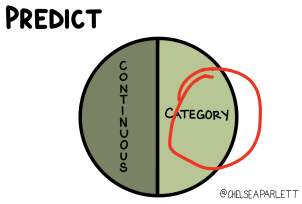
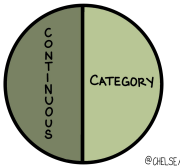


Decision Trees

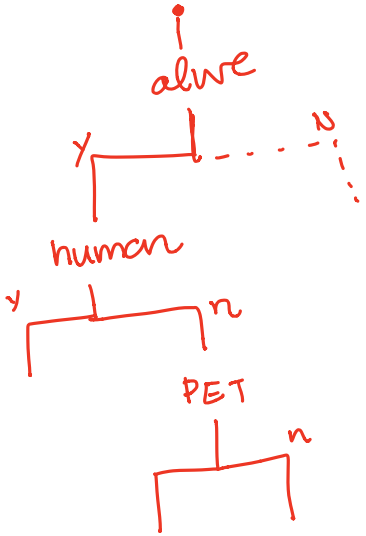
Chelsea Parlett-Pelleriti

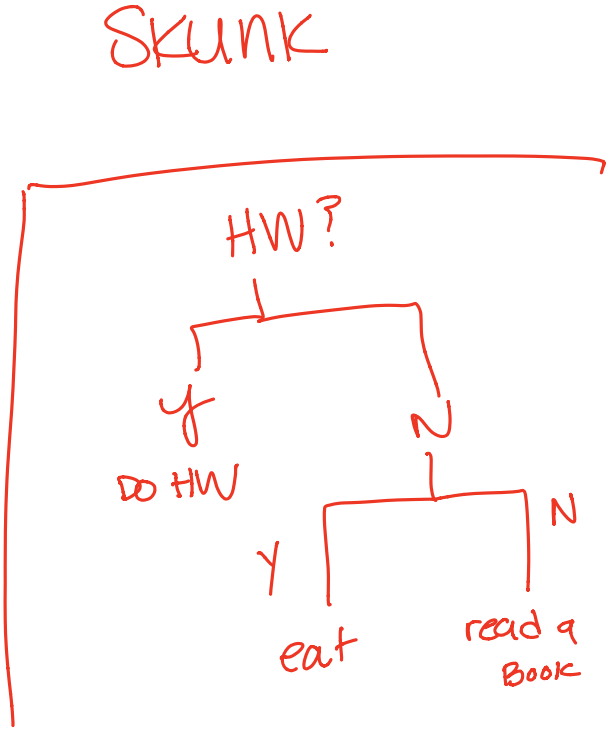

I BE-LEAF
IN YOU
@CHELSEAPARLETT

PREDICT

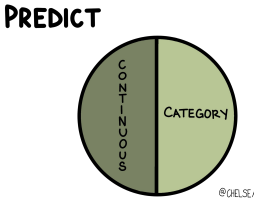

CONTINUOUS
CATEGORY
@CHELSEAPARLETT

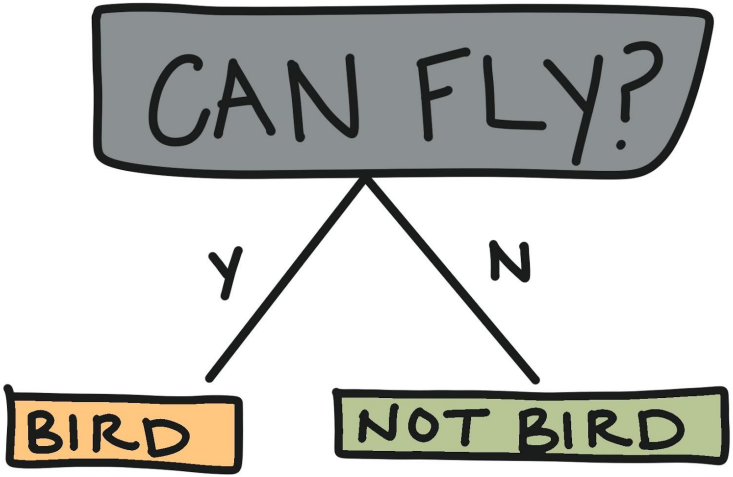
Twenty Questions



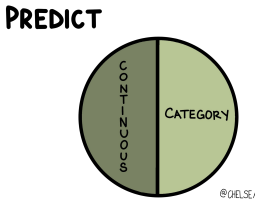


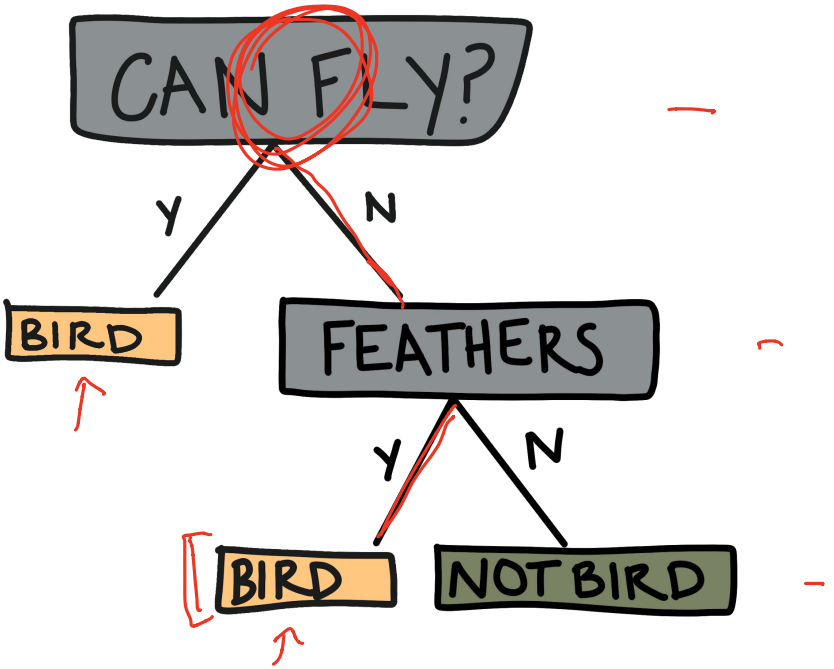
Simple Tree


PREDICT
CONTINUOUS
CATEGORY
@MELSEAPARLETT


CAN FLY?
Y
N
BIRD
NOT BIRD

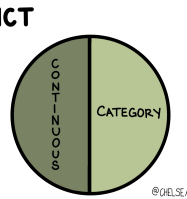
More Complicated Tree

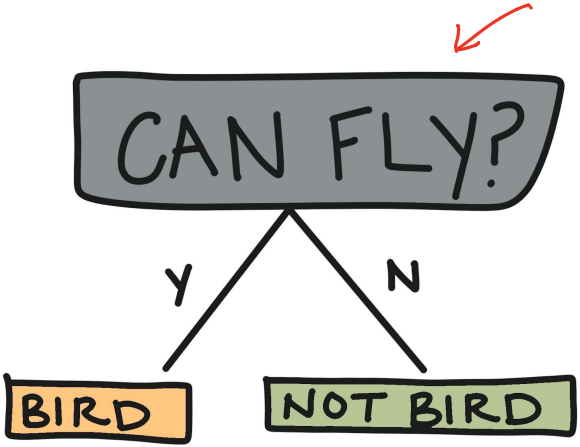


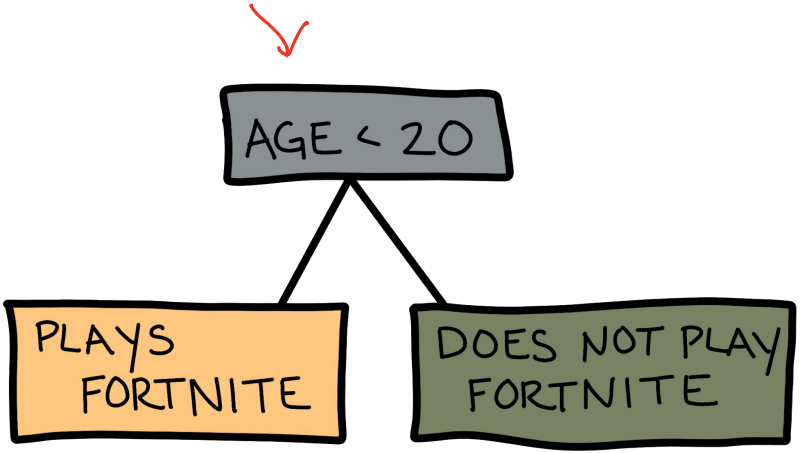


