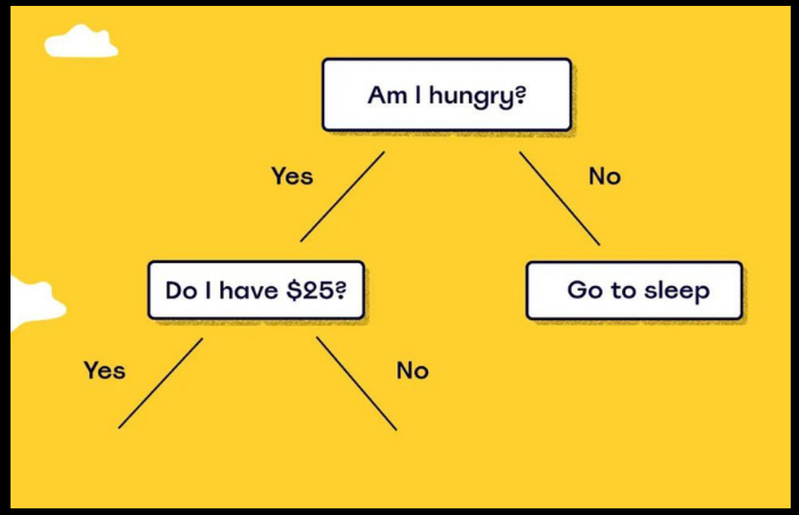
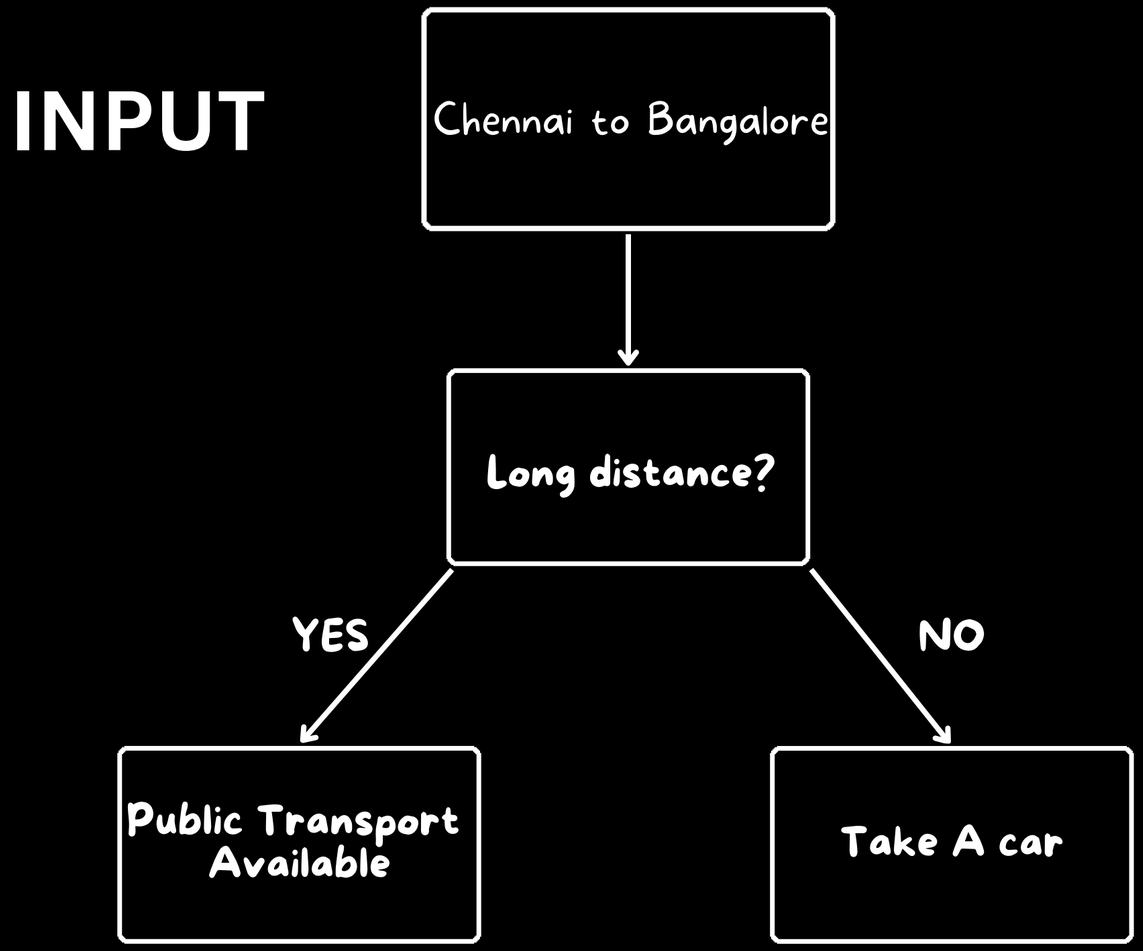
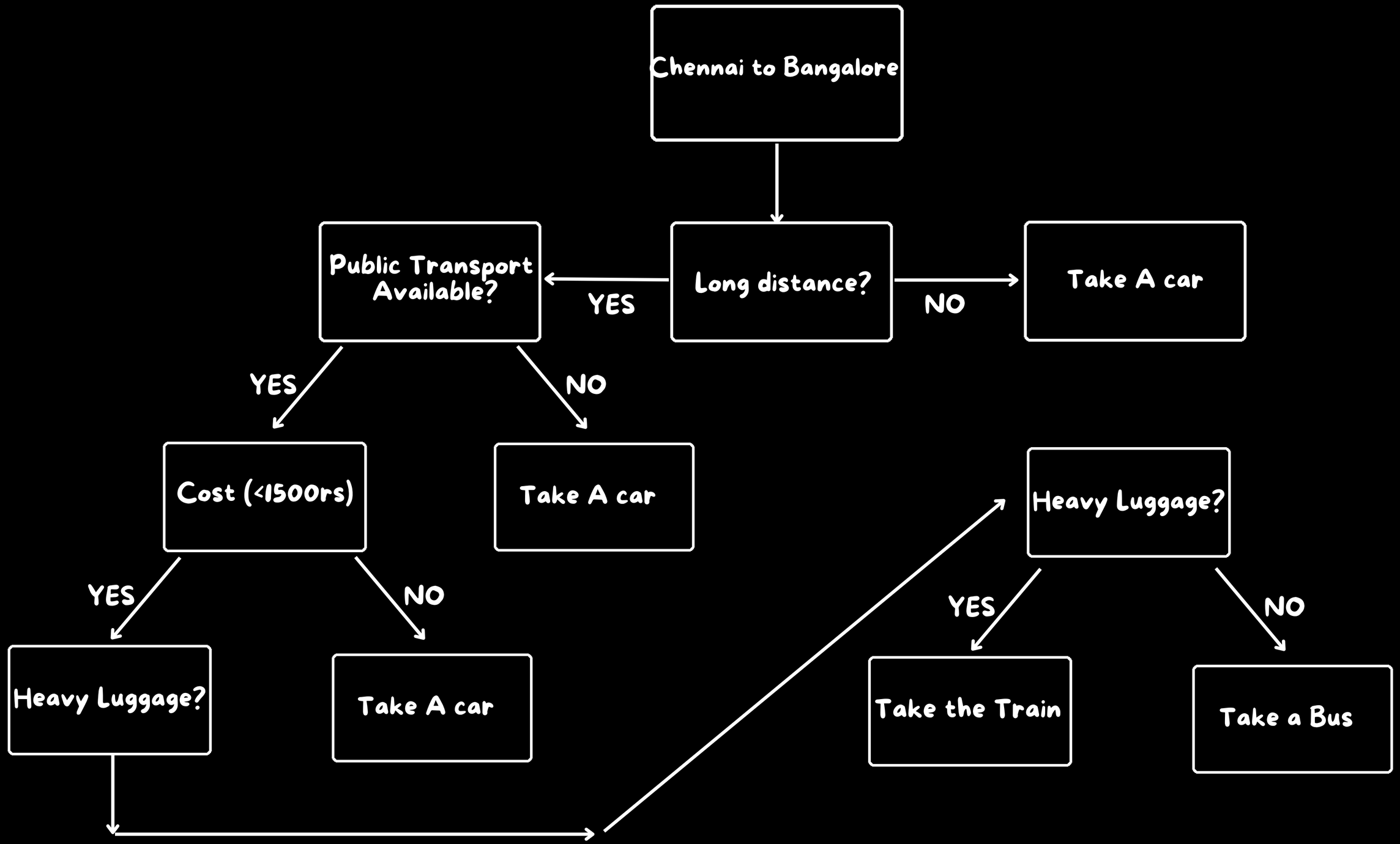
DECISION TREES



Agenda

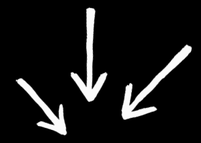
The Problem


INPUT
Chennai to Bangalore
Long distance?
