

$$-\sum_{k=1}^n p_k \log p_k$$

+ve	-ve
20	10

$$H(Y) = - \left(\frac{20}{30} \log \frac{20}{30} + \frac{10}{30} \log \frac{10}{30} \right)$$

Entropy before splitting

$$x_j = 1$$

+ve	-ve
12	8

$$x_j = 0$$

+ve	-ve
8	2

$$H(Y | x_j = 0) = - \left(\frac{8}{10} \log \frac{8}{10} + \frac{2}{10} \log \frac{2}{10} \right)$$

$$H(Y | x_j = 1) = - \left(\frac{12}{20} \log \frac{12}{20} + \frac{8}{20} \log \frac{8}{20} \right)$$

Combine:

$$H(Y | x_j) = \sum_g P(x_j = g) H(Y | x_j = g)$$

$$= P(x_j = 1) \cdot H(Y | x_j = 1) + P(x_j = 0) \cdot H(Y | x_j = 0)$$

$$= - \left(\frac{20}{30} \left(\frac{12}{20} \log \frac{12}{20} + \frac{8}{20} \log \frac{8}{20} \right) + \frac{10}{30} \left(\frac{8}{10} \log \frac{8}{10} + \frac{2}{10} \log \frac{2}{10} \right) \right)$$

↓
Entropy after the split

① **Mutual Information** = $H(Y) - H(Y | x_j)$

Select the feature which gives the maximum MI.

② **Gini Index**

$$Gini(x) = 1 - \sum_{k=1}^n p_k^2$$

$$\rightarrow \sum_{k=1}^n p_k \log p_k$$

Decision Tree Example:

Entropy (ID3 Algorithm)

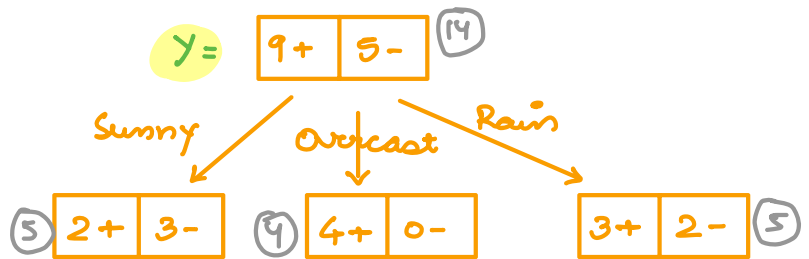
Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes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



$$H(y) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94 \text{ Initial Entropy (Before split)}$$

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

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

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

outlook

$$MI(y, x_j) = H(y) - H(y|x_j)$$

Before split

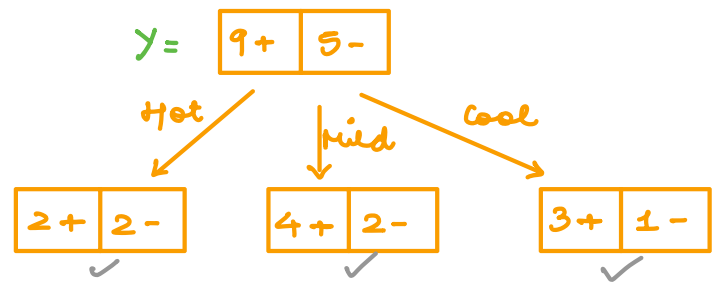
$$\begin{aligned}
 MI(Y, x_j = \text{outlook}) &= H(Y) - P(x_j = \text{Sunny}) \cdot H(Y | x_j = \text{Sunny}) + \text{After split} \\
 &\quad + P(x_j = \text{Overcast}) \cdot H(Y | x_j = \text{Overcast}) + \\
 &\quad + P(x_j = \text{Rain}) \cdot H(Y | x_j = \text{Rain}) \\
 &= 0.94 - \frac{5}{14} (0.971) - \frac{4}{14} \cdot 0 - \frac{5}{14} (0.971)
 \end{aligned}$$

$$= 0.2464$$

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



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

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

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

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

$$\begin{aligned}
 MI(Y | x_j = \text{Temp}) &= H(Y) - \left(P(x_j = \text{Hot}) \cdot H(Y | x_j = \text{Hot}) + \right. \\
 &\quad \left. P(x_j = \text{Mild}) \cdot H(Y | x_j = \text{Mild}) + \right. \\
 &\quad \left. P(x_j = \text{Cool}) \cdot H(Y | x_j = \text{Cool}) \right)
 \end{aligned}$$

