Decision Tree

05 June 2023 20:38

* It comes under Supervised learning.
* It is used to solve both classification and Regression problems.

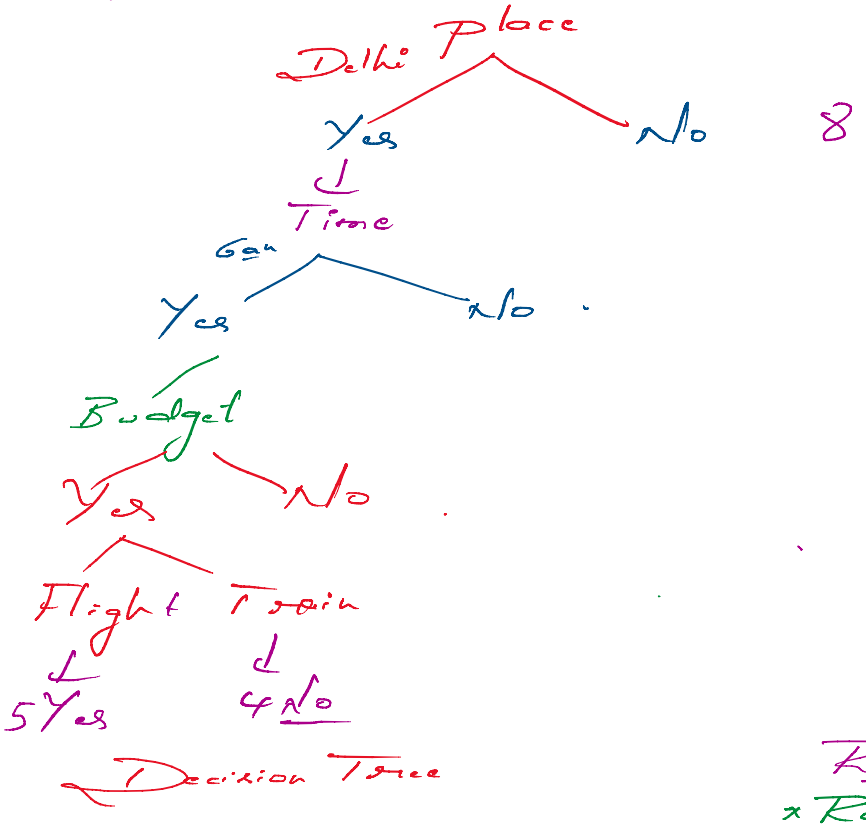
Types of Decision Tree

1. Decision Tree Regression
   It is used when target is continuous.
2. Decision Tree Classifier
   It is used when target is categorical / discrete

Decision Trees use logic to understand data and to make predictions.

Logic → if-else conditions

Eg:- Place, Time, Budget, Transportation



Decision Tree

Root Node

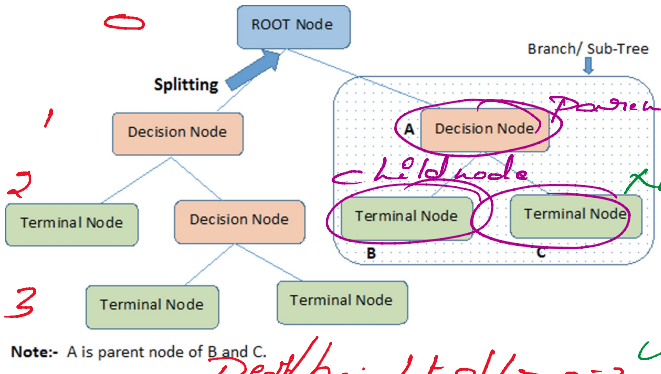
x Root node represents entire population and this further gets splitted into two or more homogeneous sets.

Splitting

It is a process of dividing node into two or more sub nodes

Decision node

. . . Further



Note:- A is parent node of B and C.

Depth/height of tree = 3

node into two -

Decision node

* when subnode splits into further subnodes. we call it as Decision node
* Decision node will have child nodes.

Parent

Child node

Terminal node / Leaf node.

Node which has no children [no further splits]

Branch / Subtree

A subsection of decision tree

Pruning

The process of removing nodes inorder to reduce the size of tree.

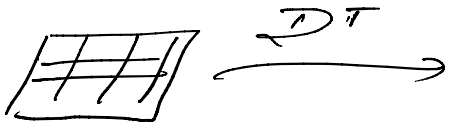
Parent node / child node

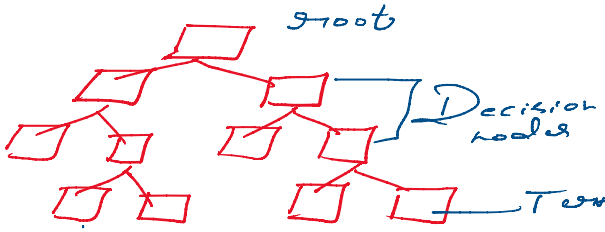
A node which is divided into subnodes are called as parent node.

