp

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$H(Y) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$H(Y \mid x_j = \text{Hot}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$H(Y \mid x_j = \text{Mild}) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183$$

$$H(Y \mid x_j = \text{Cool}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

$$MI(Y \mid x_j = \text{temp}) = H(Y) - \boxed{\Big(P(x_j = \text{Hot}) \cdot H(Y \mid x_j = \text{Hot}) + P(x_j = \text{Mild}) \cdot H(Y \mid x_j = \text{Mild}) + P(x_j = \text{Cool}) \cdot H(Y \mid x_j = \text{Cool})\Big)}$$

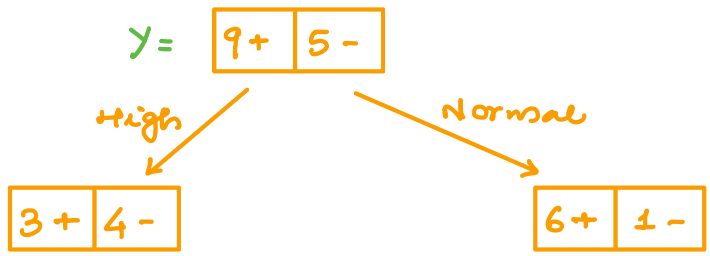
weighted avg of entropy after split

$$= 0.94 - \frac{4}{14}(1) - \frac{6}{14}(0.9183) - \frac{4}{14}(0.8113)$$

$$= 0.0289$$

Attribute: Humidity

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$H(y) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$H(y|x_j = high) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$

$$= 0.9852$$

$$H(y|x_j = normal) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

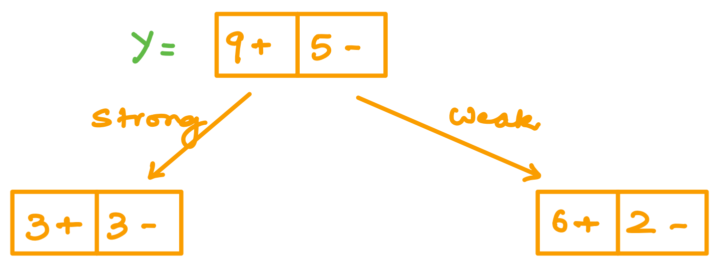
$$= 0.5916$$

$$MI(y|x_j = humidity) = 0.94 - \frac{7}{14}(0.9852) - \frac{7}{14}(0.5916)$$

$$= 0.1516$$

Attribute: Wind

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$H(y) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$H(y|x_j = strong) = 1.0$$

$$H(y|x_j = weak) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$$= 0.8113$$

$$MI(y|x_j = wind) = 0.94 - \frac{6}{14}(1.0) - \frac{8}{14}(0.8113) = 0.0478$$

$MI(y \mid x_j: \text{outlook}) = 0.2464$

$MI(y \mid x_j = \text{temp}) = 0.0289$

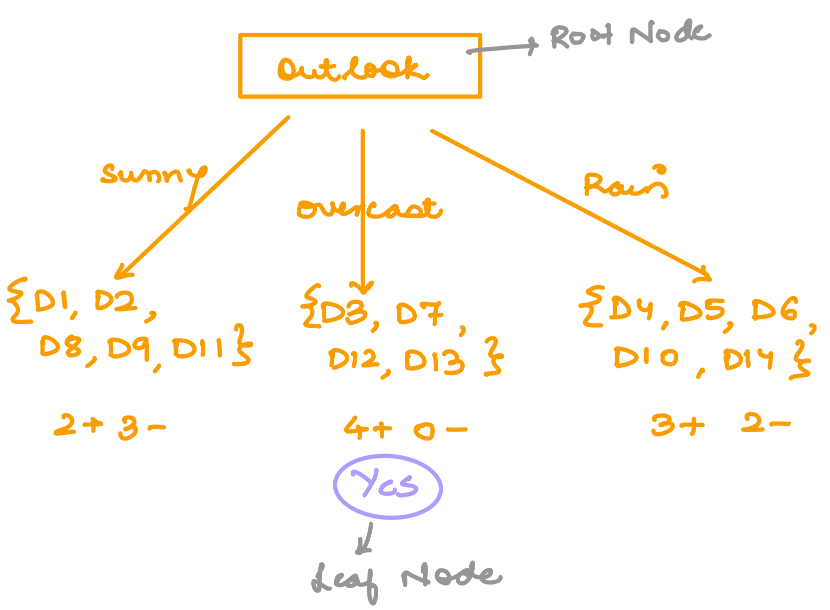
$MI(y \mid x_j = \text{humidity}) = 0.1516$

$MI(y \mid x_j = \text{wind}) = 0.0478$

which attribute has maximum MI (information gain)?

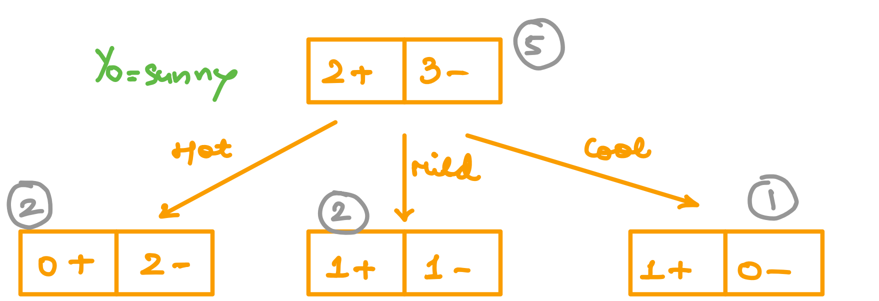
Take outlook as root node

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Attribute: Temp

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes



$H(Y_0: \text{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$ $\Big\}$ $\dfrac{2 \cdot \frac{2}{5} \cdot \frac{3}{5}}{\frac{2}{5}+\frac{3}{5}}$ $\qquad \dfrac{2 \cdot P \cdot R}{P+R}$

$H(Y_0: \text{sunny} \mid x_j: \text{hot}) = 0$ $\Big\}$ $\left(\dfrac{2 \cdot \frac{0}{2} \cdot \frac{2}{2}}{\frac{0}{2}+\frac{2}{2}}\right)\frac{2}{5}$ $+ \left(\dfrac{2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}+\frac{1}{2}}\right)\frac{2}{5}$ $+ \left(\dfrac{2 \cdot \frac{1}{1} \cdot \frac{0}{1}}{\frac{1}{1}+\frac{0}{1}}\right)\frac{1}{5}$