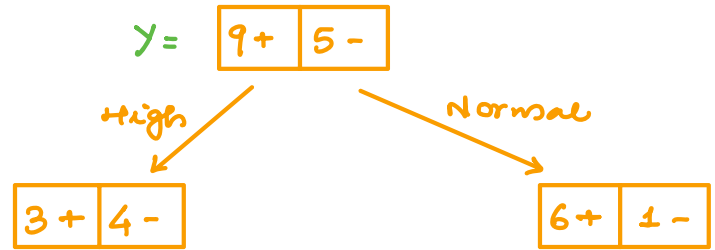
weighted avg of entropy after split

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

$$= 0.0289$$

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



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

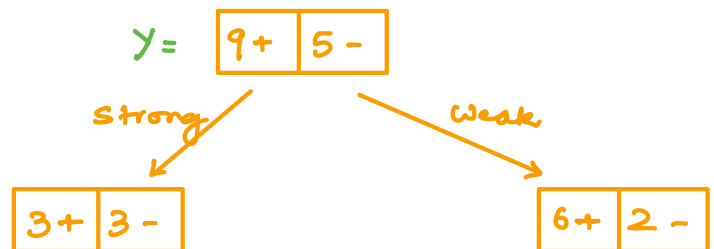
$$H(y|x_j = \text{high}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.9852$$

$$H(y|x_j = \text{normal}) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.5916$$

$$MI(y|x_j = \text{humidity}) = 0.94 - \frac{7}{14} (0.9852) - \frac{7}{14} (0.5916) = 0.1516$$

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



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

$$H(y|x_j = \text{strong}) = 1.0$$

$$H(y|x_j = \text{weak}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.8113$$

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

$$MI(Y|X_j = \text{outlook}) = 0.2464$$

$$MI(Y|X_j = \text{temp}) = 0.0289$$

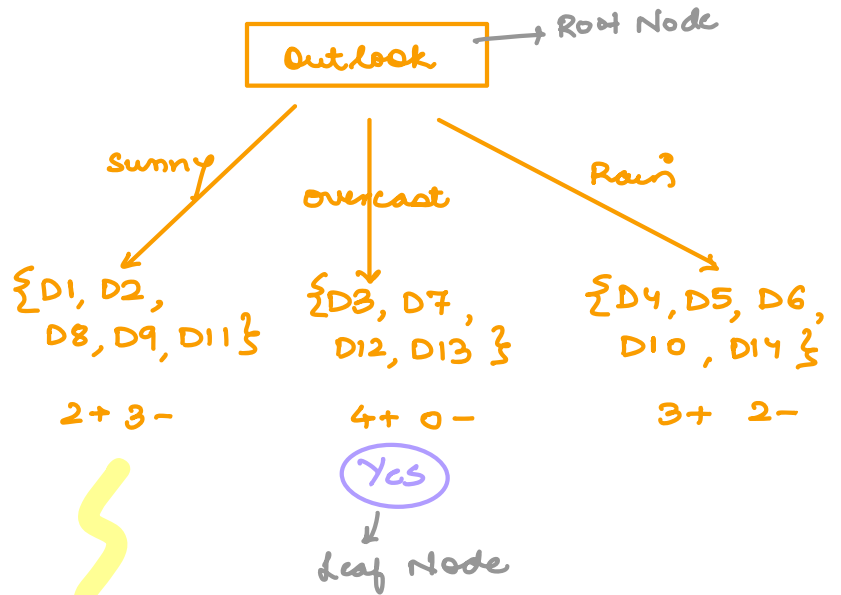
$$MI(Y|X_j = \text{humidity}) = 0.1516$$

$$MI(Y|X_j = \text{wind}) = 0.0478$$

which attribute has maximum MI (information gain)?

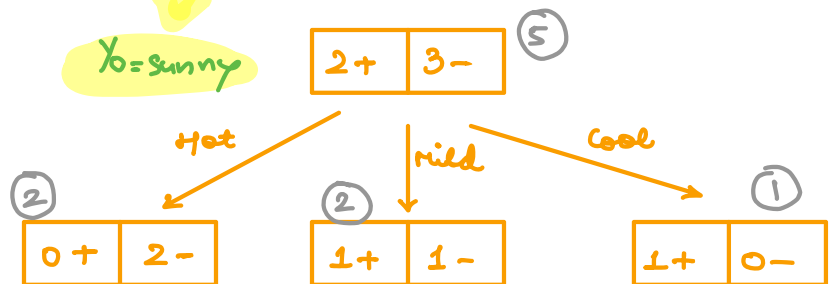
Take outlook as root node

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



$$H(Y_0: \text{sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$H(Y_0: \text{sunny} | X_j = \text{hot}) = 0$$