Data Types

PRED ⇝ OUTCOME

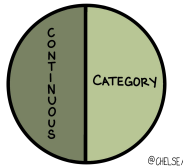


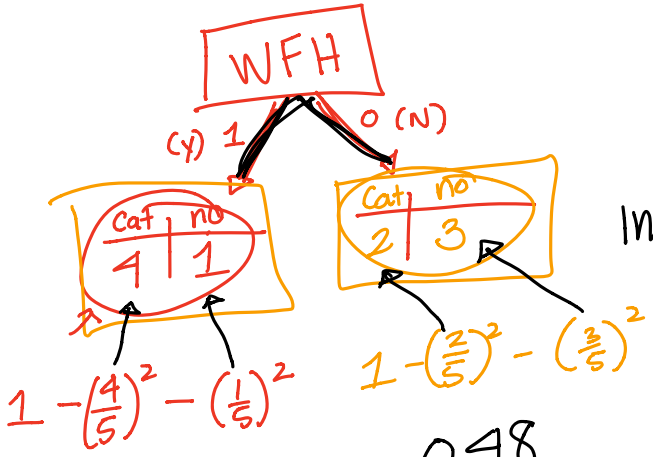




Categorical

$1 - \sum p_i^2$

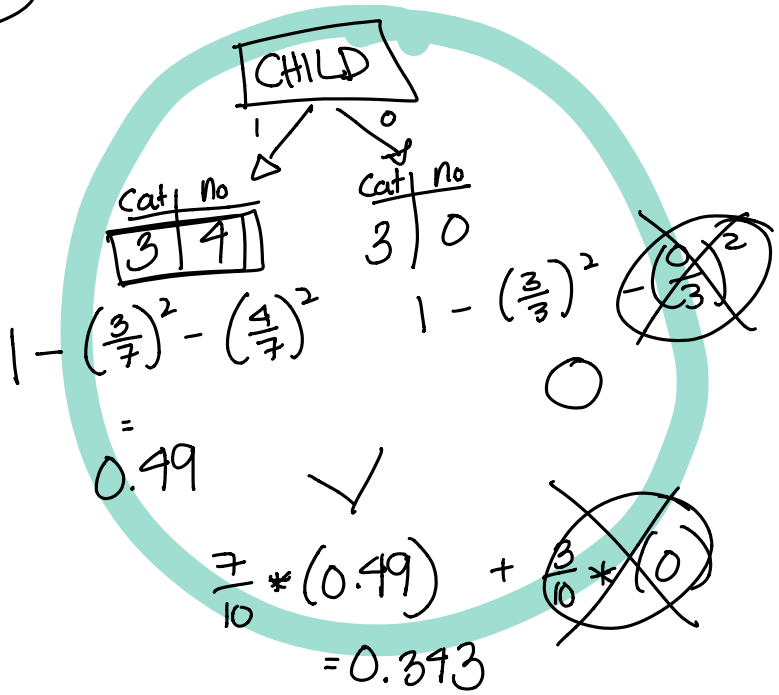




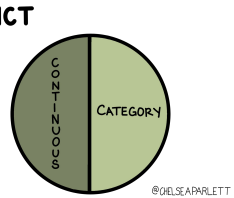
cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	1	71.48068	0	0
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1
1	1	85.56270	1	0
1	1	82.31722	1	0
0	0	75.31565	0	1
1	0	34.01671	1	1
1	1	95.43196	1	1

Categorical

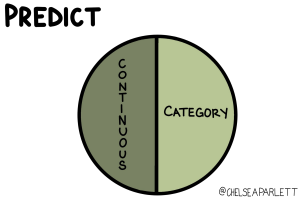
$1 - \sum p_i^2$



cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	1	71.48068	0	0
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1
1	1	85.56270	1	0
1	1	82.31722	1	0
0	0	75.31565	0	1
