Decision Tree

Play Tennis Dataset

- To find out the root node we will follow 2 approaches.

  ① Information Gain → Max Value

  ② Gini Impurity. → Min Value

Information Gain

$$\boxed{\text{Entropy} = \sum -P \log_2(P)}$$

Label → Play Tennis

Output → Y=9 N=5

$E(L) = -P(Y) \log P(Y) - P(N) \log P(N)$

$= -\frac{9}{14} \log(\frac{9}{14}) - \frac{5}{14} \log(\frac{5}{14})$

$= -\frac{9}{14} \times (-0.64) - \frac{5}{14} (-1.49)$

$= 0.41 + 0.54$

$= 0.95$

Entropy of Individual class in features

Outlook	Play Tennis		TV	P(Y)	P(N)
sunny	Y=2	N=3	5	2/5	3/5
overcast	Y=4	N=0	4	4/4	0/4
rainfall	Y=3	N=2	5	3/5	2/5

$E(\text{sunny}) = -(\frac{2}{5}) \log(\frac{2}{5}) - (\frac{3}{5}) \log(\frac{3}{5})$

$= -0.4 \times (-1.32) - 0.6 \times (-0.74)$

$= 0.53 + 0.44$

$= 0.97$

$E(\text{overcast}) = -(\frac{4}{4}) \log(\frac{4}{4}) - \frac{0}{4} \log(\frac{0}{4})$

$= 0$

$E(\text{rainfall}) = -\left(\frac{3}{5}\right)\log\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log\left(\frac{2}{5}\right)$

$= -0.6 \times (-0.74) - 0.4 \times (-1.32)$

$= 0.42 + 0.53$

$= 0.97$

$$\boxed{E(\text{class}) = \sum \frac{\text{No of Observations}}{\text{Total no. of observation}} \times E_i}$$

$E(\text{outlook}) = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$

$= 0.35 + 0.35$

$= 0.70$

$$\boxed{\text{Information Gain} = E_{before} - E_{after}}$$

$IG = 0.95 - 0.70$

$= \underline{0.25}$

Temperature	Play Tennis		TV	P(Y)	P(N)
hot	Y = 2	N = 2	4	2/4 = 1/2	2/4 = 1/2
mild	Y = 4	N = 2	6	4/6 = 2/3	2/6 = 1/3
cool	Y = 3	N = 1	4	3/4	1/4

$E(\text{mild}) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)$

$= -\frac{2}{3} \times (-0.59) - \frac{1}{3} \times (-1.59)$

$= 0.39 + 0.53 = 0.92.$

$E(\text{hot}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$

$= -\frac{1}{2}(-1) - \frac{1}{2}(-1)$

$= 1.$

$E(\text{cool}) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)$

$= -\frac{3}{4}(-0.42) - \frac{1}{4}(-2)$

$= 0.32 + 0.5$

$= 0.82$

$$E(\text{Temperature}) = \frac{4}{14}\times 1 + \frac{6}{14}\times 0.92 + \frac{4}{14}\times 0.82$$
$$= 0.28 + 0.39 + 0.23$$
$$= 0.90$$

$$I.G = 0.95 - 0.90$$
$$= \underline{0.05}$$

Humidity	Play Tennis	TV	P(Y)	P(N)
high	Y = 3 N = 4	7	3/7	4/7
normal	Y = 6 N = 1	7	6/7	1/7

$$E(\text{high}) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)$$
$$= -\frac{3}{7}(-1.22) - \frac{4}{7}(-0.81)$$
$$= 0.52 + 0.46$$
$$= 0.98$$

$$E(\text{normal}) = -\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right)$$
$$= -\frac{6}{7}(-0.22) - \frac{1}{7}(-2.80)$$
$$= 0.19 + 0.4$$
$$= 0.59$$

$$E(\text{Humidity}) = \frac{7}{14}\times 0.98 + \frac{7}{14}\times 0.59$$
$$= 0.49 + 0.295$$
$$= 0.79$$

$$IG = 0.95 - 0.79$$
$$= \underline{0.16}$$

Wind	Tennis play		TV	P(Y)	P(N)
Weak	Y = 6	N = 2	8	$6/8 = \frac{3}{4}$	$2/8 = \frac{1}{4}$
Strong	Y = 3	N = 3	6	$3/6 = \frac{1}{2}$	$3/6 = \frac{1}{2}$

$$E(\text{Weak}) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)$$

$$= -\frac{3}{4}\log(-0.42) - \frac{1}{4}\log(-2)$$

$$= 0.32 + 0.5$$

$$= 0.82$$

$$E(\text{Strong}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)$$

$$= \frac{1}{2} + \frac{1}{2} = 1$$

$$E(\text{Wind}) = \frac{8}{14} \times 0.82 + \frac{6}{14} \times 1$$

$$= 0.47 + 0.43$$

$$= 0.90$$

$$IG = 0.95 - 0.90$$

$$= \underline{0.05}$$

- As outlook feature has maximum Information Gain, we will consider it as Root Node.