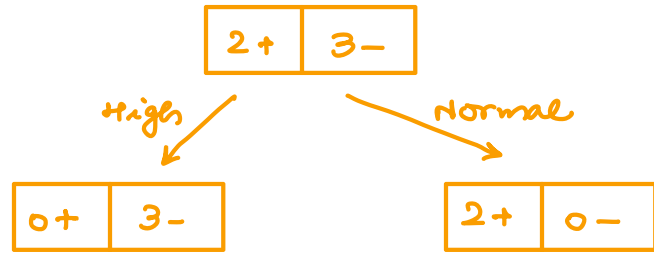
$$H(Y_0: \text{sunny} | X_j = \text{mild}) = 1$$

$$H(Y_0: \text{sunny} | X_j = \text{cool}) = 0$$

$$MI(Y_0 = \text{outlook} | x_j = \text{temp}) = 0.97 - \frac{2}{5}(0) - \frac{2}{5}(1) - \frac{1}{5}(0) = 0.570$$

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



$$H(Y_0 = \text{sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

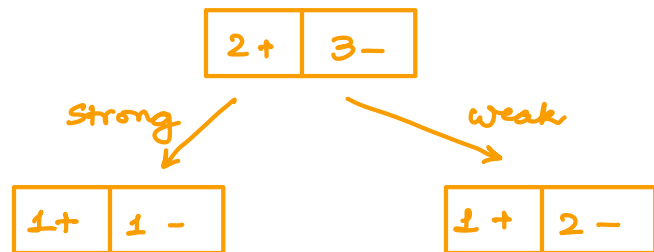
$$H(Y_0 = \text{sunny} | x_j = \text{high}) = 0.0$$

$$H(Y_0 = \text{sunny} | x_j = \text{normal}) = 0.0$$

$$MI(Y_0 = \text{sunny} | x_j = \text{humidity}) = 0.97 - \frac{3}{5}(0) - \frac{2}{5}(0) = 0.97$$