1	0	34.01671	1	1
1	1	95.43196	1	1

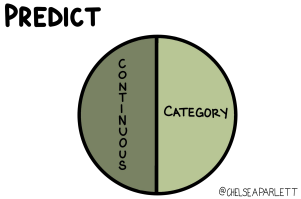


Continuous $1 - \sum p_i^2$



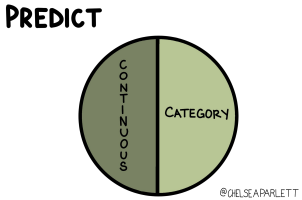
cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1

Continuous $1 - \sum p_i^2$



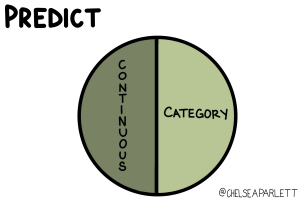
cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1

Basic Steps

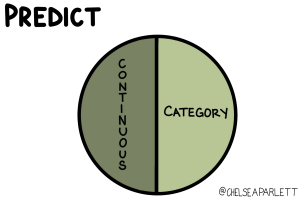


1. Calculate Gini Impurity (or Entropy/Information Gain) for each node
2. Choose Node with lowest score
3. If the parent node has the lowest score, it is a leaf.

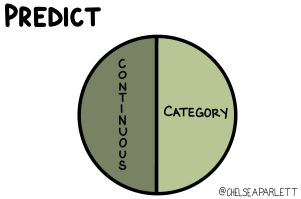
Example



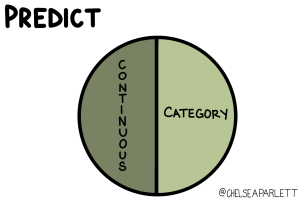
.fit()

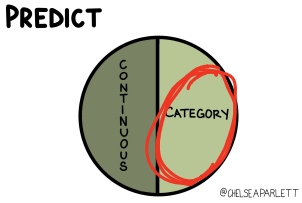


.predict()