Subnodes of parent node is called as child node.

How Decision works?

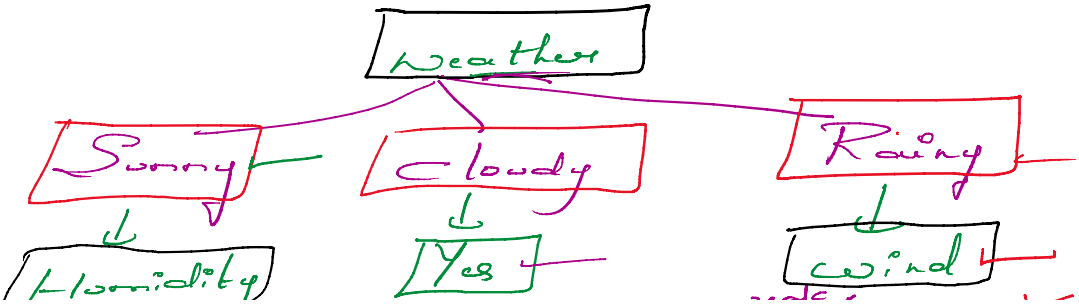
Decision tree builds classification or regression model in the form of tree structure. It break down entire data into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree which includes decision nodes and leaf nodes



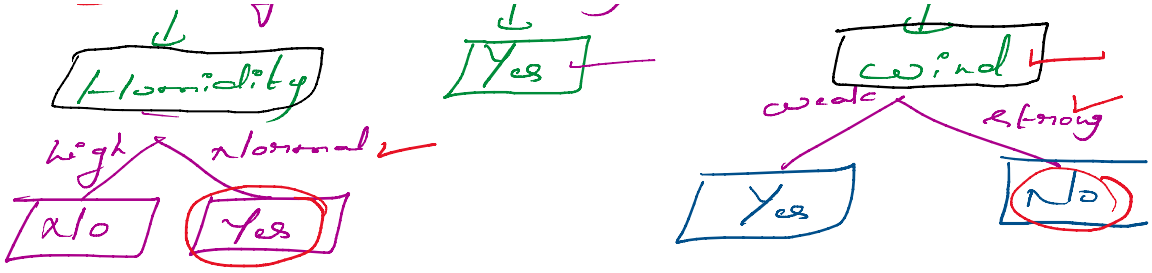


Weather	Humidity	Wind	play
sunny	high	weak	no
sunny	high	strong	no
cloudy	high	weak	yes
rainy	high	weak	yes
rainy	normal	weak	yes
rainy	normal	strong	no
cloudy	normal	strong	yes
sunny	high	weak	no

Decision Tree



cloudy	normal	strong	yes
sunny	high	weak	no
sunny	normal	weak	yes
rainy	normal	weak	yes
sunny	normal	strong	yes
cloudy	high	strong	yes
cloudy	normal	weak	yes
rainy	high	strong	no



Rainy / strong wind

cloudy, high, weak

Decision Tree working procedure

* It will search for root node among input variables.

In this example, we need to select root node among weather, windy and Humidity.

How to Select root node?

→ Information gain.

* It measures how much information is given by ith feature about data.

* We select the feature which has maximum information gain.

Entropy

Entropy is nothing but the uncertainity in a dataset.

or measure of randomness.

Eg:- Group of friends → 10

Avengers
6 votes
5 votes
10 votes

Mario
4 votes
5 votes
9

→ Purity

Entropy measure impurity of data.

Entropy = 0 → pure split

Break → 10.08
→ 10.15 Pm

Entropy

$-p \log p - P \log P$

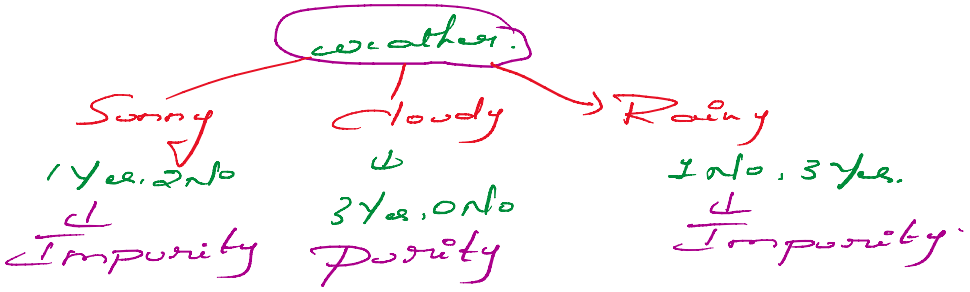
Entropy

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

P_+ → probability of +ve class

P_- → probability of -ve class.

* Entropy measure impurity of node



Imp

* Entropy will be 0 for pure split.
* Entropy will be high when impurity is high.

