Decision Trees & Random Forest

* Decision tree is a classifier.
* Simple tree like structure, model makes a decision at every node.
* Useful in simple tasks,
* One of the most popular algorithms.
* Easy explainability, easy to show how a desision process works.

why so popular ?

* Easy to implement & present.
* Well defined logic, mimic human level thought.
* Random forests, Ensembles of desicion trees are more powerful classifiers.
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.

Build Desicion Trees :

Two common algorithms :

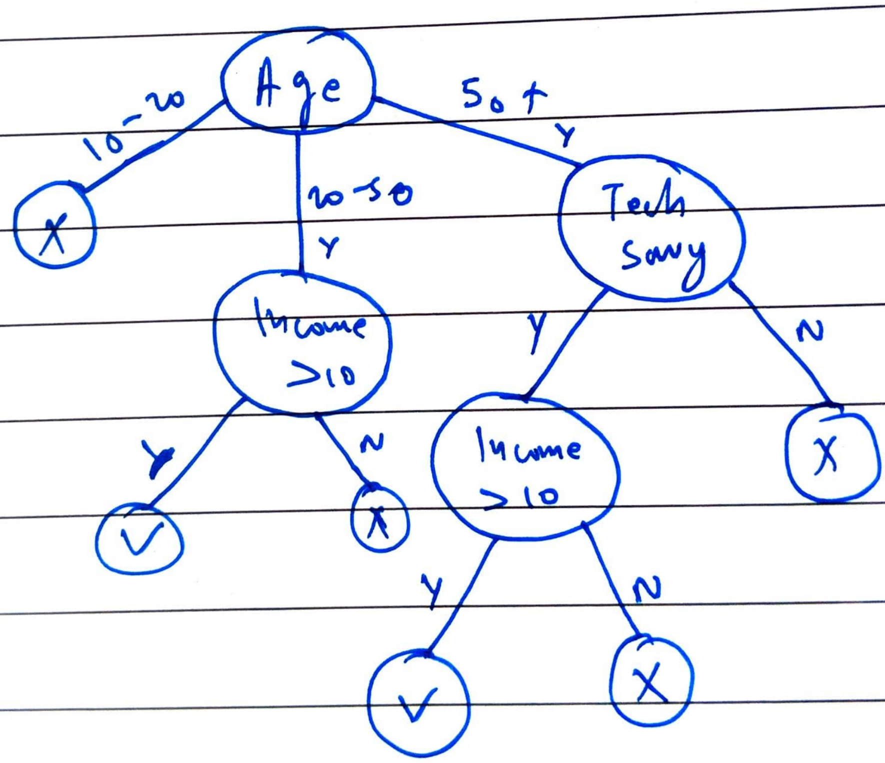
* CART (Classification & Regression Trees) → Uses Gini Index (classification) as metric.
* ID3 (Iterative Dichotomiser 3) → uses Entropy function & information gain as metrics.

Problem: To predict whether a customer will buy a self-driving car.

Given features / attributes about person:

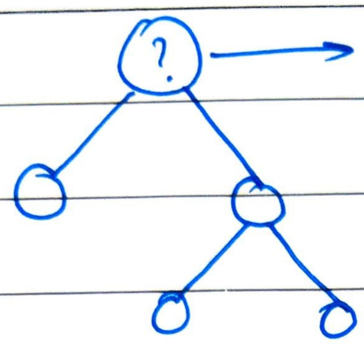
Sex	Income	Car (current)	Tech Savvy	Age
M	<5 lac	Yes	Yes	10-20
F	5-10	No	No	20-50
	>10 lac			50+

Example:



* Important factor is to decide which feature to choose at every node.

Entropy & Information Gain:



which feature should come here? we figure that out by measuring:

① Entropy ② Information Gain

* **Entropy** measures randomness of system.

Ex: Say a box contains 3 R & 3 B balls.