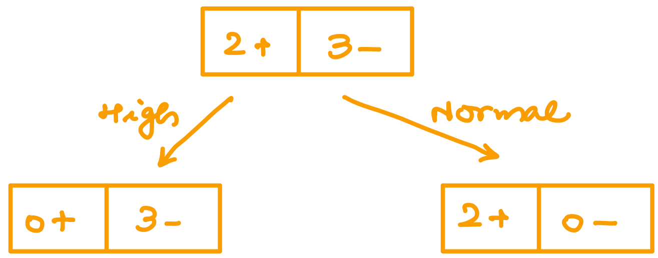
$H(Y_0: \text{sunny} \mid x_j = \text{mild}) = 1$

$H(Y_0: \text{sunny} \mid x_j: \text{cool}) = 0$

$MI(Y_{o=outlook} \mid x_j = temp) = 0.97 - \frac{2}{5}(0) - \frac{2}{5}(1) - \frac{1}{5}(0) = 0.570$

Attribute: Humidity

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes



$H(Y_{o=sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

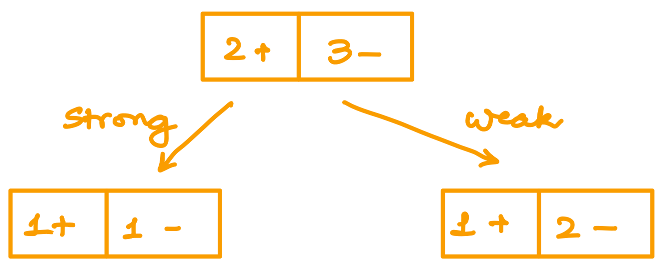
$H(Y_{o=sunny} \mid x_j = high) = 0.0$

$H(Y_{o=sunny} \mid x_j = normal) = 0.0$

$MI(Y_{o=sunny} \mid x_j = humidity) = 0.97 - \frac{3}{5}(0) - \frac{2}{5}(0) = 0.97$

Attribute: wind

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes



$H(Y) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$H(Y \mid x_j = strong) = 1.0$

$H(Y \mid x_j = weak) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183$

$MI(Y \mid x_j = wind) = 0.97 - \frac{2}{5}(1) - \frac{3}{5}(0.9183) = 0.0192$

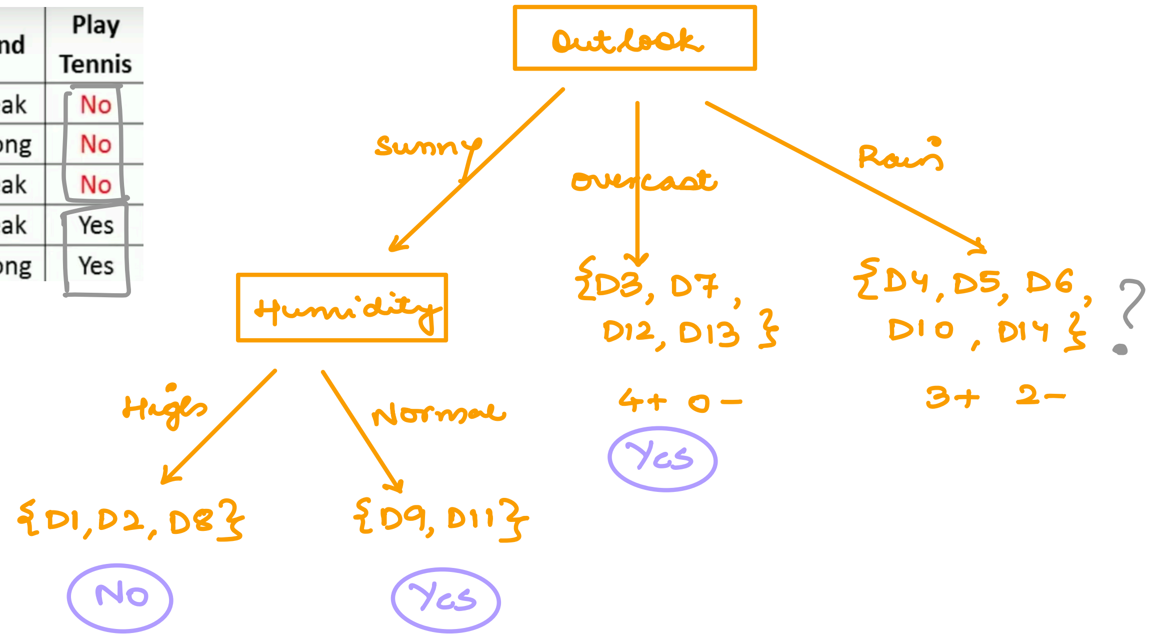
$MI(Y \mid x_j : temp) = 0.570$

$MI(Y \mid x_j : humidity) = 0.97$

$MI(Y \mid x_j : wind) = 0.0192$

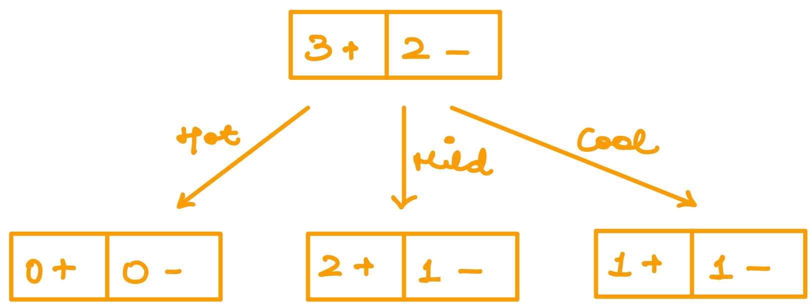
Maximum information gain is through humidity.

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes



Attribute: temp

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No



$$H(Y) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$$

$$H(Y \mid x_j = hot) = 0$$

$$H(Y \mid x_j = mild) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183$$

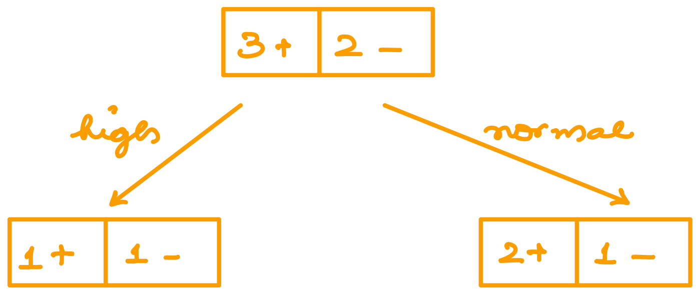
$$H(Y \mid x_j = cool) = 1.0$$

$$MI(y|x_j = temp) = 0.97 - \frac{0}{5}(0) - \frac{3}{5}(0.918) - \frac{2}{5}(1.0)$$

