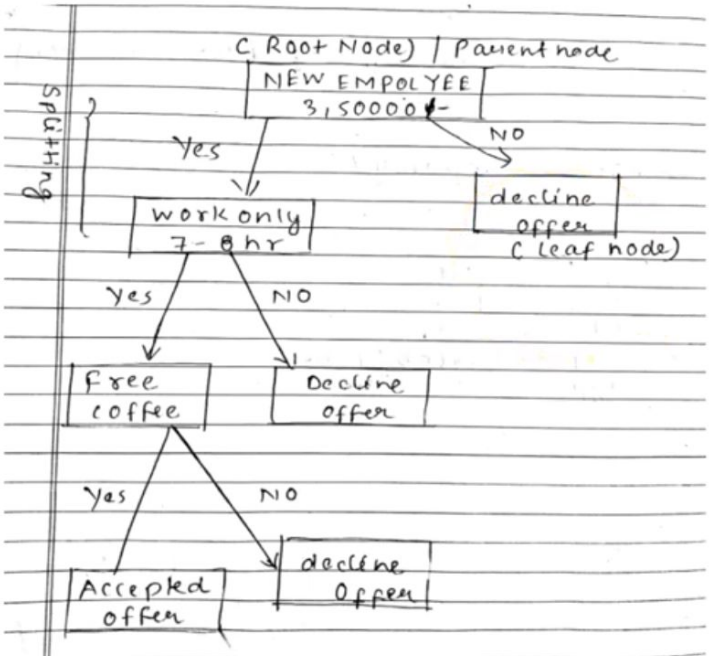
What is Decision tree

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

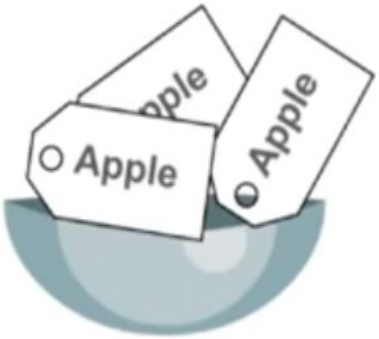
If your data is pure then less variance in data.