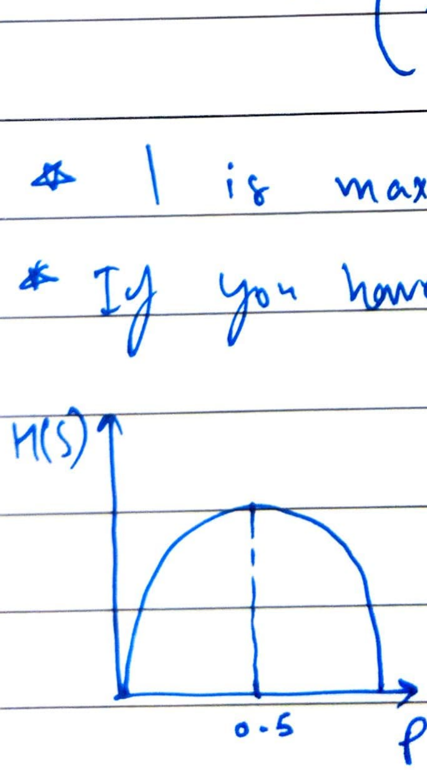
Randomness / entropy = $\boxed{H(S) = -\sum p_c \log p_c}$

where p_c = Prob of class c

In ex, $p_R = \frac{1}{2} = p_B$

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

* 1 is max entropy (when same no. of examples)
* If you have balls of one type, then $H(S) = 0$



Example of training data:

outlook	temp	humidity	windy	play (outcome)
Sunny	Hot	High	F	No
S	H	H	T	N
overcast	H	H	F	Yes
Rainy	Mild	H	F	Y
R	Cool	Normal	F	Y
R	C	N	T	N
O	C	N	T	Y
S	M	H	F	N
S	C	N	F	Y
R	M	N	F	Y
S	M	N	T	Y
O	M	H	T	Y
O	H	N	F	Y
R	M	H	T	N

Entropy of System in beginning:

Class $\begin{bmatrix} \text{Yes} \\ \text{No} \end{bmatrix} \begin{matrix} \rightarrow 9 \\ \rightarrow 5 \end{matrix}$

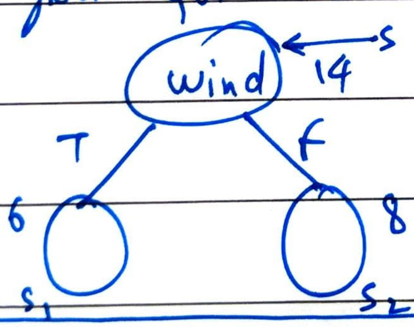
$$H(s) = -\left(\frac{9}{14} \log \frac{9}{14} + \frac{5}{14} \log \frac{5}{14} \right)$$

$$= 0.41 + 0.53 = 0.94$$

Now, when we construct DT:

* We will try to put all attributes one by one as root & see how much is reduction in entropy. This reduction is called information gain.

Say, you put wind:



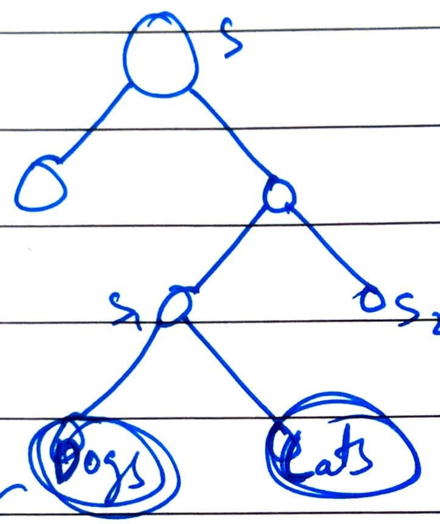
$$\boxed{IG(S, A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)}$$

Splitting about A

$$= H(S) - \left[\frac{6}{14} H(S_1) + \frac{8}{14} H(S_2) \right]$$

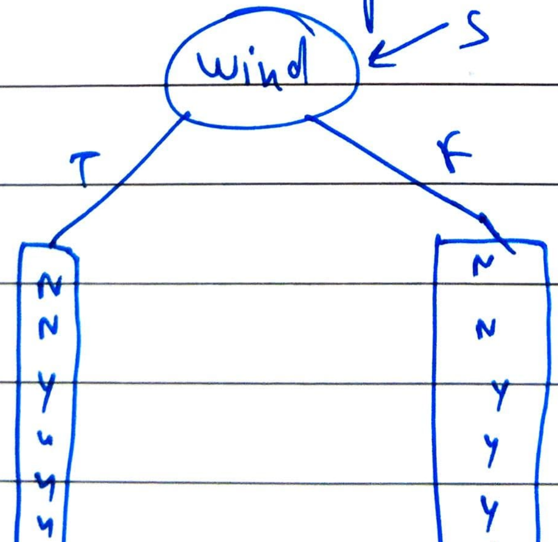
* Goal is to maximize Information Gain (IG).