$$= 0.0192$$

Attribute: humidity

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No



$$H(y) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$$

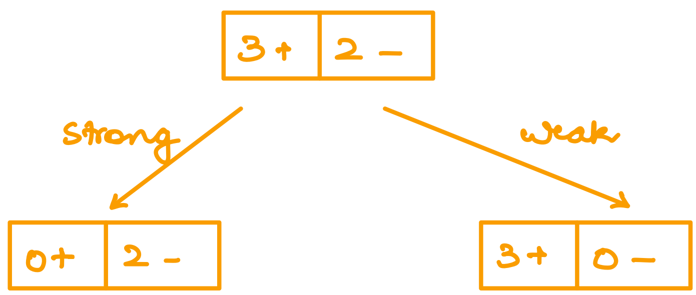
$$H(y|x_j = high) = 1.0$$

$$H(y|x_j = normal) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.9183$$

$$MI(y|x_j = humidity) = 0.97 - \frac{2}{5}(1.0) - \frac{3}{5}(0.9183) = 0.0192$$

Attribute: Wind

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No



$$H(y) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$$

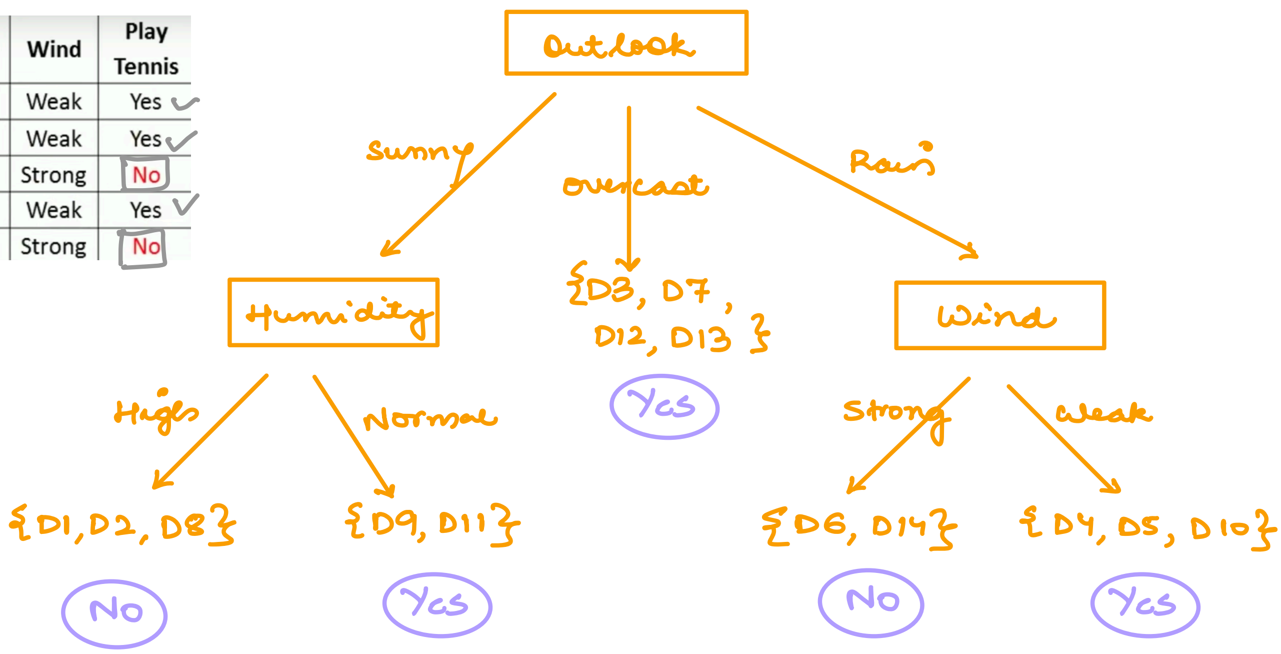
$$H(y|x_j = strong) = 0$$

$$H(y|x_j = weak) = 0$$

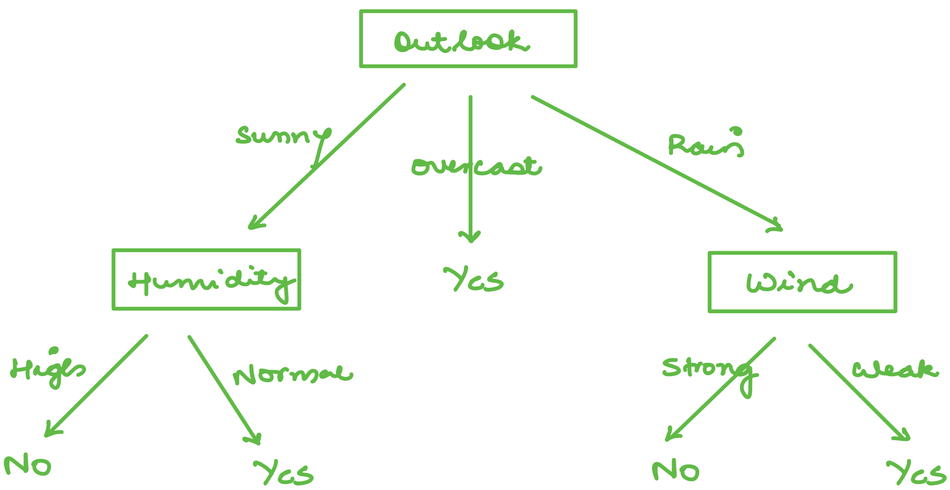
$$MI(y \mid x_j : wind) = 0.97 - \frac{2}{5}(0) - \frac{3}{5}(0) = 0.97$$

$$\left.\begin{aligned} MI(y \mid x_j = temp) &= 0.0192 \\ MI(y \mid x_j : humidity) &= 0.0192 \\ MI(y \mid x_j : wind) &= 0.97 \end{aligned}\right\} \text{maximum for wind}$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes ✓
D5	Cool	Normal	Weak	Yes ✓
D6	Cool	Normal	Strong	$\boxed{\text{No}}$
D10	Mild	Normal	Weak	Yes ✓
D14	Mild	High	Strong	$\boxed{\text{No}}$



Final Decision Tree:



Decision Tree using Gini Index

$\longleftarrow x \longrightarrow$ $\longleftarrow y \longrightarrow$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Calculate Gini Index of every attribute, attribute having minimum gini index will be selected.