Gini Impurity

$$\boxed{G = 1 - \sum_{i=1}^{c} (P_i)^2}$$

$$G(L) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$= 1 - \frac{81}{196} - \frac{25}{196}$$

$$= \frac{90}{196} = 0.46$$

Outlook

$$G(\text{sunny}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 1 - \frac{9}{25} - \frac{4}{25} = \frac{12}{25} = 0.48$$

$$G(\text{overcast}) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$G(\text{rainfall}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\boxed{G(\text{class}) = \sum \frac{\text{No. of observations}}{\text{Total no. of observations}} \times G_i}$$

$$G(\text{outlook}) = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$$

$$= 0.17 + 0 + 0.17$$

$$= 0.34$$

$$IG = 0.48 - 0.34$$

$$= 0.14$$

Temperature

$$G(\text{hot}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2$$

$$= 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2} = 0.5$$

$$G(\text{mild}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$$

$$= 1 - \frac{4}{9} - \frac{1}{9} = \frac{4}{9} = 0.44$$

$$G(\text{cool}) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= 1 - \frac{9}{16} - \frac{1}{16} = \frac{6}{16} = 0.375$$

$$G(\text{Temperature}) = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.44 + \frac{4}{14} \times 0.375$$

$$= 0.14 + 0.19 + 0.11$$

$$= 0.43$$

$$I.G = 0.48 - 0.43 = 0.05$$

Humidity

$$G(High) = 1-\left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$$
$$= 1 - \frac{9}{49} - \frac{16}{49}$$
$$= \frac{24}{49} = 0.48$$

$$G(normal) = 1-\left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2$$
$$= 1 - \frac{36}{49} - \frac{1}{49}$$
$$= \frac{12}{49} = 0.24$$

$$G(Humidity) = \frac{7}{14} \times 0.48 + \frac{7}{14} \times 0.24$$
$$= 0.24 + 0.12 = 0.36$$

$$IG = 0.48 - 0.36$$
$$= 0.12$$

Wind

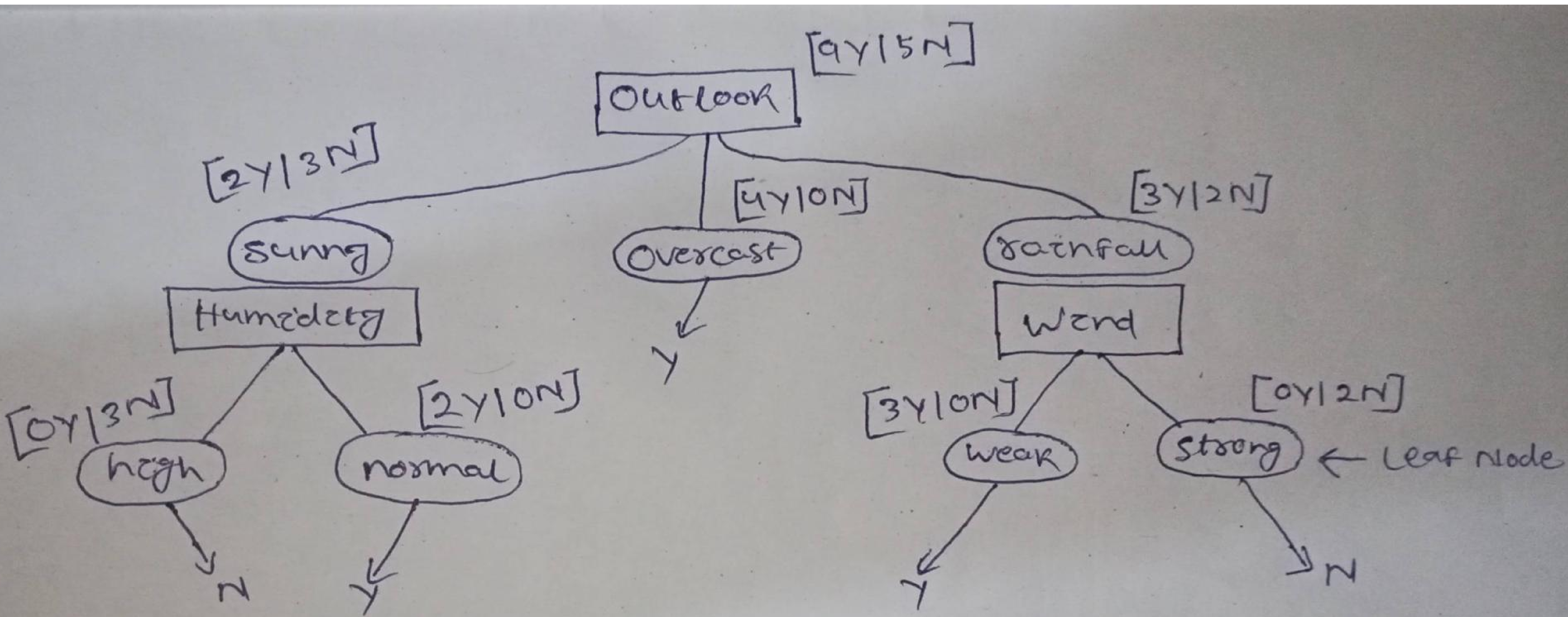
$$G(weak) = 1-\left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2$$
$$= 1 - \frac{36}{64} - \frac{4}{64}$$
$$= \frac{64-36-4}{64} = \frac{24}{64} = 0.375$$

$$G(strong) = 1-\left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$$
$$= 1 - \frac{1}{4} - \frac{1}{4} = \frac{2}{4} = 0.5$$

$$G(wind) = \frac{8}{14} \times 0.375 + \frac{6}{14} \times 0.5$$
$$= 0.21 + 0.21$$
$$= 0.42.$$

$$IG = 0.48 - 0.42$$
$$= 0.06$$

- With respect to the calculation of Gini Impurity 'outlook' has least impurity value and max information gain. So, we will consider it as root node.



- we won't split more, as we already got conclusion. Otherwise our model will be overfit.