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



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

$$H(Y | x_j = \text{strong}) = 1.0$$

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

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

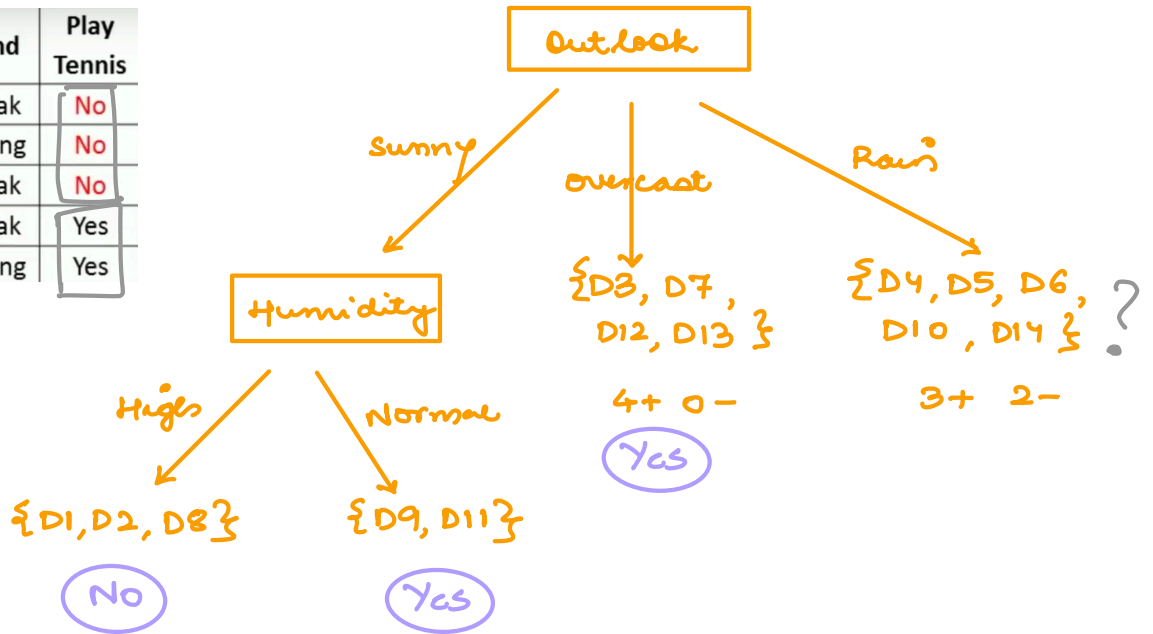
$$H(Y|X_j = \text{temp}) = 0.570$$

$$H(Y|X_j = \text{humidity}) = 0.97$$

$$H(Y|X_j = \text{wind}) = 0.0192$$

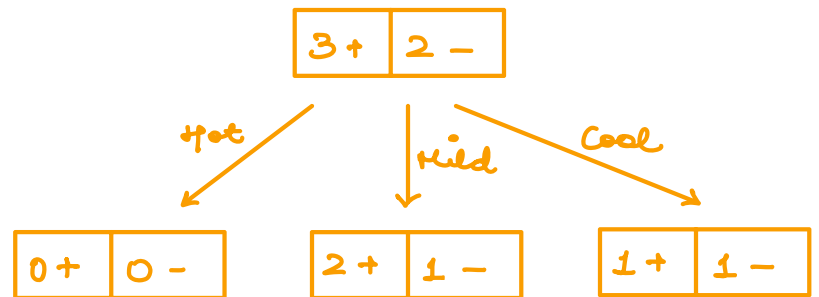
Maximum information gain is through humidity.