$$Gini(Y) = 1 - \left[\overset{\text{Cinema}}{\left(\frac{6}{10}\right)^2} + \overset{\text{Tennis}}{\left(\frac{2}{10}\right)^2} + \overset{\text{Stay In}}{\left(\frac{1}{10}\right)^2} + \underset{\text{Shopping}}{\left(\frac{1}{10}\right)^2}\right]$$

$$= 0.58$$

Attribute : Weather

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

6 Cinema
2 Tennis
1 Stay In
1 Shopping
10

Sunny → 3: 1 Cinema, 2 Tennis, 0 Stay In, 0 Shopping

Rainy → 3: 2 Cinema, 0 Tennis, 1 Stay In, 0 Shopping

windy → 4: 3 Cinema, 0 Tennis, 0 Stay In, 1 Shopping

$$Gini(Y \mid x_j = Sunny) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$$

$$Gini(Y \mid x_j = Rainy) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$$

$$Gini(Y \mid x_j = windy) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

$$Gini(Y \mid x_j = weather) = \frac{3}{10} * 0.444 + \frac{3}{10} * 0.444 + \frac{4}{10} * 0.375$$

$$= 0.416$$

Attribute: Parents

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

6 Cinema
2 Tennis
1 Stay In
1 Shopping
10

Yes → 5: 5 Cinema, 0 Tennis, 0 Stay In, 0 Shopping

No → 5: 1 Cinema, 2 Tennis, 1 Stay In, 1 Shopping

$$Gini(y \mid x_j = Yes) = 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$$

$$Gini(y \mid x_j = No) = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0.72$$

$$Gini(y \mid x_j = parents) = \frac{5}{10} * 0 + \frac{5}{10} * 0.72 = 0.36$$

Attribute: Money

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

6 Cinema
2 Tennis
1 Stay In
1 Shopping

Poor → 3 Cinema, 0 Tennis, 0 Stay In, 0 Shopping

Rich → 3 Cinema, 2 Tennis, 1 Stay In, 1 Shopping

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

$$Gini(y \mid x_j = Rich) = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right] = 0.694$$

$$Gini(y \mid x_j = money) = \frac{3}{10} * 0 + \frac{7}{10} * 0.694 = 0.486$$

$$\left.\begin{aligned} Gini(y \mid x_j = weather) &= 0.416 \\ Gini(y \mid x_j = parents) &= 0.36 \\ Gini(y \mid x_j = money) &= 0.486 \end{aligned}\right\}$$

Gini for parents is minimum. We select parent as root node.

$\boxed{\text{Parents}}$

Yes

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

Decision is always Cinema

No

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

?

Not pure ?

Attribute : weather

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

1 Cinema
2 Tennis
1 Stay In
1 Shopping

Sunny: 0 Cinema, 2 Tennis, 0 Stay In, 0 Shopping

Rainy: 0 Cinema, 0 Tennis, 1 Stay In, 0 Shopping

windy: 1 Cinema, 0 Tennis, 0 Stay In, 1 Shopping

$$Gini(y_{parents=No} \mid x_j = Sunny) = 1 - \left[\left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini\left(Y_{parents=NO} \mid x_j = Rainy\right) = 1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$$

$$Gini\left(Y_{parents=NO} \mid x_j = windy\right) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$$

$$Gini\left(Y_{parents=NO} \mid x_j = weather\right) = \frac{2}{5} * 0 + \frac{1}{5} * 0 + \frac{2}{5} * 0.5 = 0.2$$

Attribute : Money

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

6 Cinema
2 Tennis
1 Stay In
1 Shopping

Poor

1 Cinema
0 Tennis
0 Stay In
0 Shopping

Rich

0 Cinema
2 Tennis
1 Stay In
1 Shopping

$$Gini\left(Y_{parents=NO} \mid x_j = poor\right) = 1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$$

$$Gini\left(Y_{parents=NO} \mid x_j = rich\right) = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.625$$

$$Gini\left(Y_{parents=NO} \mid x_j = money\right) = \frac{1}{5} * 0 + \frac{4}{5} * 0.625 = 0.5$$

$$\left.\begin{array}{l} Gini\left(Y_{parents=NO} \mid x_j = weather\right) = 0.2 \\ Gini\left(Y_{parents=NO} \mid x_j = money\right) = 0.5 \end{array}\right\}$$

Weather is smaller.

Parents

Yes → Cinema

No →

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Sunny →

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Rainy →

Weekend	Weather	Parents	Money	Decision
W5	Rainy	No	Rich	Stay In

windy →

Weekend	Weather	Parents	Money	Decision
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping

?

Final Decision Tree:

$\boxed{\text{Parents}}$

- Yes → Cinema
- NO → $\boxed{\text{Weather}}$
  - Sunny → Tennis
  - Rainy → Stay In
  - windy → $\boxed{\text{Money}}$
    - Rich → Shopping
    - Poor → Cinema

Decision Trees
(Discrete Attributes)

- Entropy
  (IG)
  (MI)
- Gini Index

Decision Tree with Continuous Attributes (Information Gain)

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

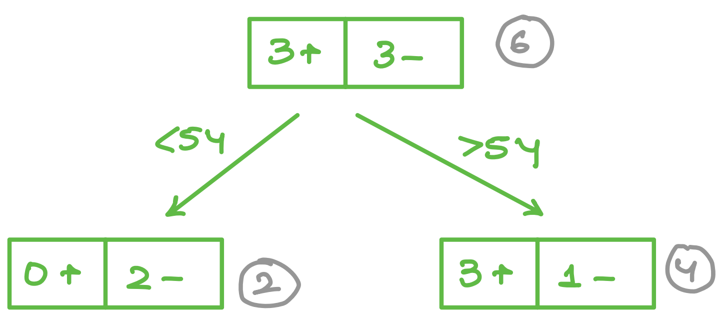
$$\frac{40+68}{2} = 54 \qquad \frac{80+90}{2} = 85$$

- 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



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

$$\left.\begin{aligned} H(y \mid x_j < 54) &= 0 \\ H(y \mid x_j > 54) &= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \end{aligned}\right] \text{After split (AS)}$$

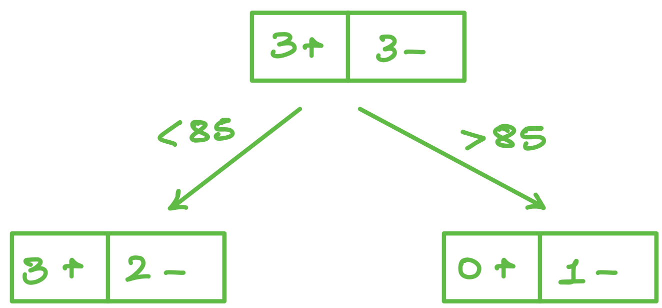
$$H(y \mid x_j = temp\ 54) = \left.\frac{2}{6} * 0 + \frac{4}{6} * 0.8113 \right] \text{weighted avg.}$$

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

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

Threshold: 85

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



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

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

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

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

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

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

$MI(y \mid x_j = temp54) = 0.4591$

$MI(y \mid x_j = temp85) = 0.1908$

IG maximized
54 is better boundary.