Information Gain

$$\text{Gain} : H(S) - \sum_{v=1}^{n} \frac{|S_v|}{|S|} H(S_v)$$

S → Sample before split

S_v → Sample after split

$H(S)$ → entropy before split

$H(S_v)$ → Entropy after split

Steps

1. Calculate Entropy of data, $H(S)$
2. Find entropy of each category in a Column
3. Find information gain for that Column.
4. Repeat 2 and 3 for all input variables.
5. Select one Column which has maximum Information Gain.
6. Repeat steps till we get decision tree.

① Find $H(S)$

$$P(\text{Yes}) = \frac{9}{14}$$

$$P(\text{No}) = \frac{5}{14}$$

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$H(S) = -P(\text{Yes}) \log_2 P(\text{Yes}) - P(\text{No}) \log_2 P(\text{No})$$

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

9Yes 5No

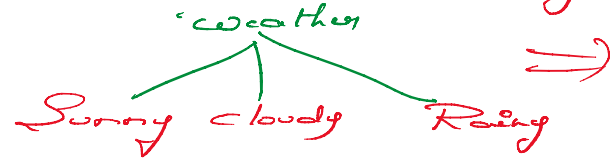
$H(S) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

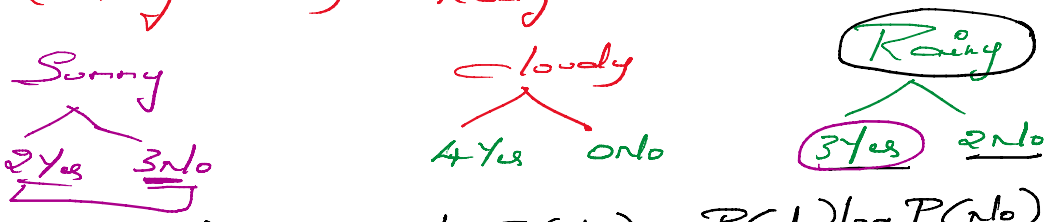
$H(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$

$H(S) = 0.92$

Weather	Humidity	Wind	play
sunny	high	weak	no
sunny	high	strong	no
cloudy	high	weak	yes
rainy	high	weak	yes
rainy	normal	weak	yes
rainy	normal	strong	no
cloudy	normal	strong	yes
sunny	high	weak	no
sunny	normal	weak	yes
rainy	normal	weak	yes
sunny	normal	strong	yes
cloudy	high	strong	yes
cloudy	normal	weak	yes
rainy	high	strong	no

② Find Entropy of categories in weather





$H(Sunny) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

$H(Sunny) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.97$

$H(Rainy) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

$H(Rainy) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.97$

Sunny
2Yes 3No

Rainy
3Yes 2No

$H(Cloudy) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

$H(Cloudy) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) - 0\log_2 0 = 0$

Cloudy
4Yes 0No

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

$= H(S) - \left(\frac{5}{5}\right)H(sunny) - \left(\frac{5}{5}\right)H(rainy) - \left(\frac{4}{5}\right)H(cloudy)$

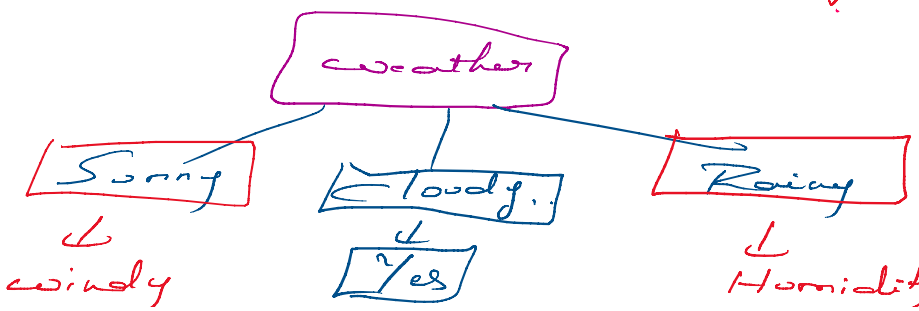
$= 0.92 - \left(\frac{5}{14}\right)(0.97) - \left(\frac{5}{14}\right)(0.97) - \left(\frac{4}{14}\right)(0)$

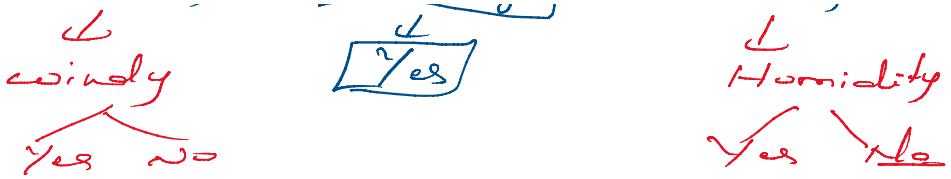
$Gain(weather) = 0.247$

$Gain(Windy) = 0.12$

$Gain(Humidity) = 0.20$

weather is root node which is giving maximum gain.




Windy
Yes
No
Yes
Humidity
Yes
No