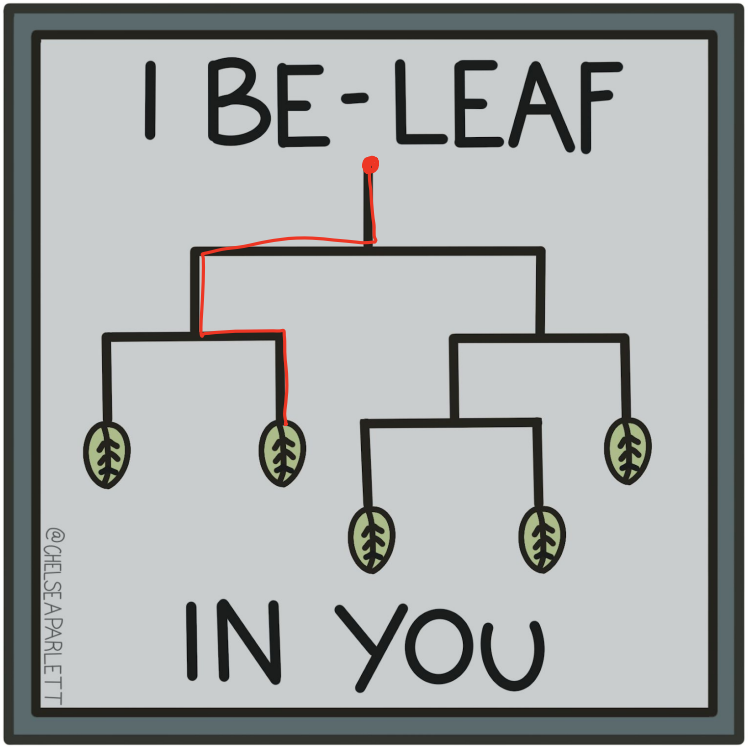
Writing a Gini Impurity Function



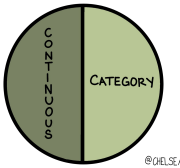


Decision Trees

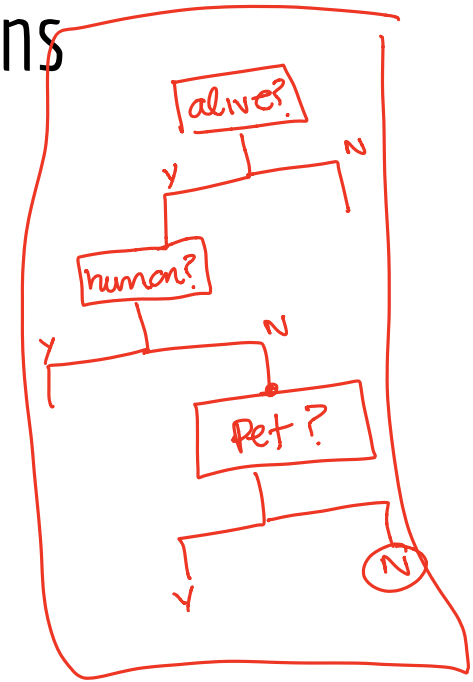
Chelsea Parlett-Pelleriti

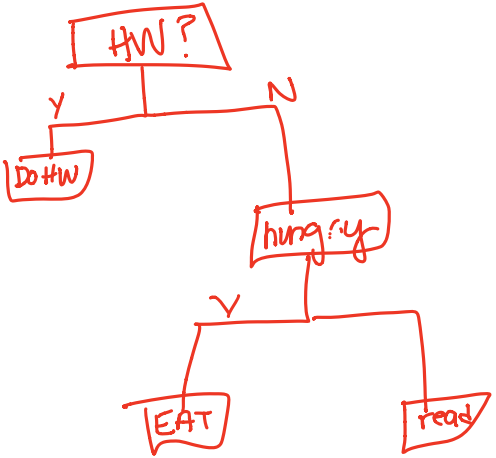

I BE-LEAF
IN YOU
@CHELSEAPARLETT

PREDICT


CONTINUOUS
CATEGORY
@CHELSEAPARLETT

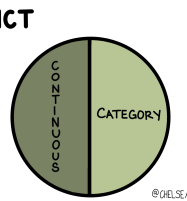
Twenty Questions

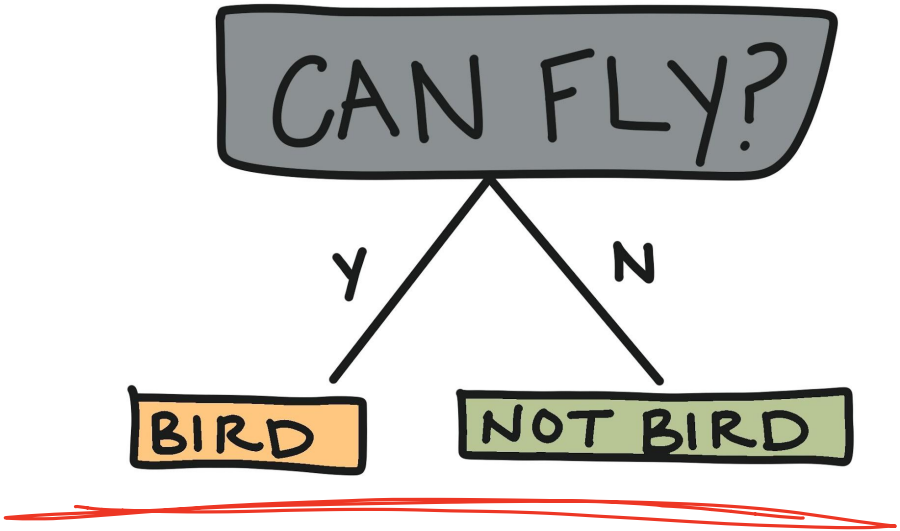




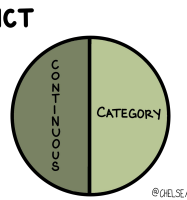
SKUNK