Now, suppose (duh) in the end you will only have unique values at leaf node, say only dogs or cats.



$H(\text{Dogs}) = 0 = H(\text{cats})$ as it only has single class and $P = 1$.

* In previous example:



$$IG = H(S) - \text{New System entropy.}$$

$$= H(S) - \sum \frac{|S_v|}{S} H(S_v)$$

$$IG = H(S) - \left[\frac{8}{14}\left(-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right) + \frac{6}{14}\left(-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}\right) \right]$$

$$= 0.94 - 0.892 = 0.048$$

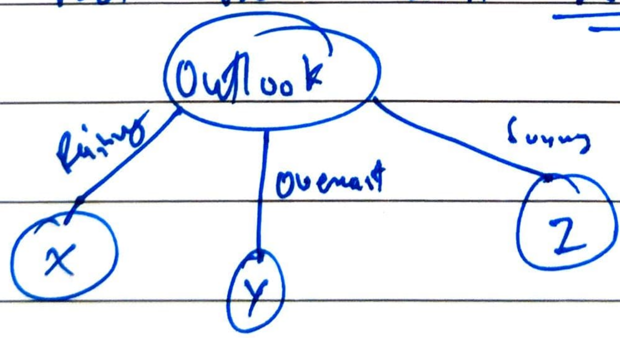
Hence, $IG(S, \text{wind}) = 0.048$

Similarly, $IG(S, \text{Outlook}) = 0.247$

$IG(S, \text{Temp}) = 0.029$

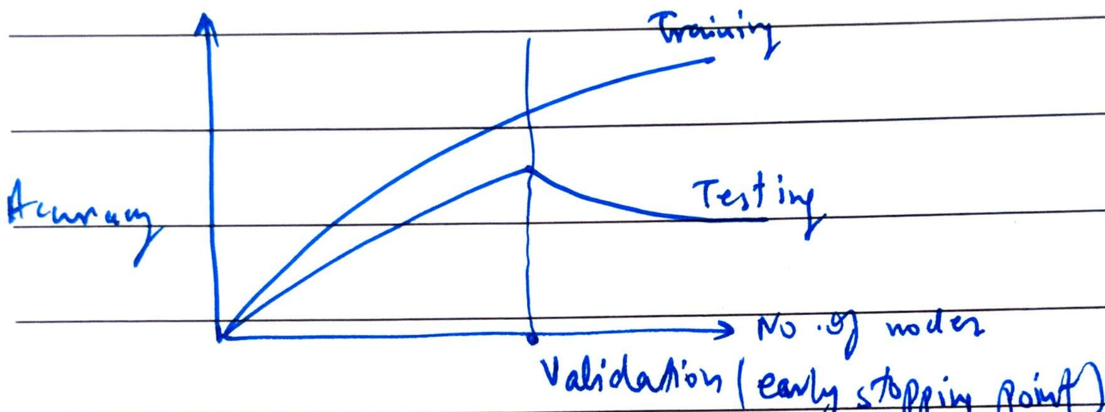
$IG(S, \text{humidity}) = 0.15$

* So, our root node will be "outlook"



* Now, repeat the same process for X, Y, Z.
* Repeat till you get nodes having all yes or No.

* Decision trees can overfit, say you have 5 examples & no. of leaf nodes = 5.
* You can reduce overfitting by restricting a tree to a certain depth.



* (I) You can also prevent overfitting by Post Prunning, create full tree then remove those nodes/subtrees that give poor generalisation.
* (II) Early stopping : Stop at a certain depth.

* Can also overfit if one node (say leaf node) has only low no. of examples like 1 or 2 etc. Hence, you can build tree such as min. no. of examples at a node are required to split that node.