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

<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.

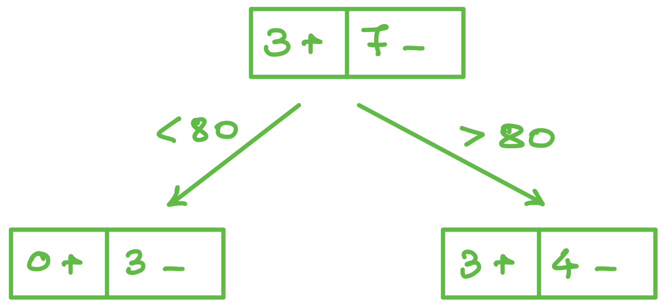
Annual Income	Label	Split Point
60	No	
70	No	
75	No	<80 >=80
85	Yes	
90	Yes	
95	Yes	<97.5 >=97.5
100	No	
120	No	
125	No	
220	No	

avg
mean

avg
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



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

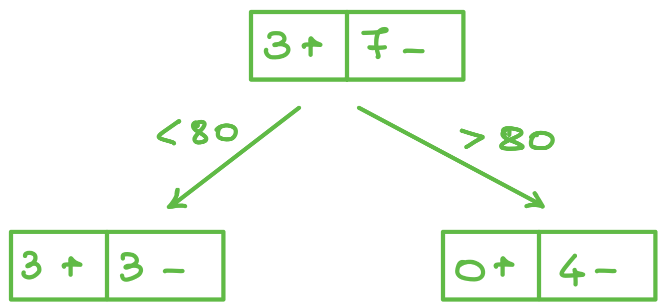
$$Gini(y|x_j>80) = 1-\left[\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right] = 0.4897$$

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

$$=0.3427$$

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



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

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

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

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

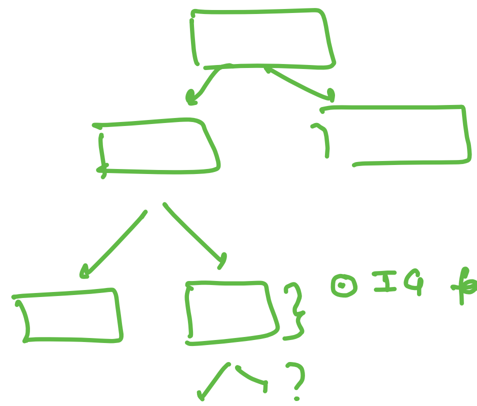
$Gini(y \mid x_j\ 97.5) = 0.3$

97.5 has smaller value.
97.5 is better splitting point.

Problem?

DT might adapt too much to the training data
↳ over fitting.

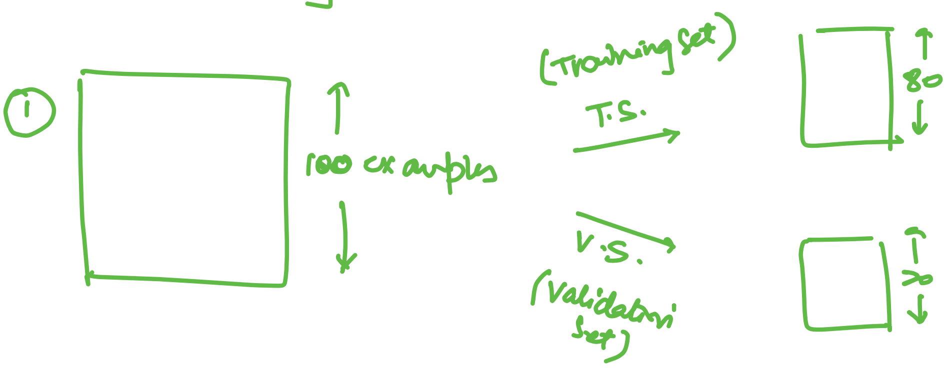
- Pre pruning
(while creating DT)



⊙ IG falls below a particular values then no more splitting

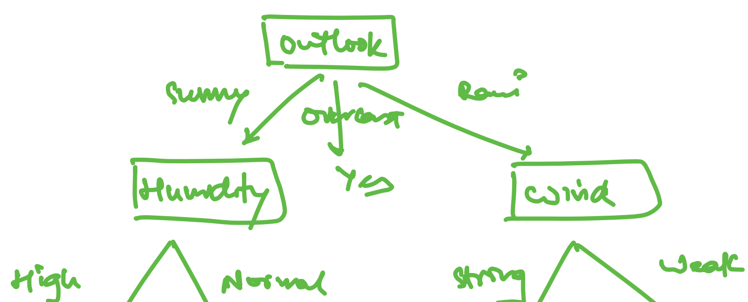
⊙ no. of examples < threshold
no more splitting

