# Decision Tree

## Play Tennis Dataset

- To findout the root node we will follow 2 approaches.

① Information Gain → Max Value

② Gini Impurity. → Min Value

## Information Gain

$$Entropy = \sum -P \log_2(P)$$

Label → Play Tennis

Output → $Y = 9$ $N = 5$

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

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

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

$$= 0.41 + 0.54$$

$$= 0.95$$

## 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 |

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

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

$$= 0.53 + 0.44$$

$$= 0.97$$

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

$$= 0$$

$$E(\text{rain fall}) = -\left(\tfrac{3}{5}\right)\log\left(\tfrac{3}{5}\right) - \left(\tfrac{2}{5}\right)\log\left(\tfrac{2}{5}\right)$$

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

$$= 0.42 + 0.53$$

$$= 0.97$$

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

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

$$= 0.35 + 0.35$$

$$= 0.70$$

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

$$IG = 0.95 - 0.70$$

$$= 0.25$$

| Temperature | Play Tennis | | TV | P(Y) | P(N) |
|---|---|---|---|---|---|
| hot | Y=2 | N=2 | 4 | $2/4 = \tfrac{1}{2}$ | $2/4 = \tfrac{1}{2}$ |
| mild | Y=4 | N=2 | 6 | $4/6 = \tfrac{2}{3}$ | $2/6 = \tfrac{1}{3}$ |
| cool | Y=3 | N=1 | 4 | $3/4$ | $1/4$ |

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

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

$$= 0.39 + 0.53 = 0.92.$$

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

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

$$= 1.$$

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

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

$$= 0.32 + 0.5$$

$$= 0.82$$

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

$$= 0.28 + 0.39 + 0.23$$

$$= 0.90$$

$$I \cdot G = 0.95 - 0.90$$

$$= \underline{\underline{0.05}}$$

| 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 |

$$E(\text{high}) = -\frac{3}{7} \log\left(\frac{3}{7}\right) - \frac{4}{7} \log\left(\frac{4}{7}\right)$$

$$= -\frac{3}{7}(-1.22) - \frac{4}{7}(-0.81)$$

$$= 0.52 + 0.46$$

$$= 0.98$$

$$E(\text{normal}) = -\frac{6}{7} \log\left(\frac{6}{7}\right) - \frac{1}{7} \log\left(\frac{1}{7}\right)$$

$$= -\frac{6}{7}(-0.22) - \frac{1}{7}(-2.80)$$

$$= 0.19 + 0.4$$

$$= 0.59$$

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

$$= 0.49 + 0.295$$

$$= 0.79$$

$$IG = 0.95 - 0.79$$

$$= \underline{\underline{0.16}}$$

| Wind | Tennis play | IV | 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 | $3/6 = \frac{1}{2}$ | $3/6 = \frac{1}{2}$ |

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

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

$$= 0.32 + 0.5$$

$$= 0.82$$

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

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

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

$$= 0.47 + 0.43$$

$$= 0.90$$

$$IG = 0.95 - 0.90$$

$$= 0.05$$

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

## Gini Impurity

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

### Outlook

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

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

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

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

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

$$= 0.17 + 0 + 0.17$$

$$= 0.34$$

### Temperature

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

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

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

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

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

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

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

$$= 0.14 + 0.19 + 0.11$$

$$= 0.43$$

## Humidity

$$G(\text{High}) = 1-\left(\tfrac{3}{7}\right)^2 - \left(\tfrac{4}{7}\right)^2$$

$$= 1 - \tfrac{9}{49} - \tfrac{16}{49}$$

$$= \tfrac{24}{49} = 0.48$$

$$G(\text{normal}) = 1-\left(\tfrac{6}{7}\right)^2 - \left(\tfrac{1}{7}\right)^2$$

$$= 1 - \tfrac{36}{49} - \tfrac{1}{49}$$

$$= \tfrac{12}{49} = 0.24$$

$$G(\text{Humidity}) = \tfrac{7}{14} \times 0.48 + \tfrac{7}{14} \times 0.24$$

$$= 0.24 + 0.12 = 0.36$$

## Wind

$$G(\text{weak}) = 1-\left(\tfrac{6}{8}\right)^2 - \left(\tfrac{2}{8}\right)^2$$

$$= 1 - \tfrac{36}{64} - \tfrac{4}{64}$$

$$= \tfrac{64-36-4}{64} = \tfrac{24}{64} = 0.375$$

$$G(\text{strong}) = 1-\left(\tfrac{1}{2}\right)^2 - \left(\tfrac{1}{2}\right)^2$$

$$= 1 - \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{2}{4} = 0.5$$

$$G(\text{wind}) = \tfrac{8}{14} \times 0.375 + \tfrac{6}{14} \times 0.5$$

$$= 0.21 + 0.21$$

$$= 0.42.$$

– With respect to the calculation of Gini Impurity 'outlook' has least value, So, we will consider it as root node.

Decision tree diagram:

Outlook [9Y15N]
- Sunny [2Y13N] → Humidity
  - High [0Y13N] → N
  - Normal [2Y10N] → Y
- Overcast [4Y10N] → Y
- Rainfall [3Y12N] → Wind
  - Weak [3Y10N] → Y
  - Strong [0Y12N] → N