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



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

$$H(Y|X_j = \text{hot}) = 0$$

$$H(Y|X_j = \text{mild}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

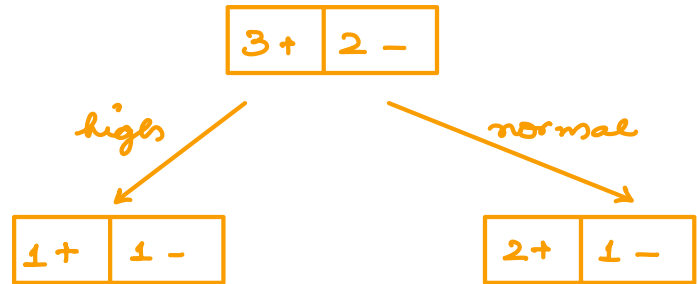
$$H(Y|X_j = \text{cool}) = 1.0$$

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

$$= 0.0192$$

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



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

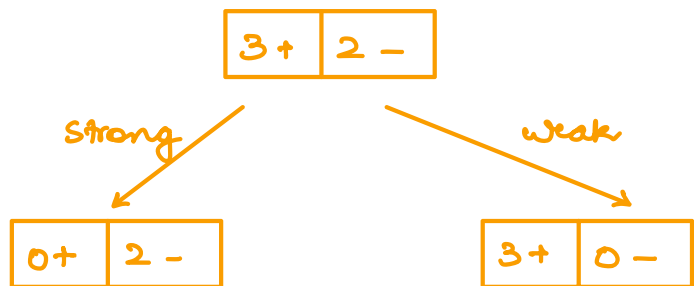
$$H(y|x_j = high) = 1.0$$

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

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

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



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

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

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

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

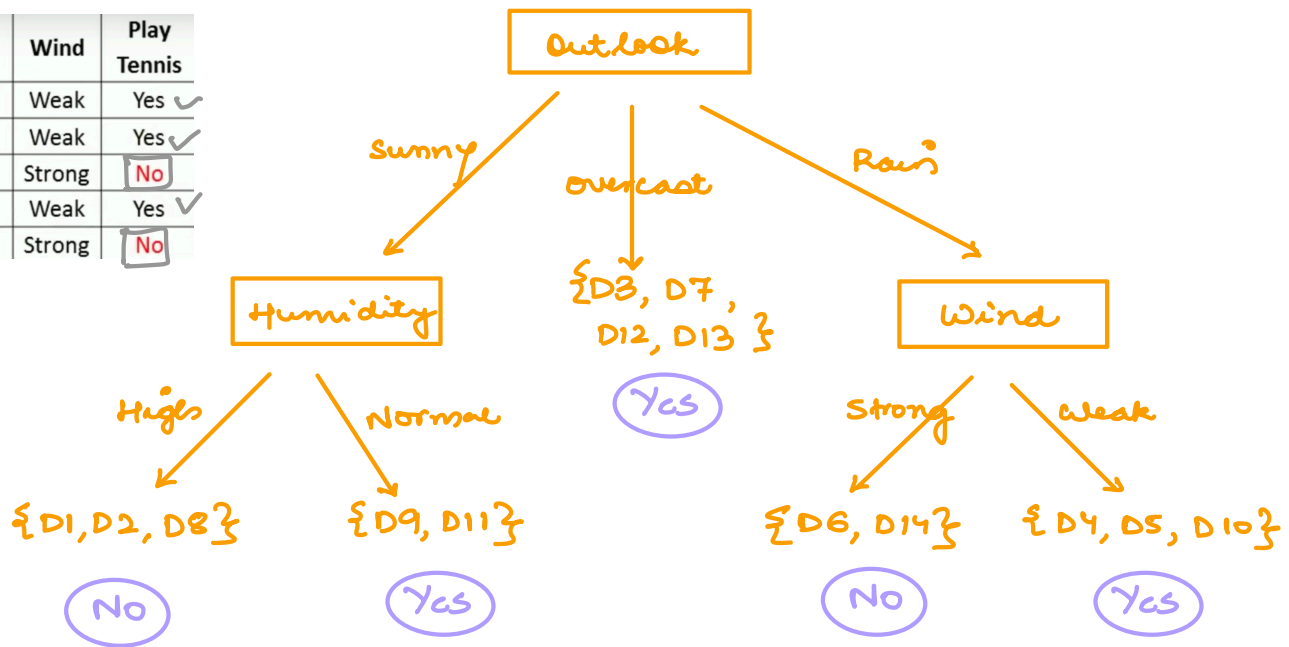
$$MI(y | x_j = \text{temp}) = 0.0192$$

$$MI(y | x_j = \text{humidity}) = 0.0192$$

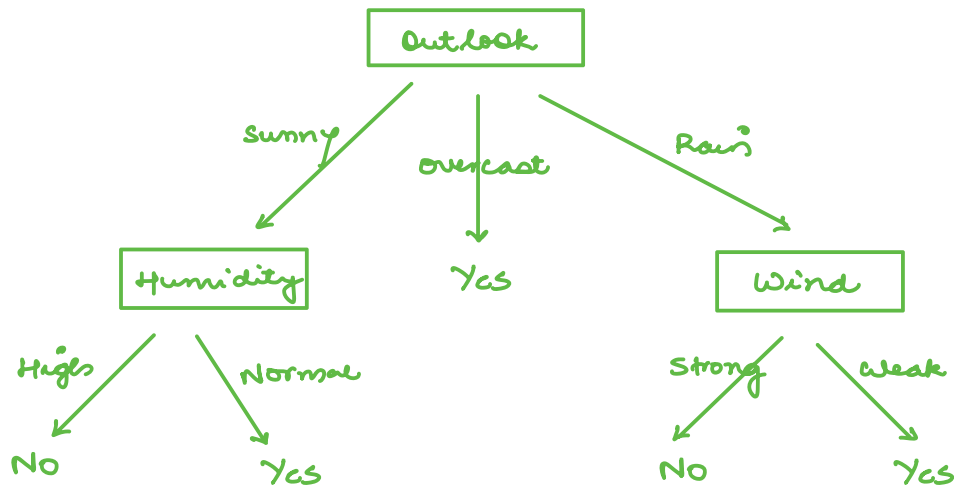
$$MI(y | x_j = \text{wind}) = 0.97$$

} maximum for 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



Final Decision Tree:



Decision Tree using Gini Index

← x → ← y →

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.

$$Gini(y) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2 \right]$$

Cinema
Tennis
Stay In

Shopping

$$= 0.58$$

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



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

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

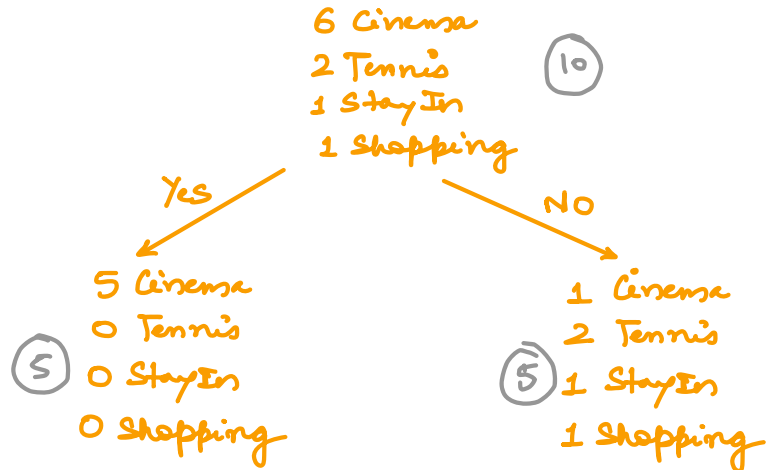
$$Gini(y|x_j = \text{windy}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right] = 0.375$$

$$Gini(y|x_j = \text{weather}) = \frac{3}{10} * 0.444 + \frac{3}{10} * 0.444 + \frac{4}{10} * 0.375$$