- Post pruning

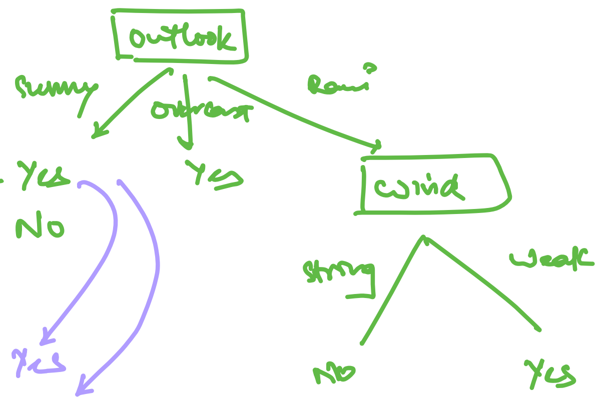


2 TS create entire tree

a



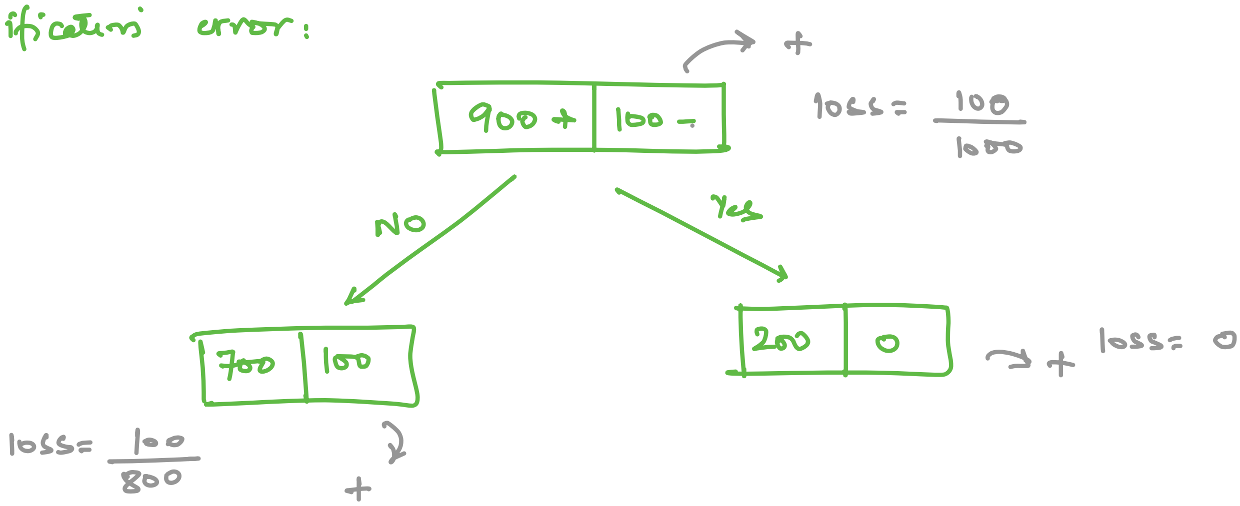
③ Pick each node in BU manner, $\boxed{\text{Humidity}}$ check if I prune this node will it improve my validation accuracy ?



ⓑ V.A. ← Yes
ⓒ V.A. ← No

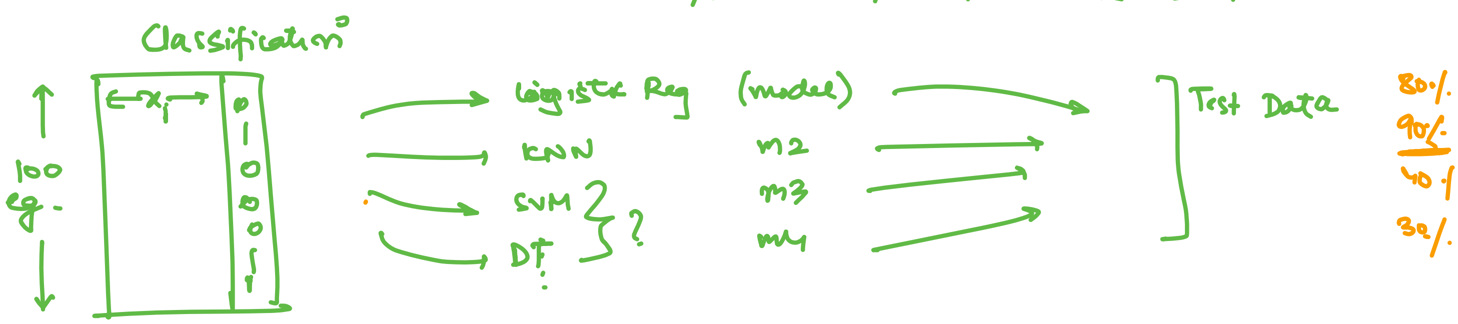
$b > a$ & $b > c$: Yes
$c > a$ & $c > b$: No
$a > b$ & $a > c$: original one (no pruning)

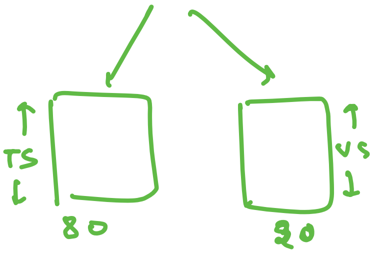
misclassification error:

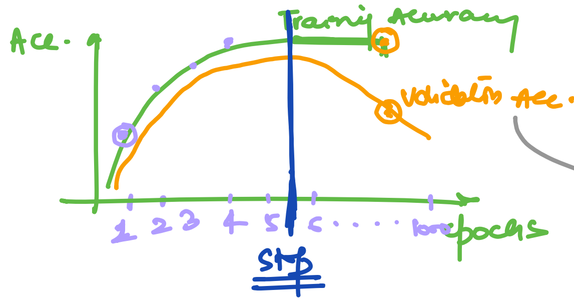


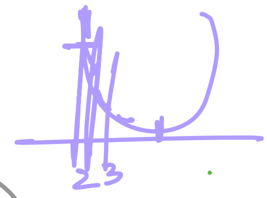
Training / validation / Test Data :

↳ you don't have access to test data.

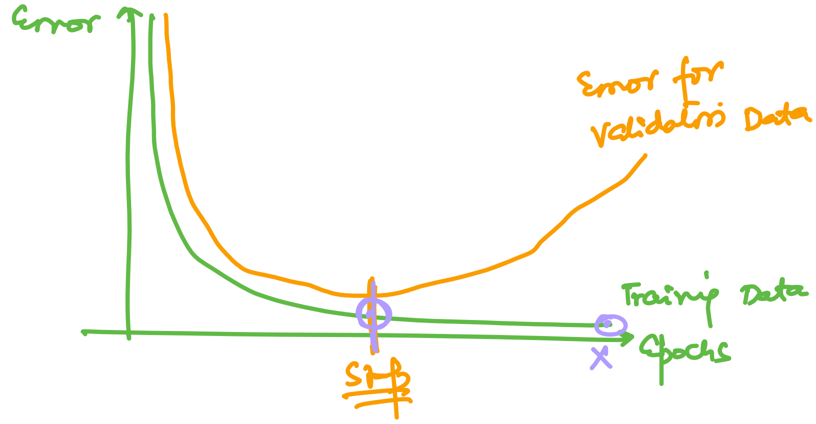








Accuracy given by my model on validation data



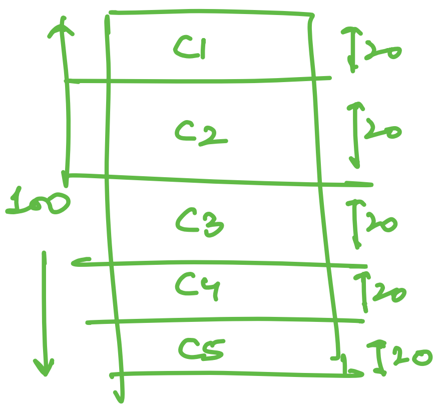
Hyper parameters

- epochs
- KNN value of K ?