YES
NO
Public Transport Available
Take A car

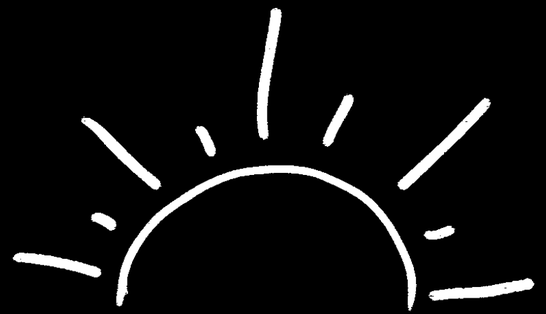

Chennai to Bangalore
Long distance?
YES
Public Transport Available?
NO
Take A car
YES
Cost (<1500rs)
NO
Take A car
YES
Heavy Luggage?
NO
Take A car
Heavy Luggage?
YES
Take the Train
NO
Take a Bus

Why Decision Trees

Is it just a bunch of if-else statements?



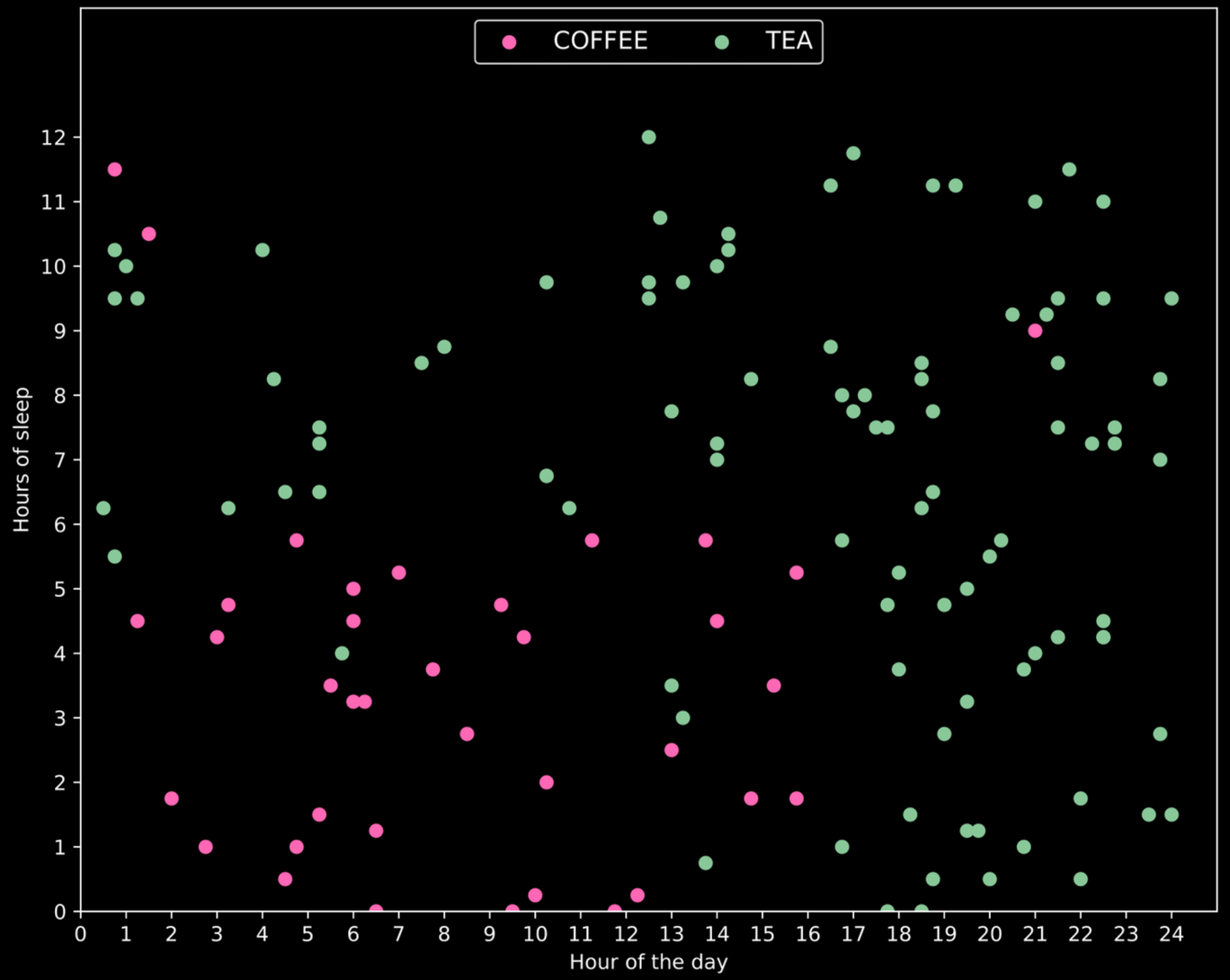