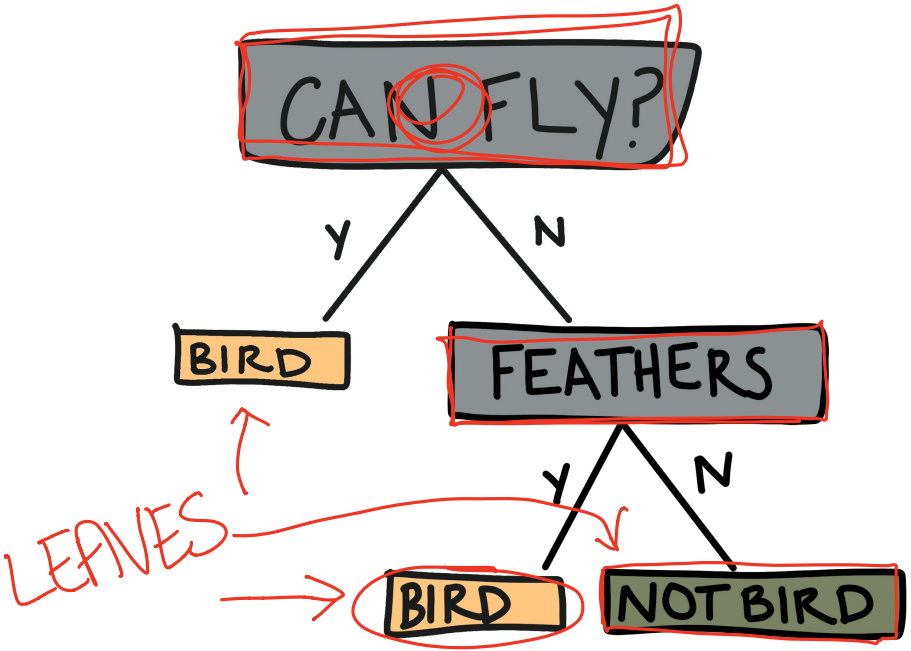
Simple Tree


PREDICT
CONTINUOUS
CATEGORY
@GIELSEAPARLETT


CAN FLY?
Y
N
BIRD
NOT BIRD

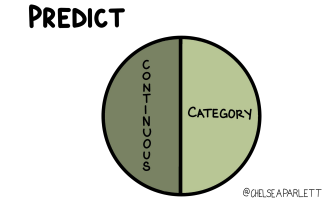
More Complicated Tree

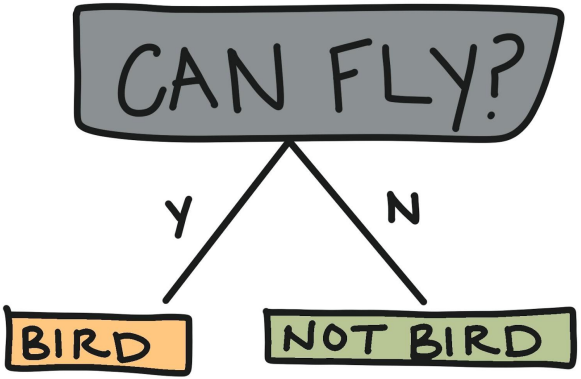


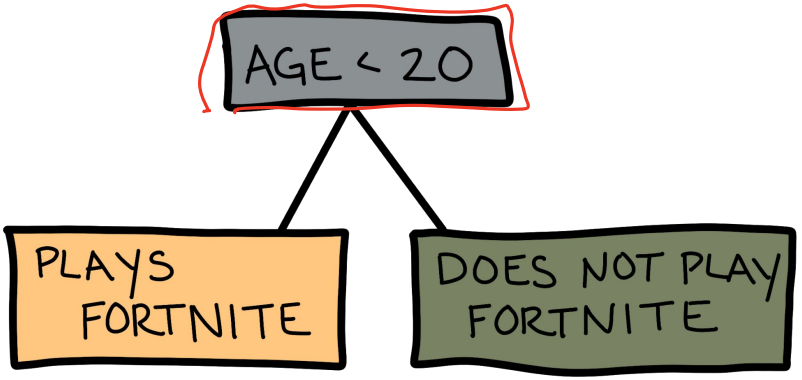


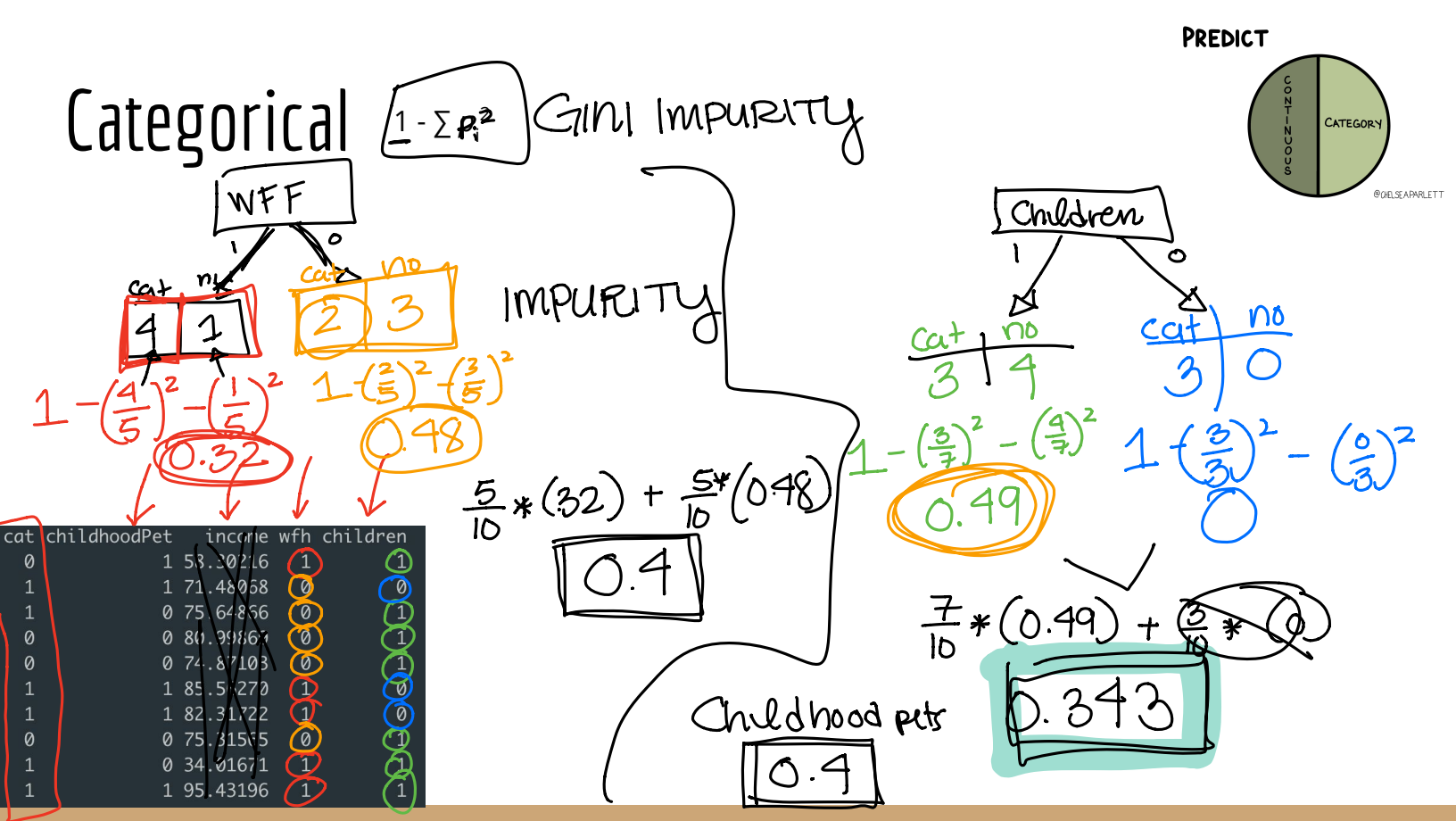
Data Types

PRED ⇝ CATEGORICAL
cat, cont



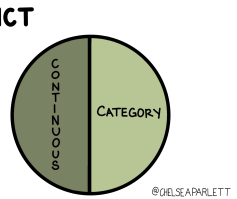





PREDICT
CONTINUOUS
CATEGORY
@GIELSEAPARLETT
Categorical
1 - Σ p_i^2
GINI IMPURITY
WFF
1
0
cat
no
4
1
2
3
1 - (4/5)^2 - (1/5)^2
0.32
1 - (2/5)^2 - (3/5)^2
0.48
IMPURITY
5/10 * (32) + 5/10 * (0.48)
0.4
Children
1
0
cat
no
3
4
cat
no
3
0
1 - (3/7)^2 - (4/7)^2
0.49
1 - (3/3)^2 - (0/3)^2
0
7/10 * (0.49) + 3/10 * (0)
0.343
Childhood pets
0.4
cat childhoodPet income wfh children
0 1 58.30216 1 1
1 1 71.48068 0 0
1 0 75.64856 0 1
0 0 80.99867 0 1
0 0 74.87103 0 1
1 1 85.5[illegible]270 1 0
1 1 82.31722 1 0
0 0 75.31595 0 1
1 0 34.01671 1 1
1 1 95.43196 1 1

