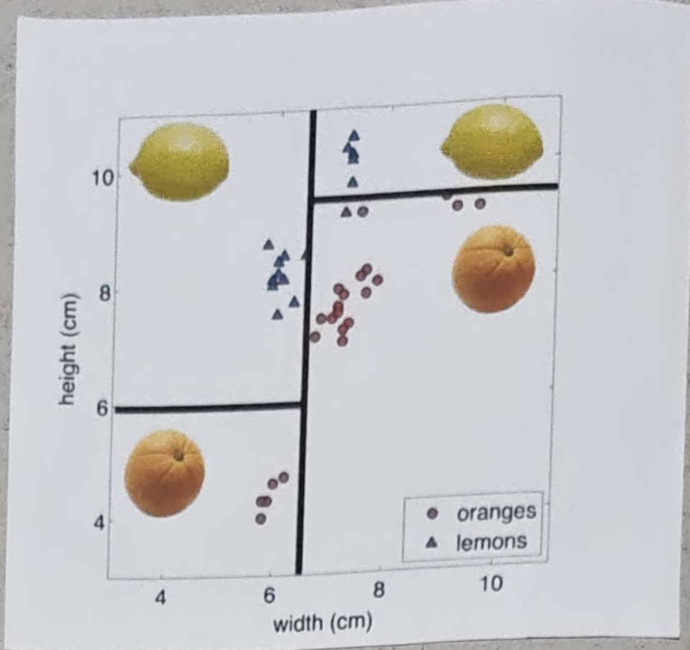
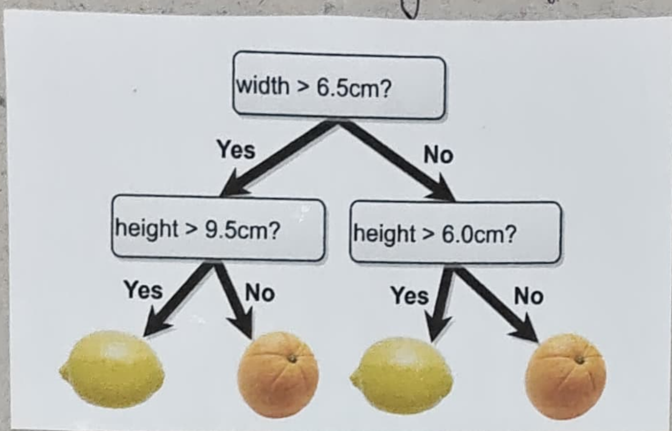
9. Decision Tree Classifier (DTC)

Consider the simple dataset shown, comprises of heights and weights of lemons and oranges.



Below DTC is built using the given dataset.



Advantages
- interpretable, intuitive
- Popular in medical diagnosis applications

How to choose the best attribute ?

One can use principles of information theory for this.

ENTROPY.

Consider a source Y with symbols y_i and probabilities $p(y_i)$.
The Entropy $H(Y)$ of this source is defined as

$$H(Y) = \sum_{i=1}^{n} p(y_i) \log_2 \frac{1}{p(y_i)}$$

Example :

$$p(0) = \frac{1}{4}, \quad p(1) = \frac{3}{4}; \quad H(Y) = ?$$

$$H(Y) = \frac{1}{4} \cdot \log_2 4 + \frac{3}{4} \log_2 \frac{4}{3}$$

$$= \frac{1}{2} + \frac{3}{4}\left(\frac{\ln\left(\frac{4}{3}\right)}{\ln(2)}\right)$$

$$\doteq 0.811 \text{ bits/symbol}$$

DTC example

Fri / sat

Hungry

Patrons

Reservation

Estimated wait time

Decision

Example	Input Attributes										Goal
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	WillWait
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	y_1 = Yes
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	y_2 = No
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	y_3 = Yes
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	y_4 = Yes
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	y_5 = No
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	y_6 = Yes
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	y_7 = No
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	y_8 = Yes
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	y_9 = No
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	y_{10} = No
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	y_{11} = No
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	y_{12} = Yes

Consider the table shown above.

<u>customer decision</u> to wait or not at restaurants.

Table columns :

1.	Alternate: whether there is a suitable alternative restaurant nearby.
2.	Bar: whether the restaurant has a comfortable bar area to wait in.
3.	Fri/Sat: true on Fridays and Saturdays.
4.	Hungry: whether we are hungry.
5.	Patrons: how many people are in the restaurant (values are None, Some, and Full).
6.	Price: the restaurant's price range ($, $$, $$$).
7.	Raining: whether it is raining outside.
8.	Reservation: whether we made a reservation.
9.	Type: the kind of restaurant (French, Italian, Thai or Burger).
10.	WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

Problem 1 :

$\boxed{1}$	$\boxed{3}$	$\boxed{4}$	$\boxed{6}$	$\boxed{8}$	$\boxed{12}$
$\boxed{2}$	$\boxed{5}$	$\boxed{7}$	$\boxed{9}$	$\boxed{10}$	$\boxed{11}$

What is the entropy of the decision $Y \in \{Yes, No\}$?

$$P(Y = Yes) = \frac{1}{2} \quad , \quad P(Y = No) = \frac{1}{2}$$

$$H(Y) = H\left(\frac{1}{2}; \frac{1}{2}\right)$$

$$= \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2$$

$$= \underline{1} \text{ bits/symbol.}$$

CONDITIONAL ENTROPY

Consider two sources : Y with symbols y_i
X with symbols x_j

The Conditional Entropy $H(Y|X)$ is defined as

$$\sum_{j=1}^{m} p(x_j)\, H(Y | X = x_j)$$

Example: Cricket

WT – Winning Toss | WG – Winning Game

LT – Losing Toss | LG – Losing Game.