Yes & NO

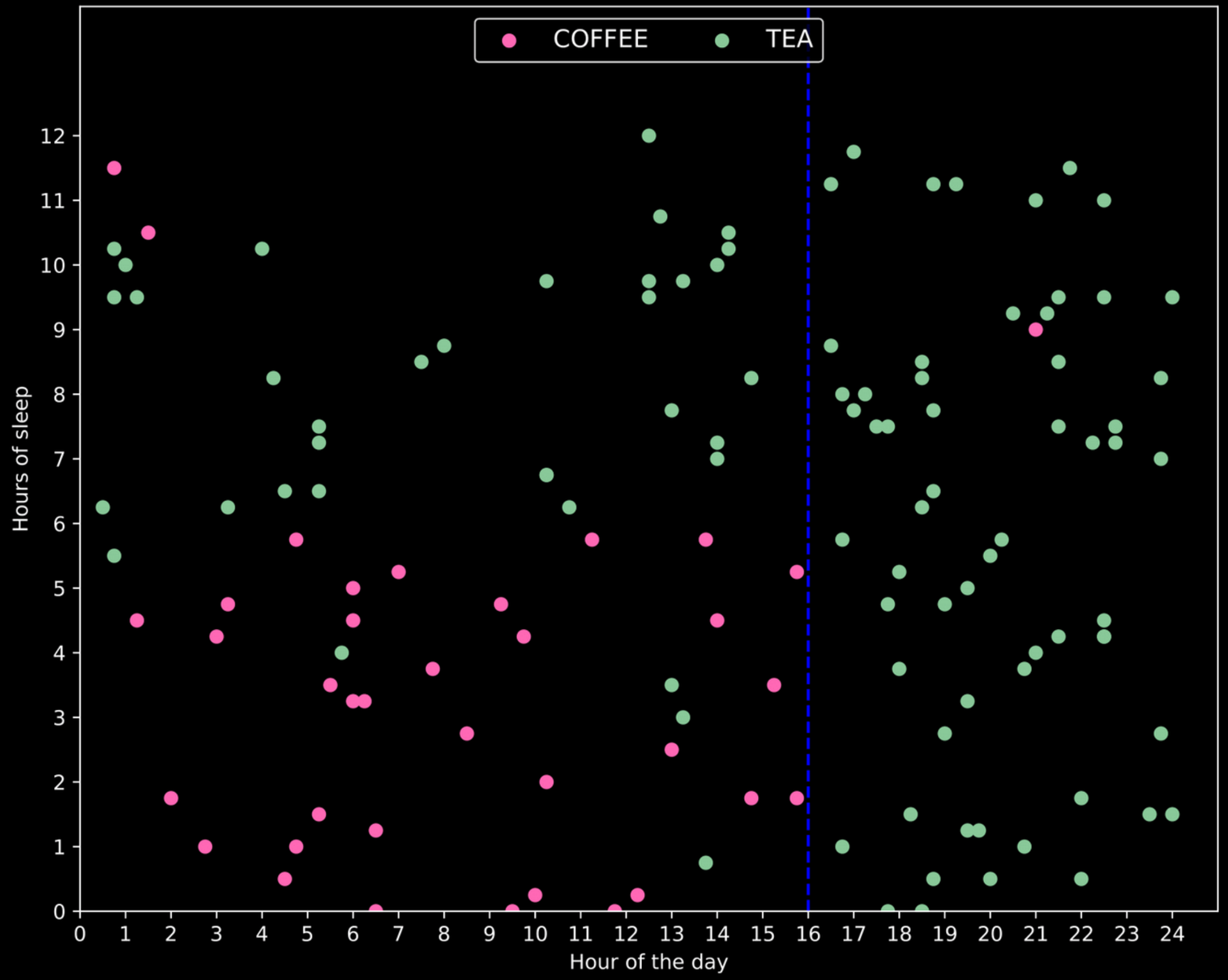


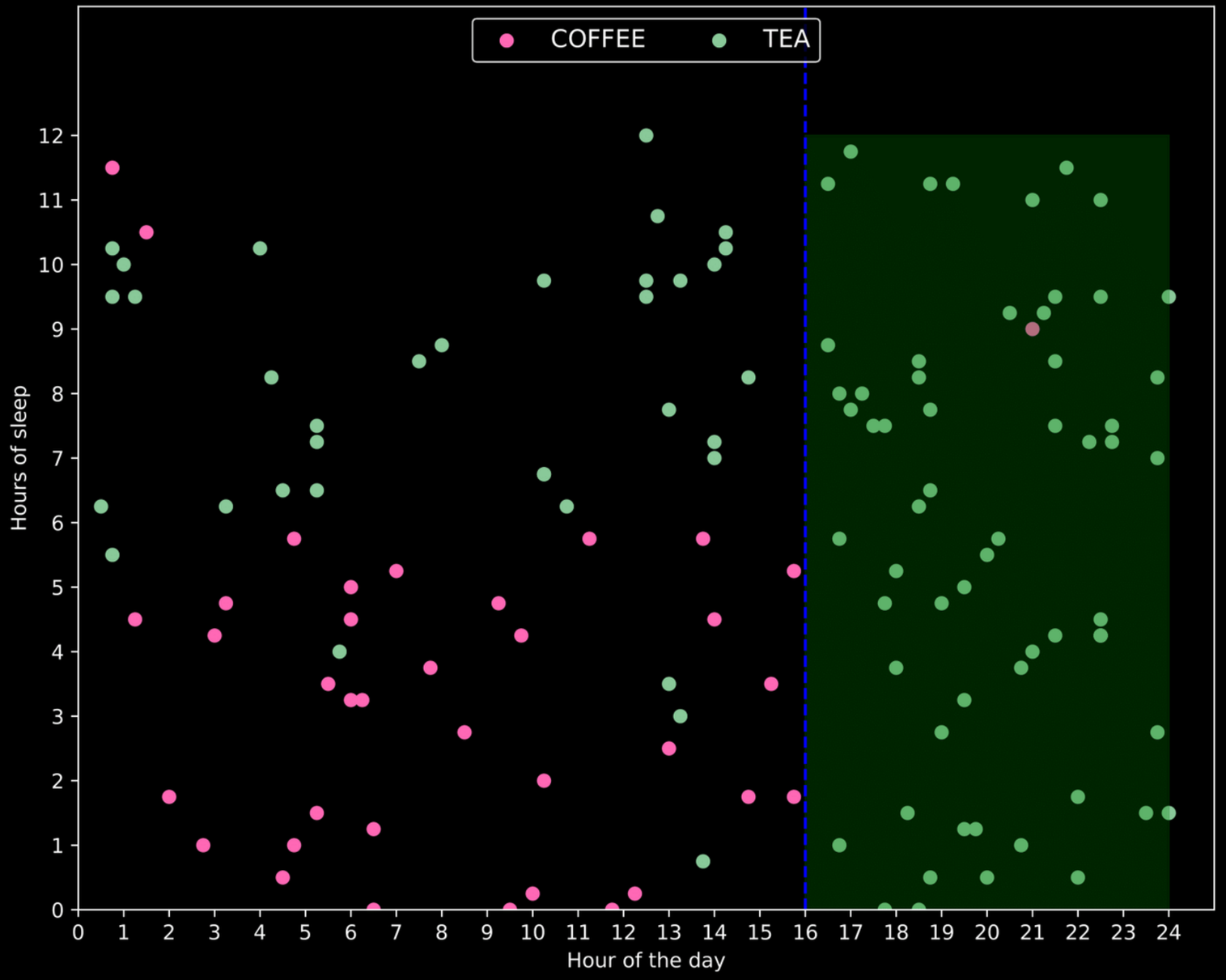
Tea or Coffee

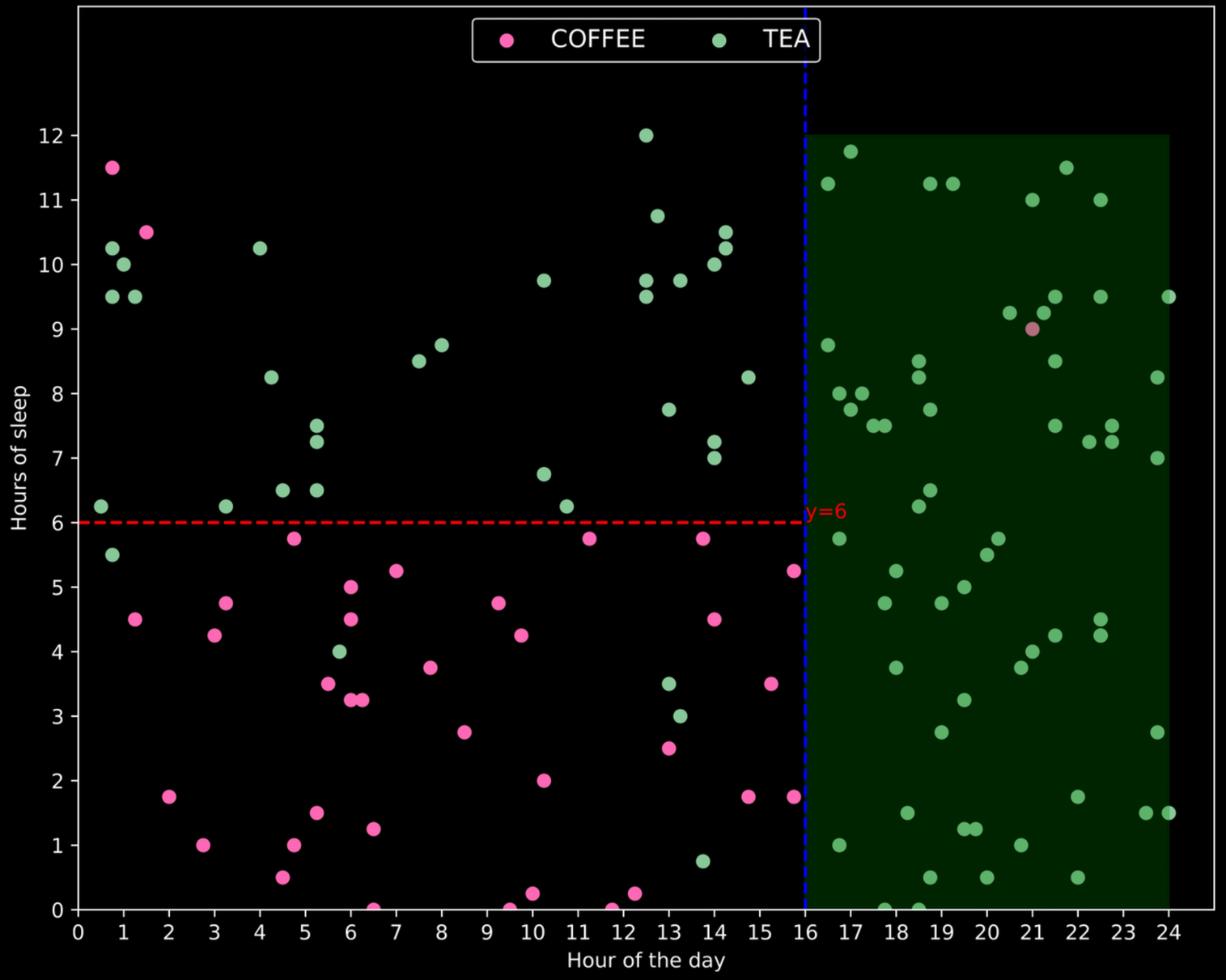