$$= 0.416$$

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



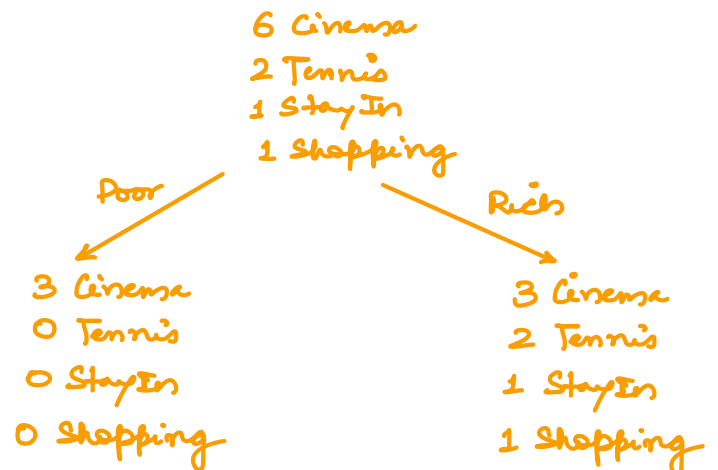
$$Gini(y|x_j = \text{Yes}) = 1 - \left[\left(\frac{5}{5} \right)^2 \right] = 0$$

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

$$Gini(y|x_j = \text{Parents}) = \frac{5}{10} * 0 + \frac{5}{10} * 0.72 = 0.36$$

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



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

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

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

$$Gini(y | x_j = \text{weather}) = 0.416$$

$$Gini(y | x_j = \text{parents}) = 0.36$$

$$Gini(y | x_j = \text{money}) = 0.486$$

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

Parents

Yes No

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

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



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

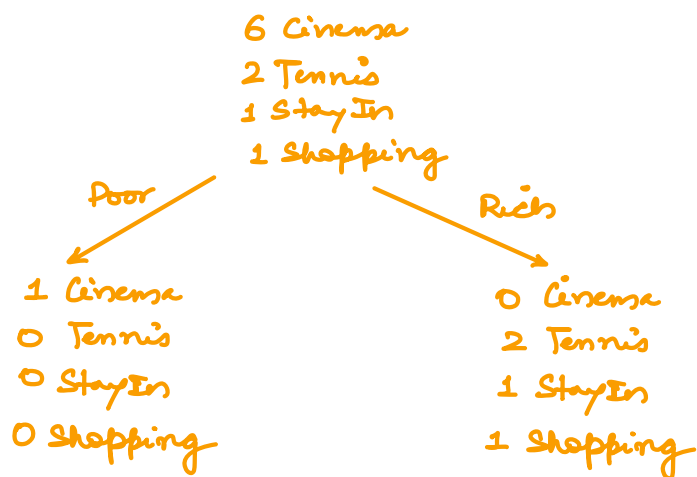
$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{Rainy}) = 1 - \left[\left(\frac{1}{1} \right)^2 \right] = 0$$

$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{Windy}) = 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] = 0.5$$

$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{Weather}) = \frac{2}{5} * 0 + \frac{1}{5} * 0 + \frac{2}{5} * 0.5 = 0.2$$

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



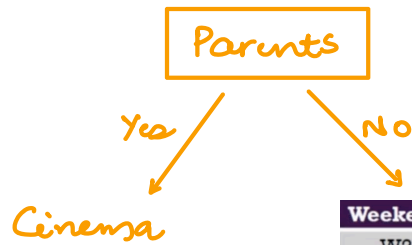
$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{poor}) = 1 - \left[\left(\frac{1}{1} \right)^2 \right] = 0$$

$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{rich}) = 1 - \left[\left(\frac{2}{4} \right)^2 + \left(\frac{1}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0.625$$

$$Gini(y_{\text{parents}=\text{No}} | x_j = \text{money}) = \frac{1}{5} * 0 + \frac{4}{5} * 0.625 = 0.5$$

$Gini(y_{\text{parents}=\text{No}} | x_j = \text{Weather}) = 0.2$
 $Gini(y_{\text{parents}=\text{No}} | x_j = \text{money}) = 0.5$

Weather is smaller.