GRID SEARCH : k = 1, 5, 10, 20, 100

k-fold Cross validation :

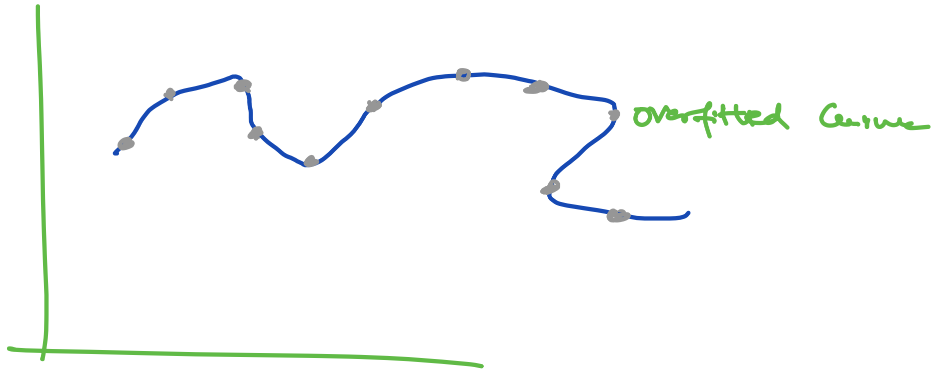
k=5



Step1:	TD: C1 C2 C3 C4	VD: C5
2:	TD: C1 C2 C3 C5	VD: C4
3:	TD: C1 C2 C4 C5	VD: C3
4:	TD: C1 C3 C4 C5	VD: C2
5:	TD: C2 C3 C4 C5	VD: C1

Overfitting:

Model performs very well on training data but doesnot perform good on test data.



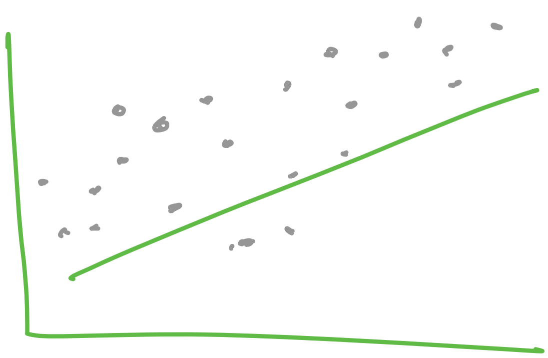
Solution:

- k fold cross validation
- Sufficient data
- Ensembling technique.
    - KNN —
    - SVM —
    - Log.R —
    - DT —

  Majority vote

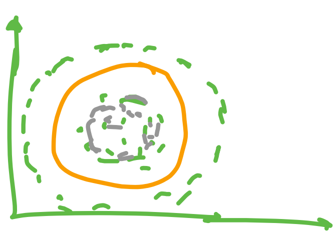
Underfitting:

Model is not going to learn patterns from training data.



underfit model will give poor performance on train as well as test data.

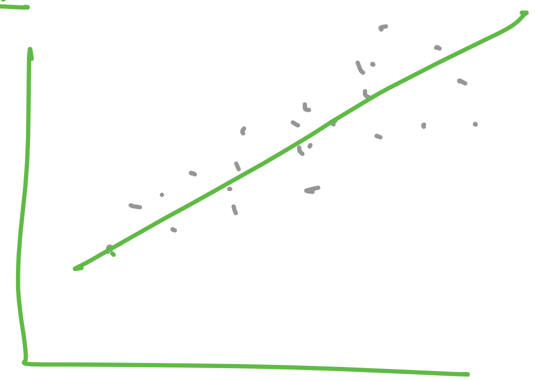
1



Solutions:

- No. of features increase
- Models Complexity
- Reduce noise
- Increase the duration of training.

Good fit ?



Error of model reduces on training data as well as on val data.

BIAS AND VARIANCE:

Bias: wrong assumptions about data — like assuming data is linear in reality is follows a complex fx^n.

It is the inability of the model bcz of that there is diff in predicted value & actual value.

Low Bias: lower assumptions make a model which closely matches the training dataset. (simple model)

High Bias: more assumptions model will not match training dataset closely. (complex model)

Variance:

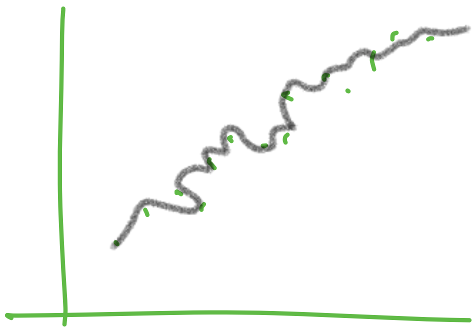
measure of spread in data from its mean position.

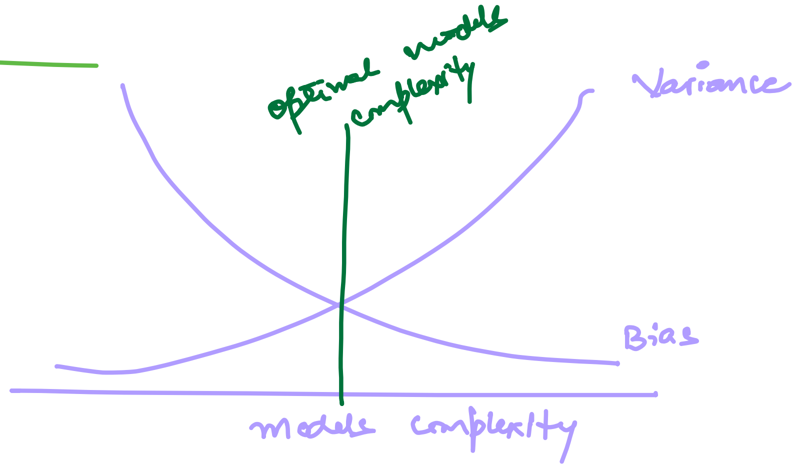
Sensitive to a subset which follows same distribution as your training dataset.

Low variance. model is very less sensitive to changes.

High variance. model is very sensitive to changes and it

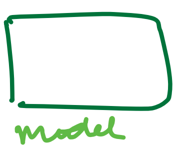
can result in significant changes if trained on a different subset.
↓
overfitting.





Bias Variance Tradeoff.

Metrics



Test how good your model is?