$$WT = \frac{1}{4}\left\{WG = \frac{5}{6},\ LG = \frac{1}{6}\right\}$$

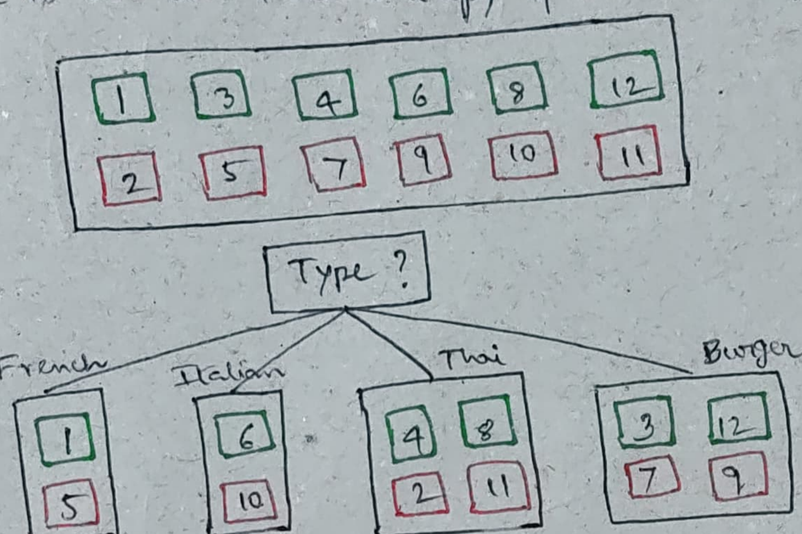
$$LT = \frac{3}{4}\left\{WG = \frac{1}{5},\ LG = \frac{4}{5}\right\}$$

$$\Rightarrow \frac{1}{4} \times \left\{\frac{5}{6}\log_2\frac{6}{5} + \frac{1}{6}\log_2 6\right\}$$

$$+ \frac{3}{4} \times \left\{\frac{1}{5}\log_2 5 + \frac{4}{5}\log_2\frac{5}{4}\right\}$$

Problem 2:

What is the conditional entropy of the TYPE feature?

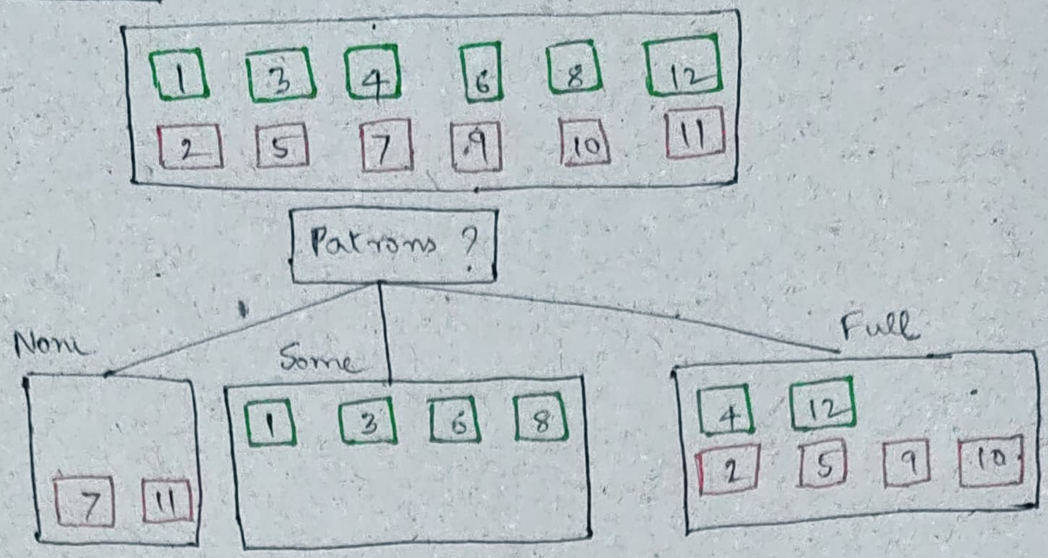


$$P(Fr) \times H(Y|Fr) + P(It) \times H(Y|It)$$
$$+ P(Th) \times H(Y|Th)$$
$$+ P(Bu) \times H(Y|Bu)$$

$$= \left(\frac{2}{12} \times 1\right) + \left(\frac{2}{12} \times 1\right) + \left(\frac{4}{12} \times 1\right) + \left(\frac{4}{12} \times 1\right)$$

$$= 1.$$

Problem 3 :



What is the conditional entropy of the PATRONS feature ?

$$P(\text{None}) * H(y \mid \text{None}) + P(\text{Some}) * H(y \mid \text{Some}) + P(\text{Full}) * H(y \mid \text{Full})$$

$$= \left(\frac{2}{12} \times 0\right) + \left(\frac{4}{12} \times 0\right) + \left(\frac{1}{2} \times H\left(\frac{1}{3}, \frac{2}{3}\right)\right)$$

$$= 0.459.$$

INFORMATION GAIN

The Information Gain (IG) is defined as

$$IG(x) = H(y) - H(y \mid x)$$

Choose the feature that maximizes the IG !

Example :

- IG for the TYPE feature is given as

$$IG(\text{TYPE}) = H(y) - H(y \mid x = \text{TYPE})$$
$$= 1 - 1 = 0$$

- IG for the PATRONS feature is given as

$$IG(\text{PATRONS}) = H(y) - H(y \mid x = \text{PATRONS})$$
$$= 1 - 0.459 = 0.541.$$

DTC example.

Since $IG(Patrons) = 0.541 > IG(Type) = 0$

Therefore, we choose PATRONS as the feature to split.

Final DTC