Lowest variance is good for creating a variance

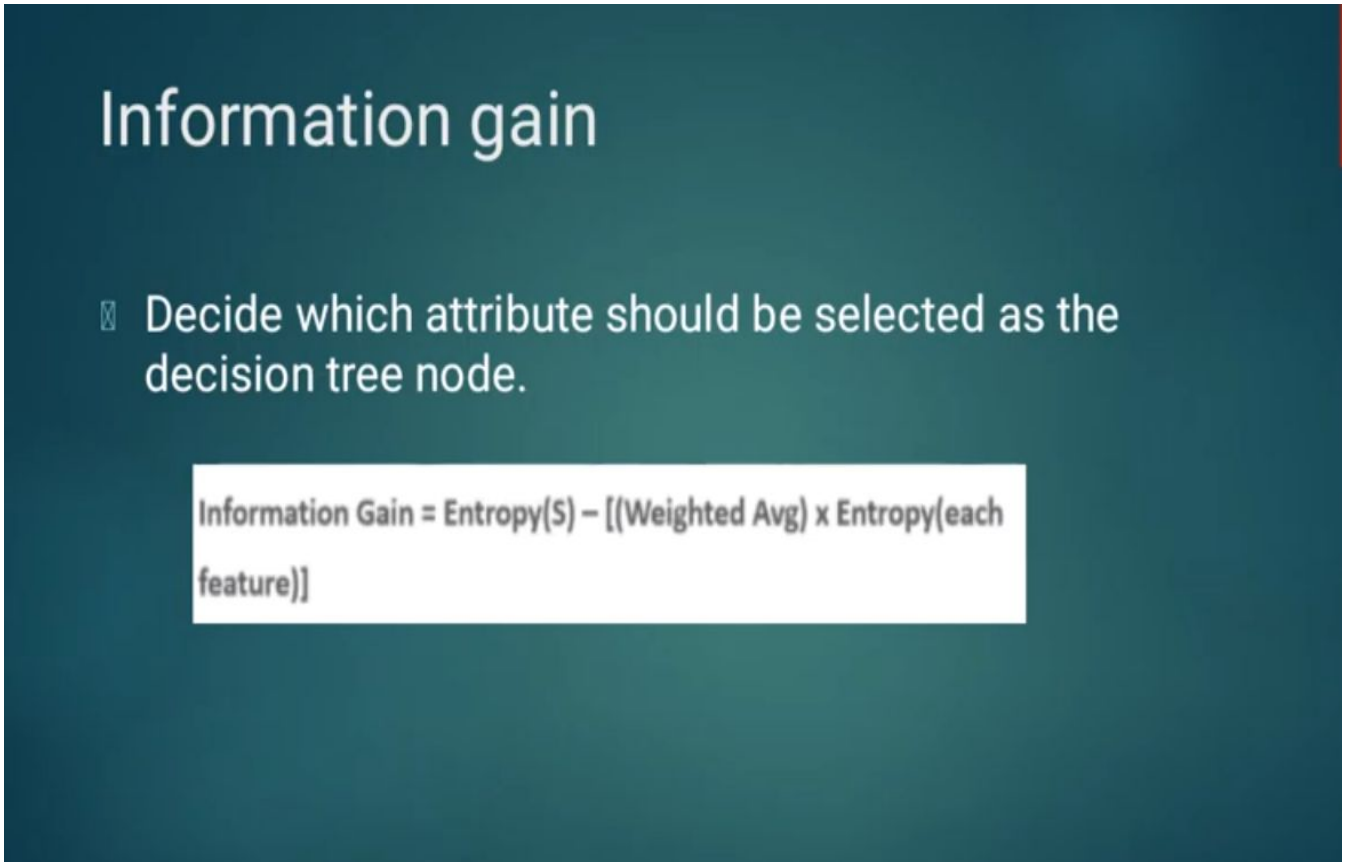
entropy

- Measure the impurity of the substances.
- What is impurity
- Impurity=0
- Impurity ≠ 0

