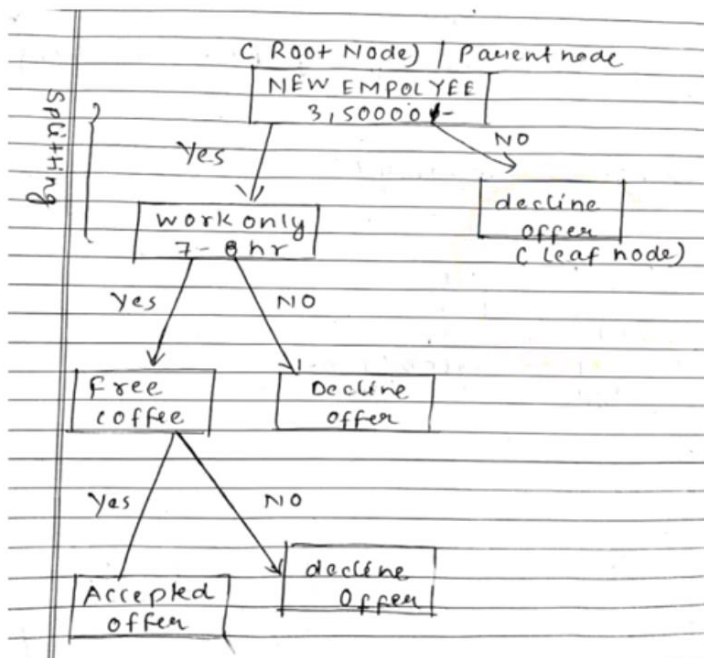# What is Decision tree

A decision tree is a graphical representation of possible solution to a decision based on certain condition it's called a decision tree.

How does a tree decide where to split

1.GINI INDEX
Measure the impurity used to build a decision tree.

2. Information Gain
Select the node will be highest information gain.

3.Reduction in variance
If your data is pure then less variance in data.
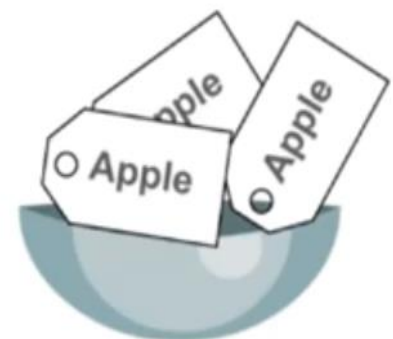Lowest variance is good for creating a variance

entropy

- Measure the impurity of the substances.

- What is impurity
- Impurity=0
- Impurity ≠ 0

$$Entropy(s) = - P(yes) \log_2 P(yes) - P(no) \log_2 P(no)$$

Entropy(features)= -P(yes)log2P(Yes) -P(No)log2P(No)
Entropy(outlook=sunny)=
P(yes)

Information gain

⬚ Decide which attribute should be selected as the decision tree node.

Information Gain = Entropy(S) – [(Weighted Avg) x Entropy(each feature)]

Day	outlook	Temperature	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

Soln :

Step ①) Out of 14 Instances we have 9 Yes
And 5 NO

compute the Entropy of entire data-set

$$E(S) = -P(Yes) \log_2 P(Yes) - P(NO) \log_2 P(NO)$$

$$= -\left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \frac{5}{14}$$
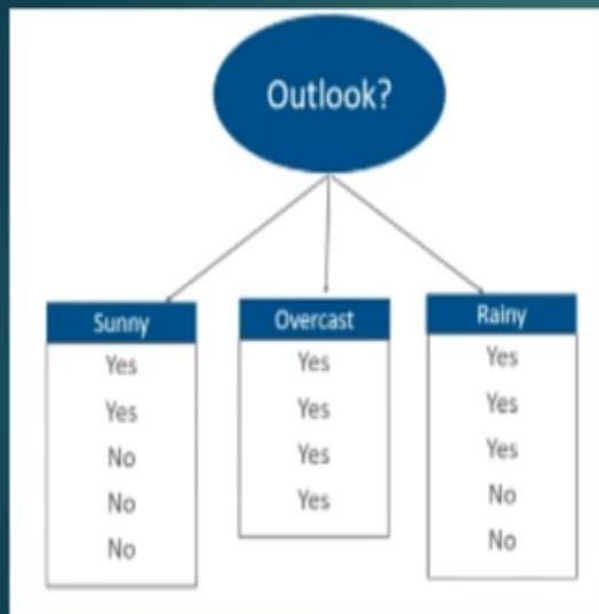
$$E(S) = 0.41 + 0.53$$

$$= 0.94$$

Find Entropy

Which node select as root node

Outlook

temperature

humidity

windy

Handwritten notes:

$$E(\text{Outlook} = \text{SUNNY})$$
$$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$
$$= 0.971$$

$$E(\text{Outlook} = \text{Overcast})$$
$$= 0$$

$$E(\text{Outlook} = \text{Rainy}) = 0.971$$

Information from Outlook

$$I(\text{Outlook}) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971$$
$$= 0.693$$