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

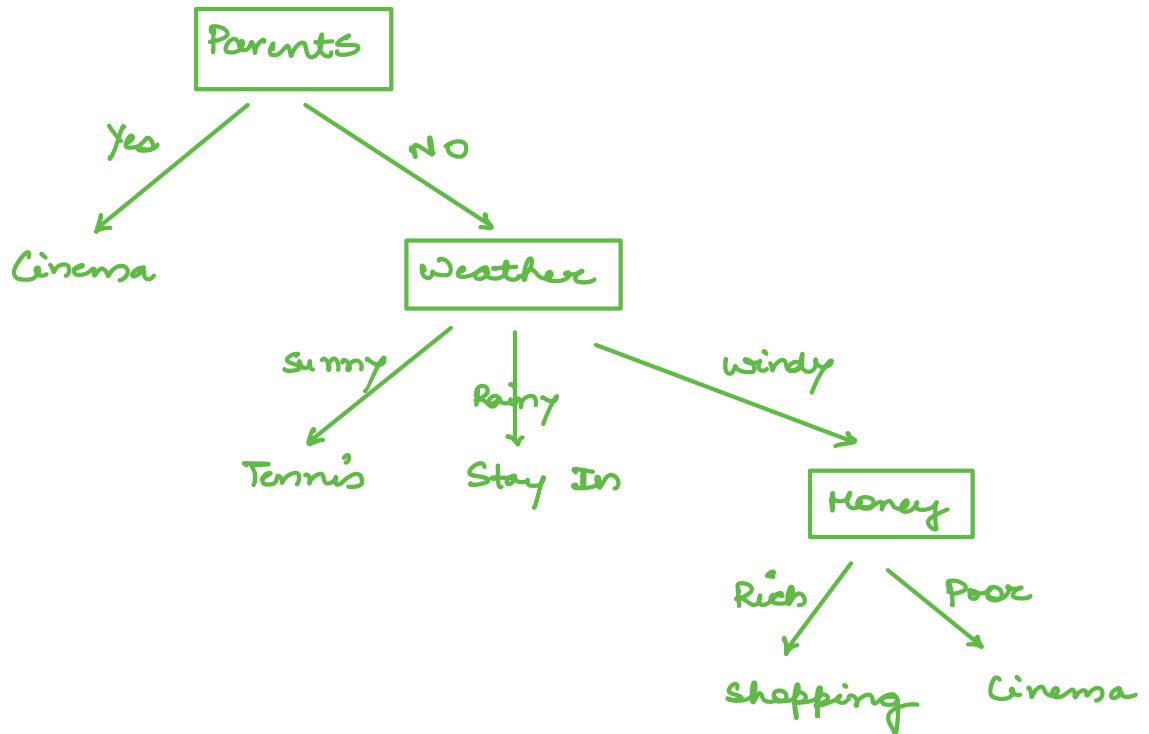
Weekend	Weather	Parents	Money	Decision
W5	Rainy	No	Rich	Stay In

windy

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

?

Final Decision Tree:



Decision Trees
(Discrete Attributes)

Entropy
(IG)
(MI)

Gini Index

Decision Tree with Continuous Attributes (Information Gain)

Discrete Y Value

x: Temperature 40 48 60 72 80 90 → Continuous Values

y: PlayTennis No No Yes Yes Yes No

- 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

$$\frac{40+60}{2} = 50$$

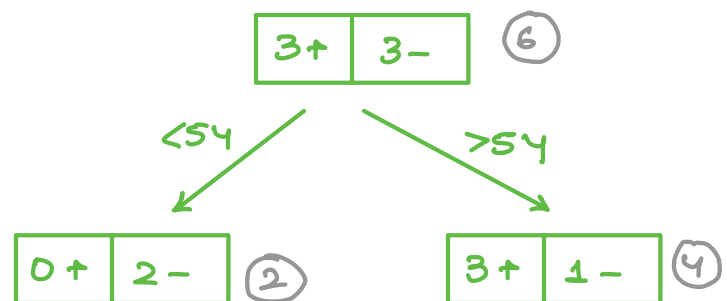
$$\frac{80+90}{2} = 85$$

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

Calculate MI through 50 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



$$H(Y) = -\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} = 1 \quad \left. \vphantom{H(Y)} \right\} \text{Before split (BS)}$$

$$H(Y | x_j < 54) = 0$$

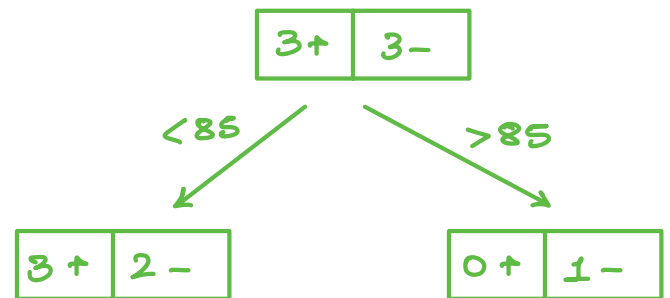
$$H(Y | x_j > 54) = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8113 \quad \left. \vphantom{H(Y | x_j > 54)} \right\} \text{After split (AS)}$$

$$H(Y | x_j = \text{temp } 54) = \frac{2}{6} * 0 + \frac{4}{6} * 0.8113 \quad \left. \vphantom{H(Y | x_j = \text{temp } 54)} \right\} \text{weighted avg.}$$

$$\begin{aligned} MI(Y | x_j = \text{temp } 54) &= \overbrace{H(Y)}^{\text{BS.}} - \overbrace{H(Y | x_j = \text{temp } 54)}^{\text{AS.}} \\ &= 1 - \left(\frac{2}{6} * 0 + \frac{4}{6} * 0.8113 \right) = 0.4591 \end{aligned}$$