Decision trees help us decide the splitting condition

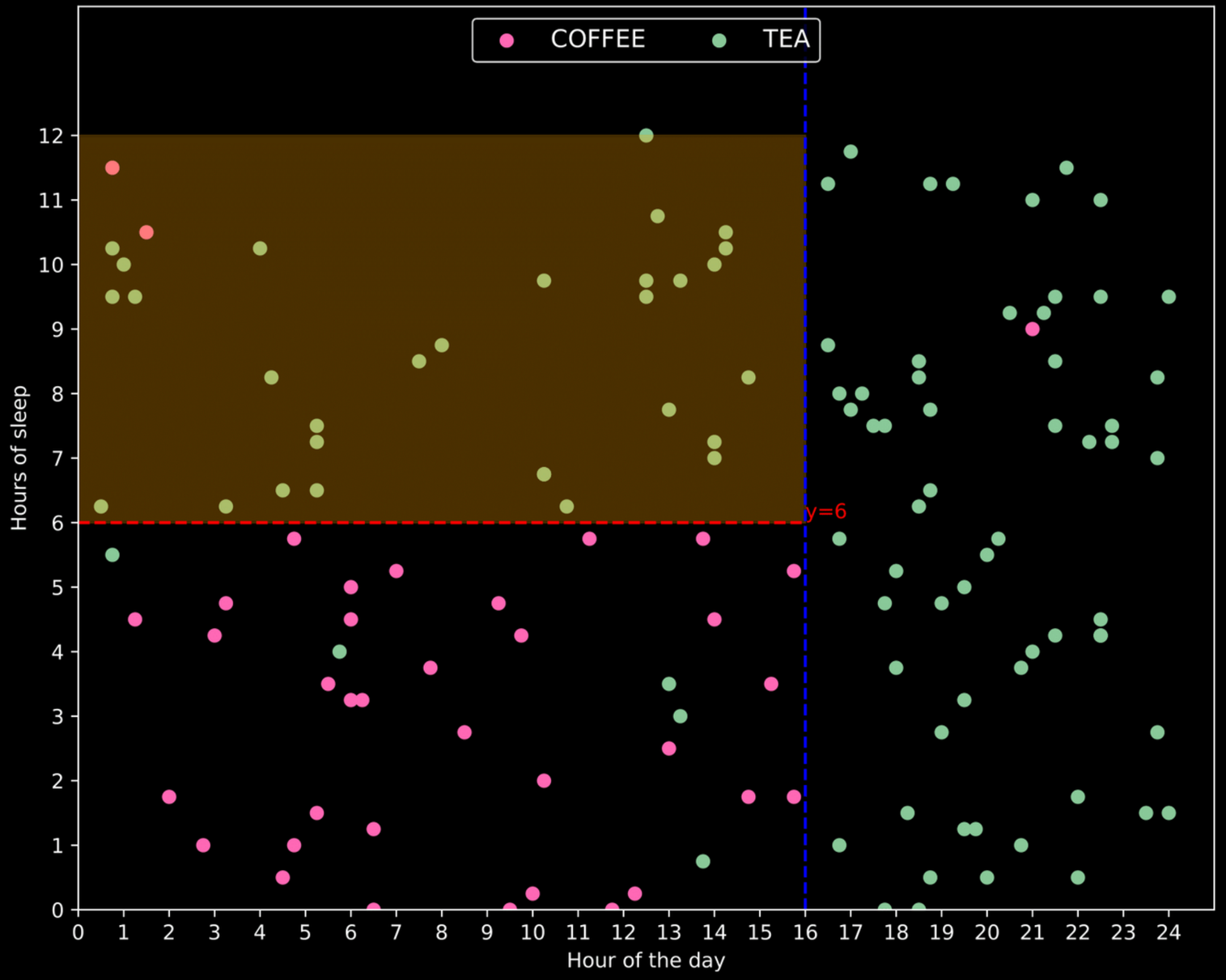