Information gained from Outlook

$$\text{Gain}(\text{Outlook}) = E(S) - I(\text{Outlook})$$
$$= 0.94 - 0.693$$

Entropy(features)= -P(yes)log2P(Yes) -P(No)log2P(No)
Entropy(outlook=sunny)= -⅖*log(⅖)-⅗*log(⅗) =0.971
Entropy(outlook=overcast)= -4/4*log2(1)-0 =0
E(outlook=Rainy)= -⅗*log2(⅗)-⅖*log2(⅖)=0.971
Information from outlook=sum of(weight*E(each)
Information from outlook=5/14*0.971+4/14*0+5/14*0.971=0.693
Information gained from outlook =E(S)-Information(Outlook)

Entropy(target)
Entropy(Play/tennis)=-P(yes)log2(P(yes)-P(NO)*log2(No)

P(yes)=9/14   P(no) =4/14

Entropy(Target)=-9/14*log2(9/14)-(5/14)*log2(5/14)=0.94

Information gained from outlook =E(S)-Information(Outlook)

E(S)=entropy of target

IG=0.94-0.693=0.247(outlook)
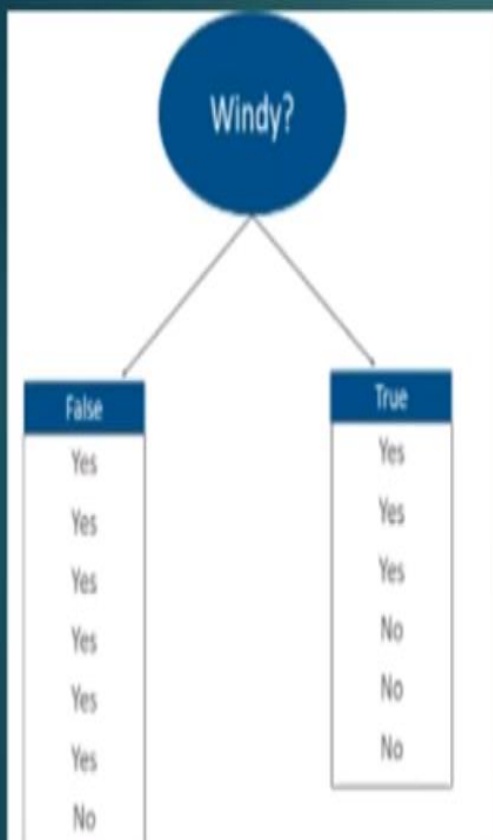
$E(windy = TRUE) = 1$

$E(wind = FALSE) = 0.811$

Information from windy

$I(windy) = \dfrac{8}{14} \times 0.811 + \dfrac{6}{14} \times 1$

$= 0.892$

Information gained from windy

gain(windy) =

$E(S) - I(windy)$

$= 0.94 - 0.892$

$= 0.048$

find out Rest of two with Same Producer.

Outlook		Temperature	
	max	Information :	0.811
Information : 0.693		Gain	: 0.029
Gain : 0.247			

humidity		windy	
Information: 0.788		Information :	0.892
Gain : 0.152		Gain	: 0.048

Windy?

False: Yes, Yes, Yes, Yes, Yes, Yes, No

True: Yes, Yes, Yes, No, No, No

```
                    ┌──────────┐
                    │ OUTLOOK  │
                    └──────────┘
          SUNNY       │          RAINY
     ┌──────────┐ ┌──────────┐ ┌────────┐
     │ humidity │ │ overcast │ │ windy  │
     └──────────┘ └──────────┘ └────────┘
                      yes
                              ┌────────┐   ┌────────┐
                              │ STRONG │   │ WEAK   │
     ┌──────┐  ┌────────┐     └────────┘   └────────┘
     │ HIGH │  │ NORMAL │
     └──────┘  └────────┘
       NO        YES            NO            YES
```