Categorical

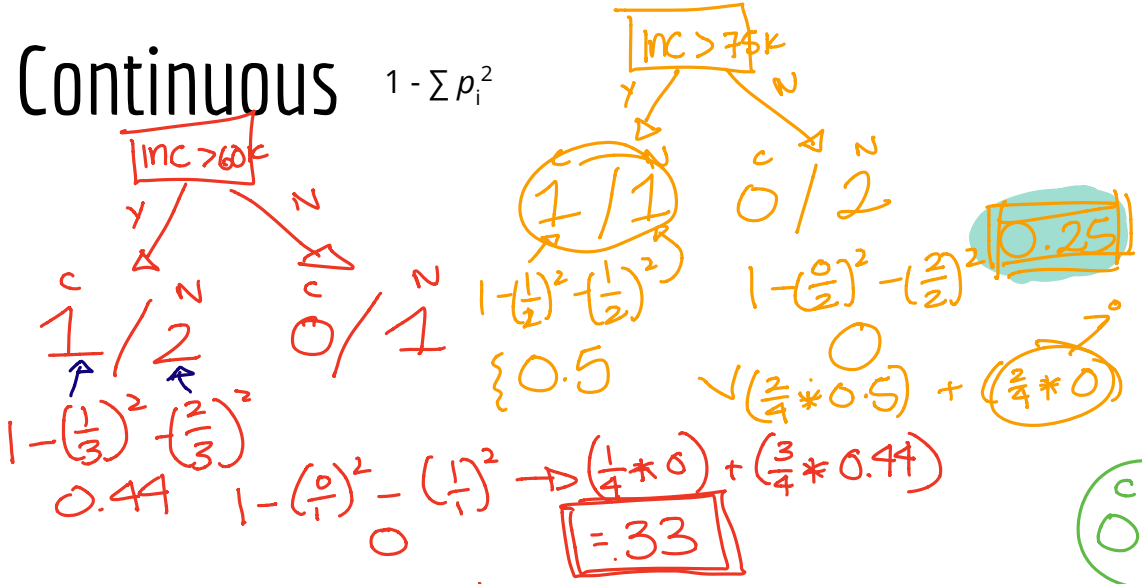
$1 - \sum p_i^2$



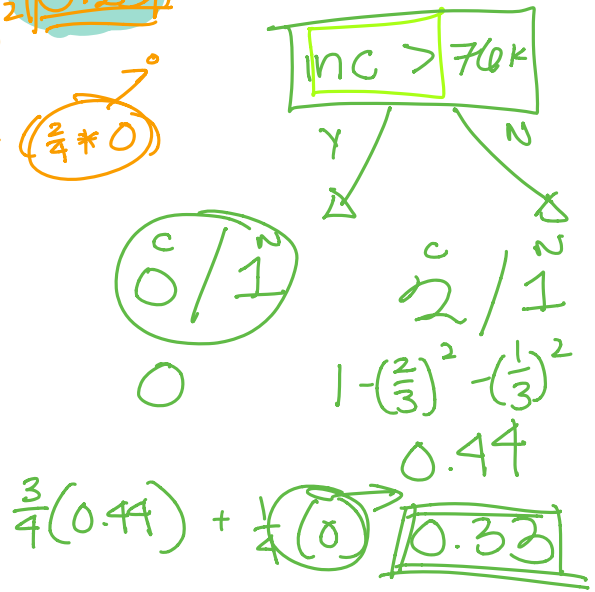
cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	1	71.48068	0	0
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1
1	1	85.56270	1	0
1	1	82.31722	1	0
0	0	75.31565	0	1
1	0	34.01671	1	1
1	1	95.43196	1	1

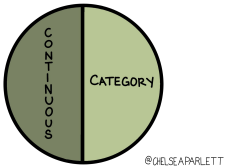
Continuous

$1 - \sum p_i^2$

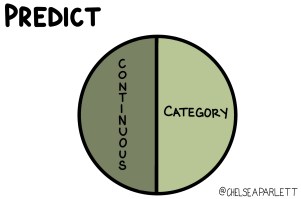


	cat	childhoodPet	income	wfh	children
7	0	1	58.30216	1	1
0	0	0	74.87103	0	1
	1	0	75.64866	0	1
0	0	0	80.99869	0	1



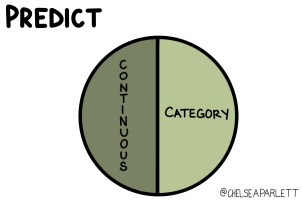


Continuous $1 - \sum p_i^2$



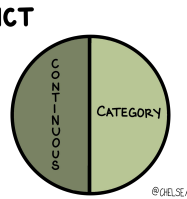
	cat	childhoodPet	income	wfh	children
7	0	1	58.30216	1	1
9	0	0	74.87103	0	1
	1	0	75.64866	0	1
0	0	0	80.99869	0	1