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



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



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

Step 0) Out of 14 Instances we have 9 Yes And 5 NO

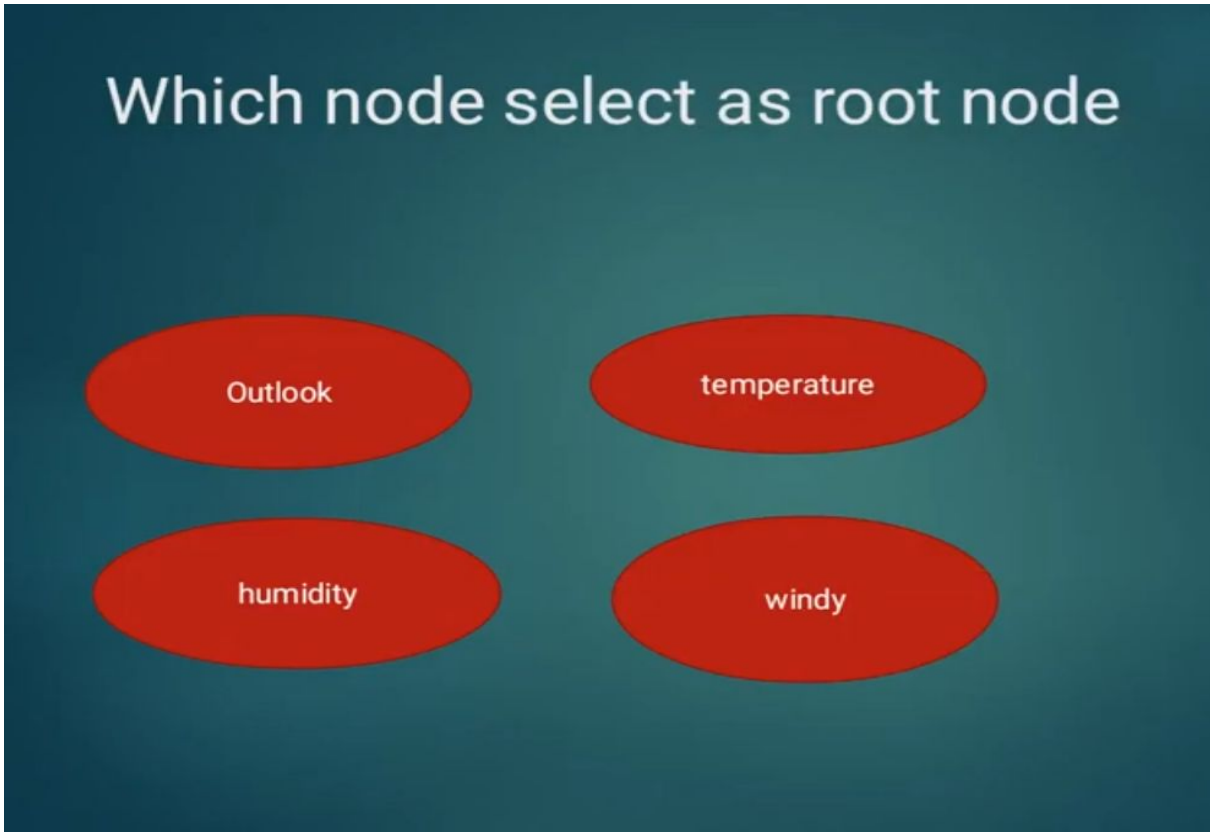
compute the Entropy of entire data-set

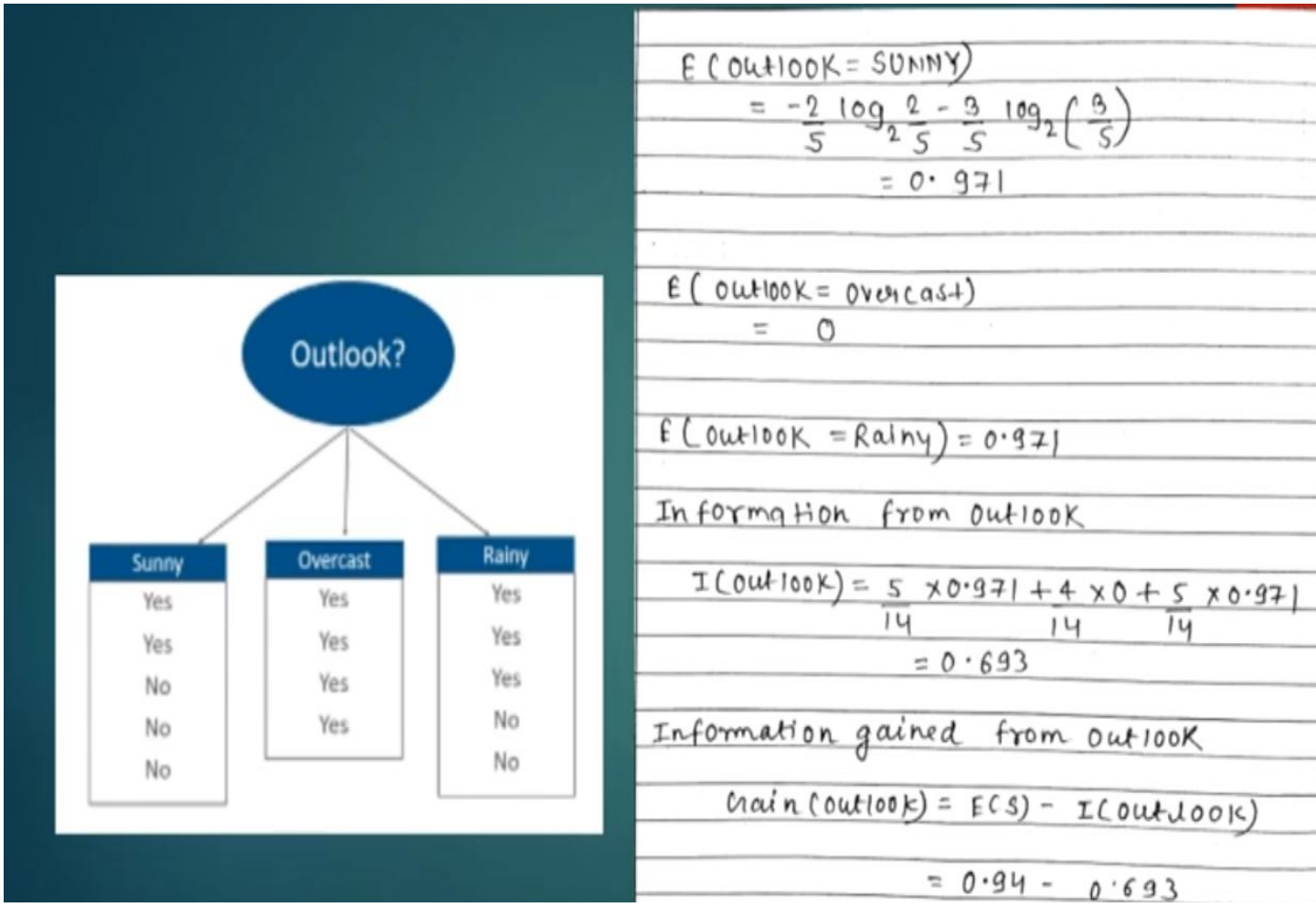
$$E(S) = -P(Yes)\log_2 P(Yes) - P(NO)\log_2 P(NO)$$