Threshold: 85

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



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

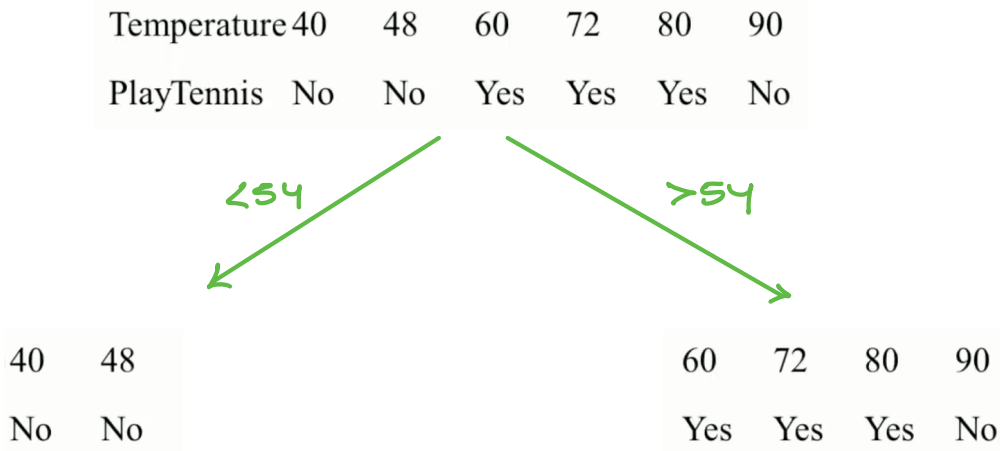
$$H(Y | x_j < 85) = -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} = 0.971$$

$$H(Y | x_j > 85) = 0$$

$$H(Y | x_j = \text{temp } 85) = \frac{5}{6} * 0.971 + \frac{1}{6} * 0$$

$$\begin{aligned} MI(Y | x_j = \text{temp } 85) &= H(Y) - H(Y | x_j = \text{temp } 85) \\ &= 1 - \left(\frac{5}{6} * 0.971 + \frac{1}{6} * 0 \right) = 0.1908 \end{aligned}$$

$$\begin{aligned}
 MI(y|x_j = \text{temp}54) &= 0.4591 \\
 MI(y|x_j = \text{temp}85) &= 0.1908
 \end{aligned}
 \left. \vphantom{\begin{aligned} MI(y|x_j = \text{temp}54) \\ MI(y|x_j = \text{temp}85) \end{aligned}} \right\} \begin{array}{l} \text{IG maximized} \\ 54 \text{ is better boundary.} \end{array}$$



Overfitting

Continuous values with Gini Index

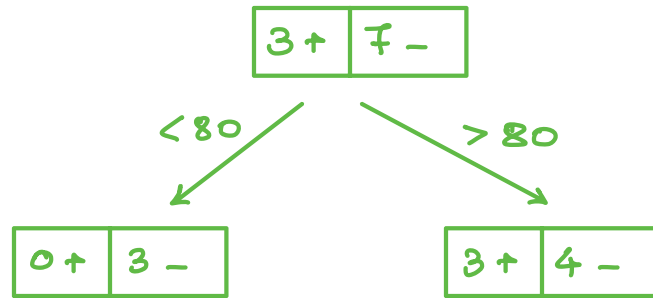
Annual Income	Label
60	No
70	No
75	No
85	Yes
90	Yes
95	Yes
100	No
120	No
125	No
220	No

- 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 → mean
85	Yes	≥ 80
90	Yes	
95	Yes	< 97.5 → mean
100	No	≥ 97.5
120	No	
125	No	
220	No	

- 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



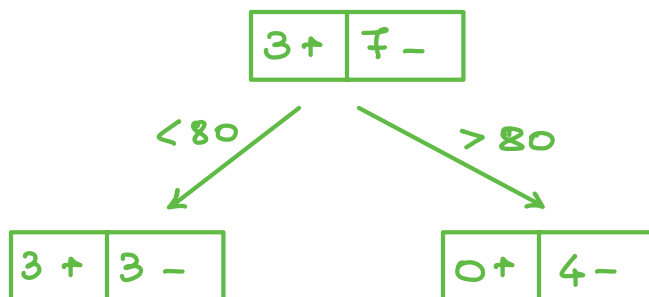
$$Gini(y | x_j < 80) = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2 \right] = 0$$

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

$$Gini(y | x_j 80) = \frac{3}{10} * 0 + \frac{7}{10} * 0.4897 = 0.3427$$

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



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

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

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

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

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

} 97.5 has smaller value.
97.5 is better splitting point.