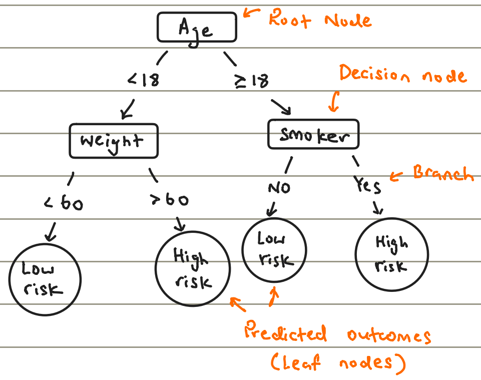
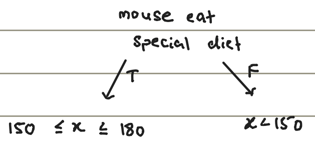
Decision Tree

- can be used for classification & Regression

Decision Tree Classifier



Decision Tree Regressor



Classification Trees

- can use both continuous & discrete features
- numeric threshold can be different for same data

eg: ↙T F↘
Exercise <20 min Exercise <30

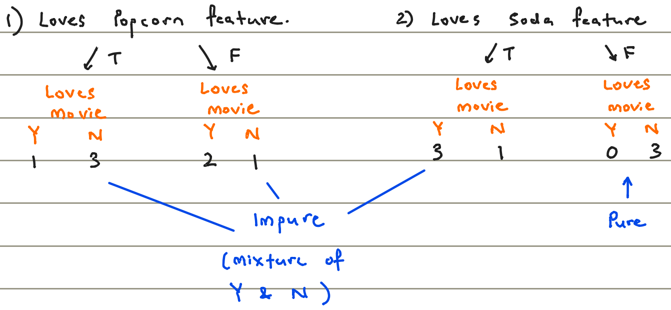
- Prediction nodes can be repeated
- Default : Left branch = True , right branch = False

Building classification Trees

1) Determine the feature at root node

↳ Steps : Build simple tree for each feature

Loves Popcorn	Loves Soda	Age	Loves movie
Y	Y	7	N
Y	N	12	N
N	Y	18	Y
N	Y	35	Y
Y	Y	38	Y
Y	N	50	N
N	N	83	N



Quantifying Impurity of leaf nodes

- Gini Impurity (popular)
- Entropy
- Information Gain

Gini Impurity (categorical)

$$G = 1 - (\text{Probability Yes})^2 - (\text{Probability No})^2$$

$$G\left(\begin{matrix}\text{Loves}\\ \text{popcorn}\\ \text{(T)}\end{matrix}\right) = 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$$= 0.375$$

$$G\left(\begin{matrix}\text{Loves}\\ \text{popcorn}\end{matrix}\ F\right) = 1 - \left(\frac{2}{1+2}\right)^2 - \left(\frac{1}{1+2}\right)^2$$

$$= 0.444$$

Weighted Gini Impurity

- To account for different sample size

$$G\left(\text{Loves popcorn}\right) = \frac{4}{4+3}(0.375) + \frac{3}{4+3}(0.444)$$
$$= 0.405$$

Gini Impurity (numerical)

1) sort numerical col ascending
2) calculate avg of adjacent
3) calculate gini Impurity for each avg

Age	Loves movie
7	N
9.5	
12	N
15	
18	Y
26.5	
35	Y

Age < 9.5

T: Loves movie T	T: Loves movie F	F: Loves movie T	F: Loves movie F
0	1	2	1

$$G\left(\underset{(T)}{\text{Age} < 9.5}\right) = 1 - (0)^2 - \left(\frac{1}{1}\right)^2$$
$$= 0$$

↑ No Impurity

$$G(\text{Age} < 9.5\ (F)) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$$
$$= \frac{4}{9}$$

Weighted Avg (Age < 9.5) $= \frac{1}{4}(0) + \frac{3}{4}\left(\frac{4}{9}\right)$
$= \frac{1}{3}$

* Repeat for all other avg ages.

eg: $\boxed{\text{Age} < 9.5 : 0.33}$ ← Pick lowest Impurity (if > 2 Ages have same Impurities, can choose either)

Age < 15 : 0.343

Age < 26.5 : 0.476

$G(\text{Love Popcorn}) = 0.405$

$G(\text{Love soda}) = 0.214$ ← Choose this as root node

$G(\text{Age} < 9.5) = 0.33$

2) Perform similar splitting & calculate Gini index

3) Assign output values for each leaf node (threshold)
↳ Typically whichever category that has most votes.

Handling overfit of Trees

1) Pruning
2) Limit tree growth (require minimum No. of data in each leaf node)
↳ Test optimal No. using cross validation