Confusion Matrix:

0, 1

Predicted label

True label	0	1
0	True Negative (TN)	false Positive (FP)
1	false Negative (FN)	True Positive (TP)

Cancer patients

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$ } out of all predicted true how many were correct

$$\text{Recall} = \frac{TP}{TP + FN}$$ } out of all actual trues → how many were detected

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

↓

HM of Precision & Recall

$$\text{Sensitivity (True positive Rate)} = \frac{TP}{TP+FN} * 100$$

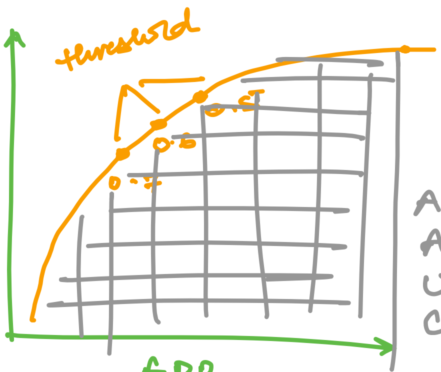
$$\text{Specificity (True Negative Rate)} = \frac{TN}{TN+FP} * 100$$

$$\text{Efficiency} = \frac{\text{Sensitivity} + \text{Specificity} + \text{Accuracy}}{3}$$

$$\text{false positive Rate} = \frac{FP}{FP+TN} * 100$$

ROC Curve (Reciever Operating Characteristics)

TPR



AUC
Area
under
Curve

FPR

Different values of threshold

↳ 0.5 $<0.5 \rightarrow 0$

$\geq 0.5 \rightarrow 1$

0.3 $<0.3 \rightarrow 0$

$\geq 0.3 \rightarrow 1$

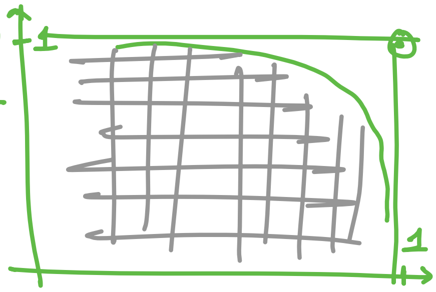
⋮

0.8 $<0.8 \rightarrow 0$

$\geq 0.8 \rightarrow 1$

PR Curve

Precision $\frac{TP}{TP+FP}$

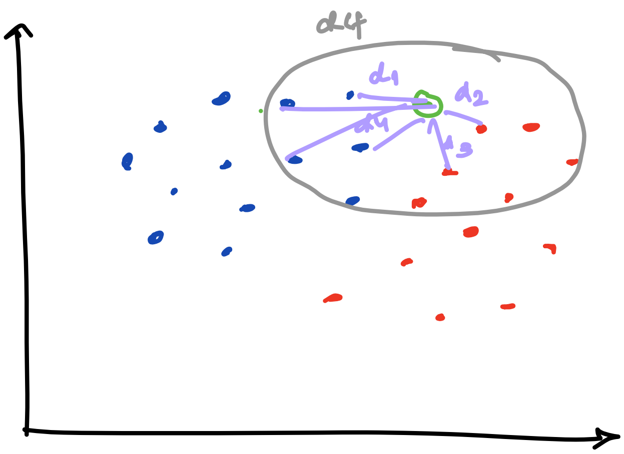


Recall $\frac{TP}{TP+FN}$

ideally: high Precision $FP=0$

high Recall $FN=0$

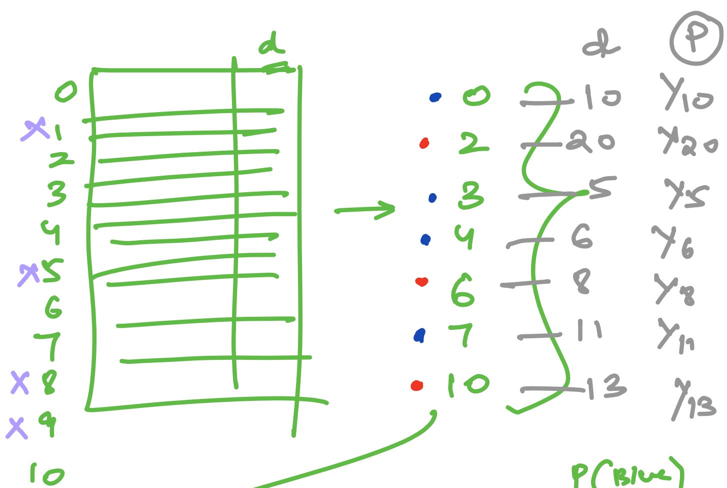
AUC measures the probability that the model will assign a randomly chosen positive instance a higher predicted probability compared to a randomly chosen -ve instance.



$\sqrt{(6-5)^2+(4-5)^2}$

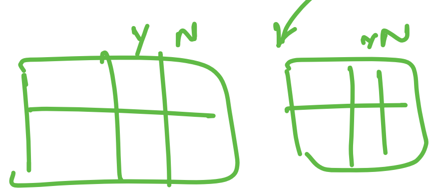
dcf

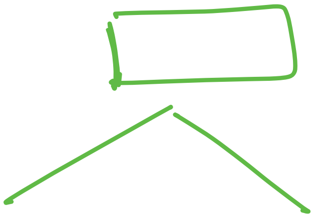
$d_1 \ldots \; > dcf$ ignore.



$\frac{1}{10}+\frac{1}{5}+\frac{1}{6}+\frac{1}{11}$: $\frac{\;}{P(\text{Blue})}$

$\frac{1}{20}+\frac{1}{8}+\frac{1}{13}$: $\frac{\;}{P(\text{Red})}$





F_1 score $= \frac{2PR}{P+R} \times = \frac{2 p_1 p_2}{p_1 + p_2}$

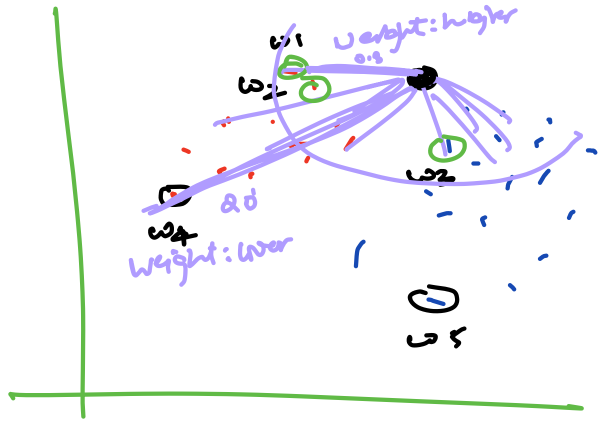
- → Linear Regression
- → Logistic Regression
- → KNN
- → Naive Bayes
- → Decision Trees
- → Todays class

Py Q's
↳ numericals

KNN:

$w_1 \to d_1$
$w_2 \to d_2$
$w_3 \to d_3$

pick Arg 3



maximize

1 − gini

minimize