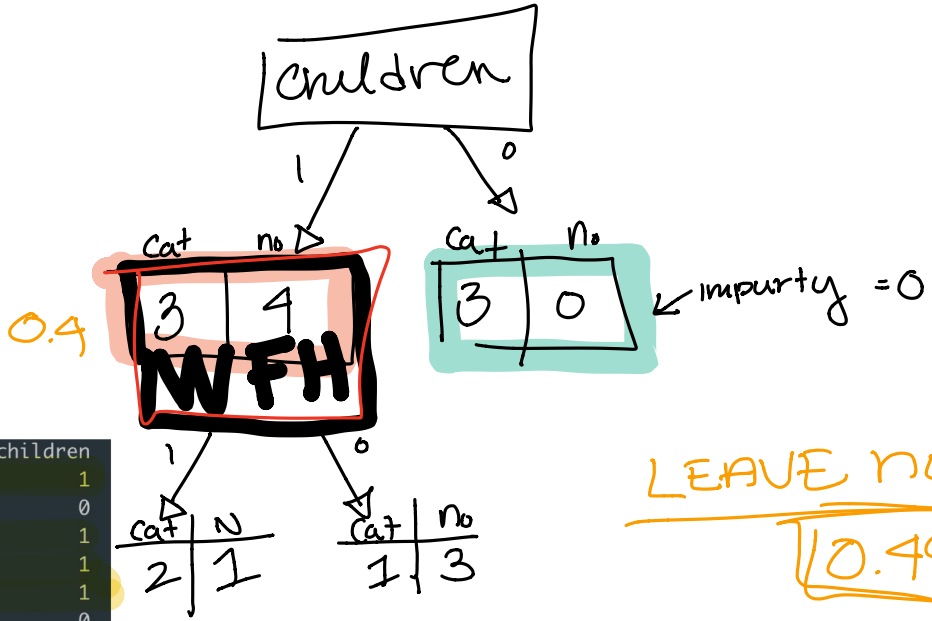
Basic Steps



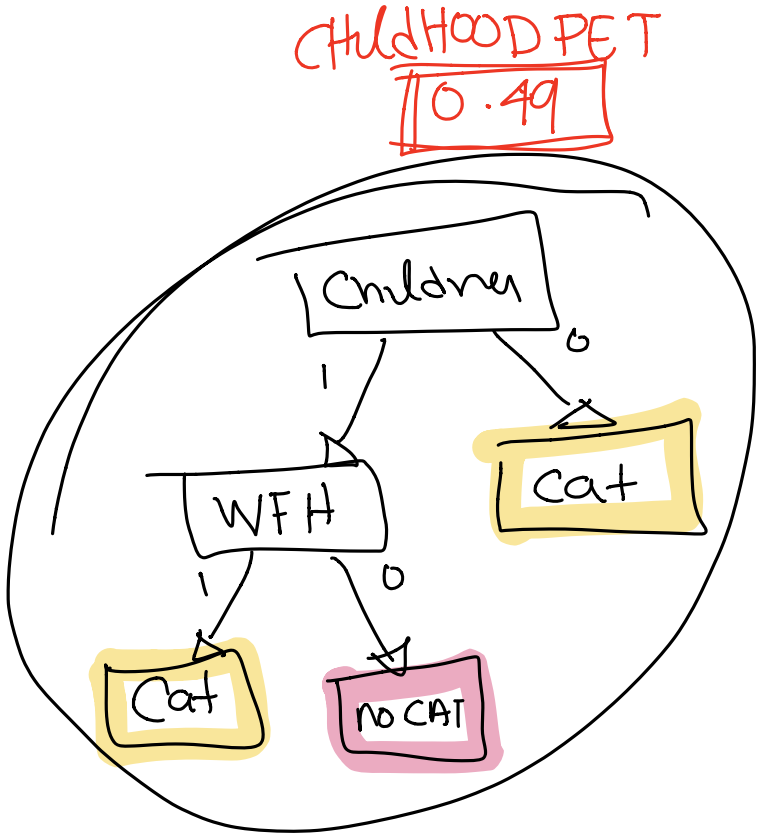
1. Calculate Gini Impurity (or Entropy/Information Gain) for each node
2. Choose Node with lowest score
3. If the parent node has the lowest score, it is a leaf.

Example

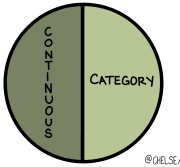




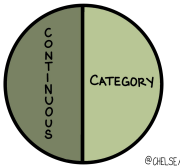
cat	childhoodPet	income	wfh	children
0	1	58.30216	1	1
1	1	71.48068	0	0
1	0	75.64866	0	1
0	0	80.99869	0	1
0	0	74.87103	0	1
1	1	85.56270	1	0
1	1	82.31722	1	0
0	0	75.31565	0	1
1	0	34.01671	1	1
1	1	95.43196	1	1



PREDICT



PREDICT



Writing a Gini Impurity Function