$$= -\left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2 \frac{5}{14}$$

$$E(S) = 0.41 + 0.53$$

$$= 0.94$$


Which node select as root node
Outlook
temperature
humidity
windy



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)

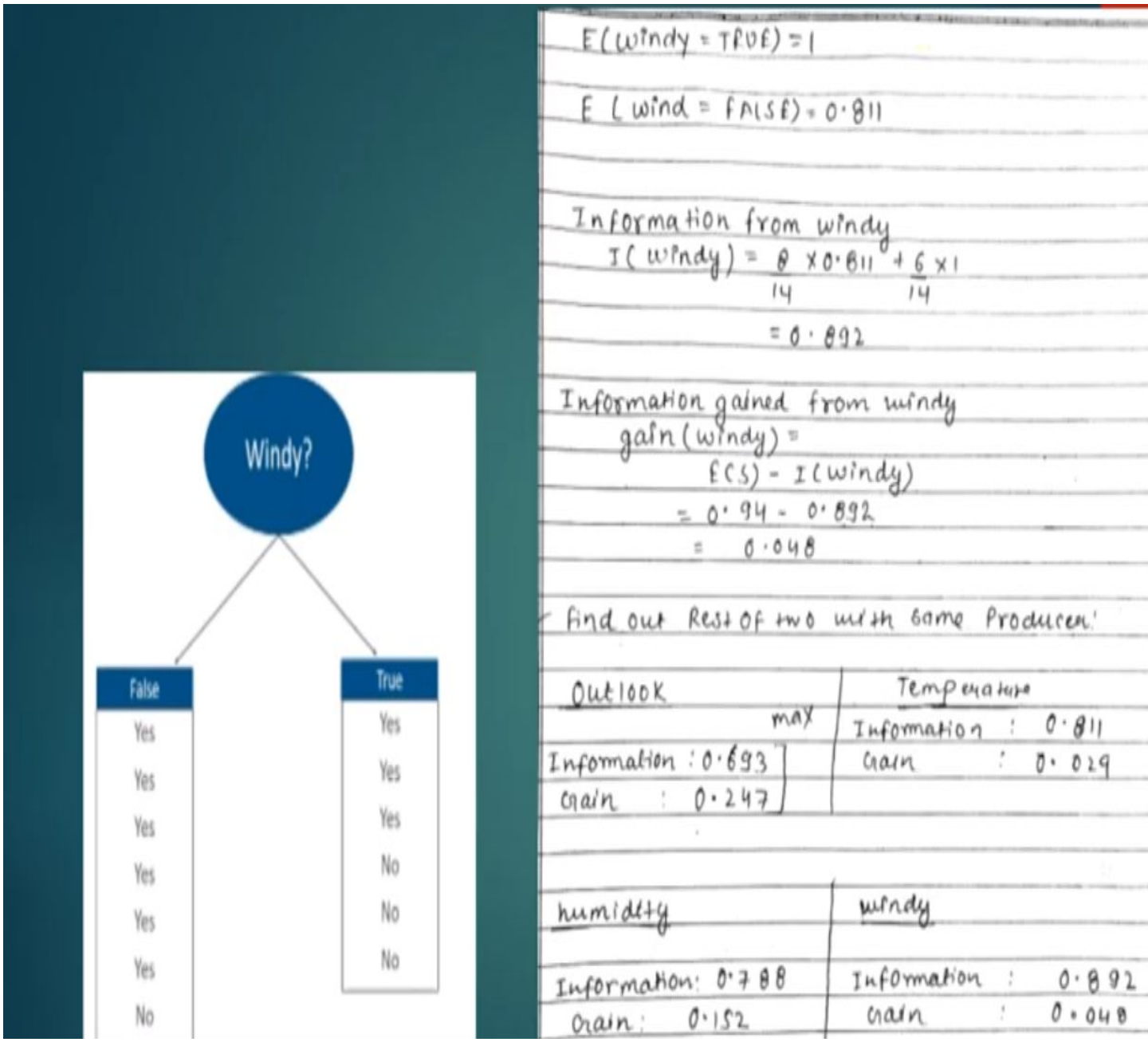
P(yes)=9/14 P(no) =4/14

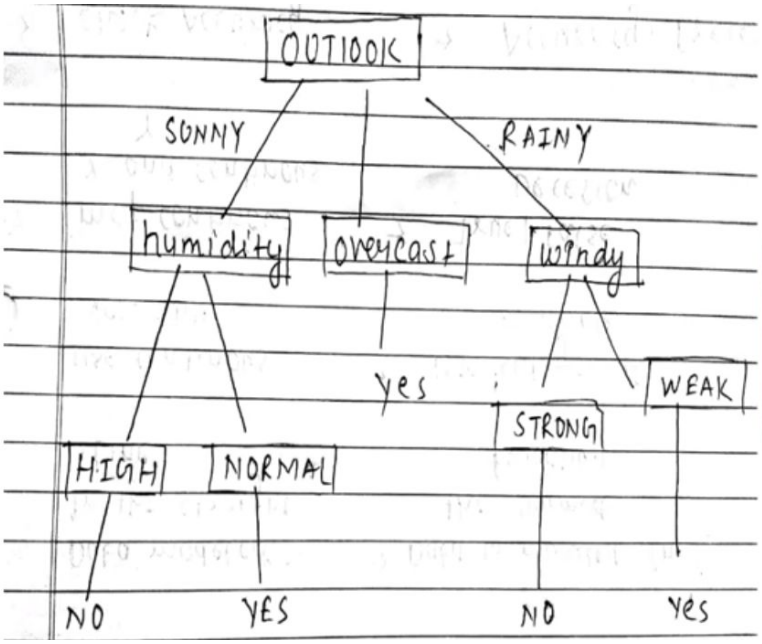
Entropy(Target)=-9/14*log2(9/14)-(5/14)*log2(5/14)=0.94

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

E(S)=entropy of target

IG=0.94-0.693=0.247(outlook)




OUTLOOK
SUNNY
RAINY
humidity
overcast
windy
yes
WEAK
STRONG
HIGH
NORMAL
NO
YES
NO
yes