Hours of sleep	Time of the Day	COFFEE/TEA
7	3pm	TEA
9	6pm	TEA
5	10am	COFFEE
6	11am	COFFEE
9	3pm	TEA

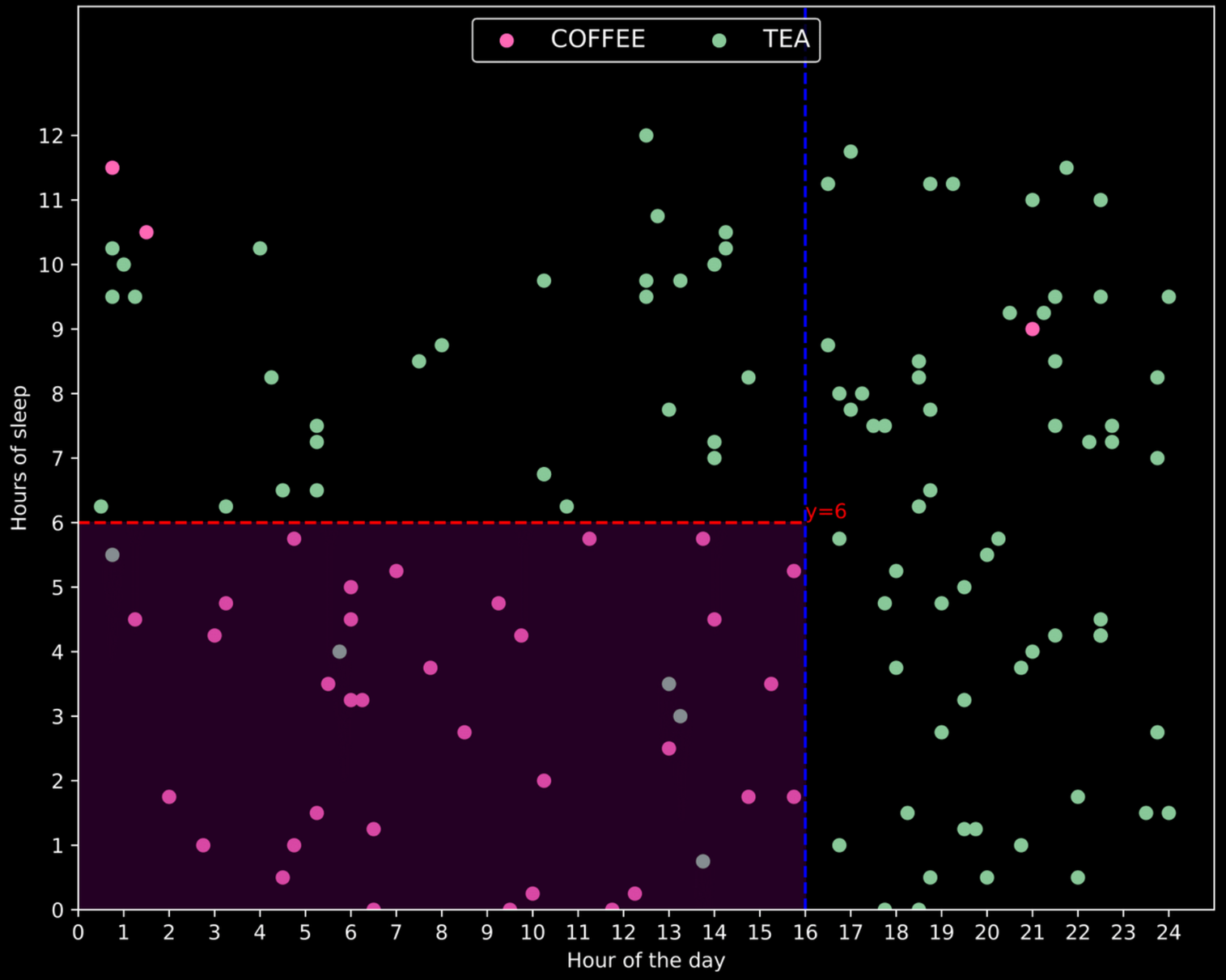


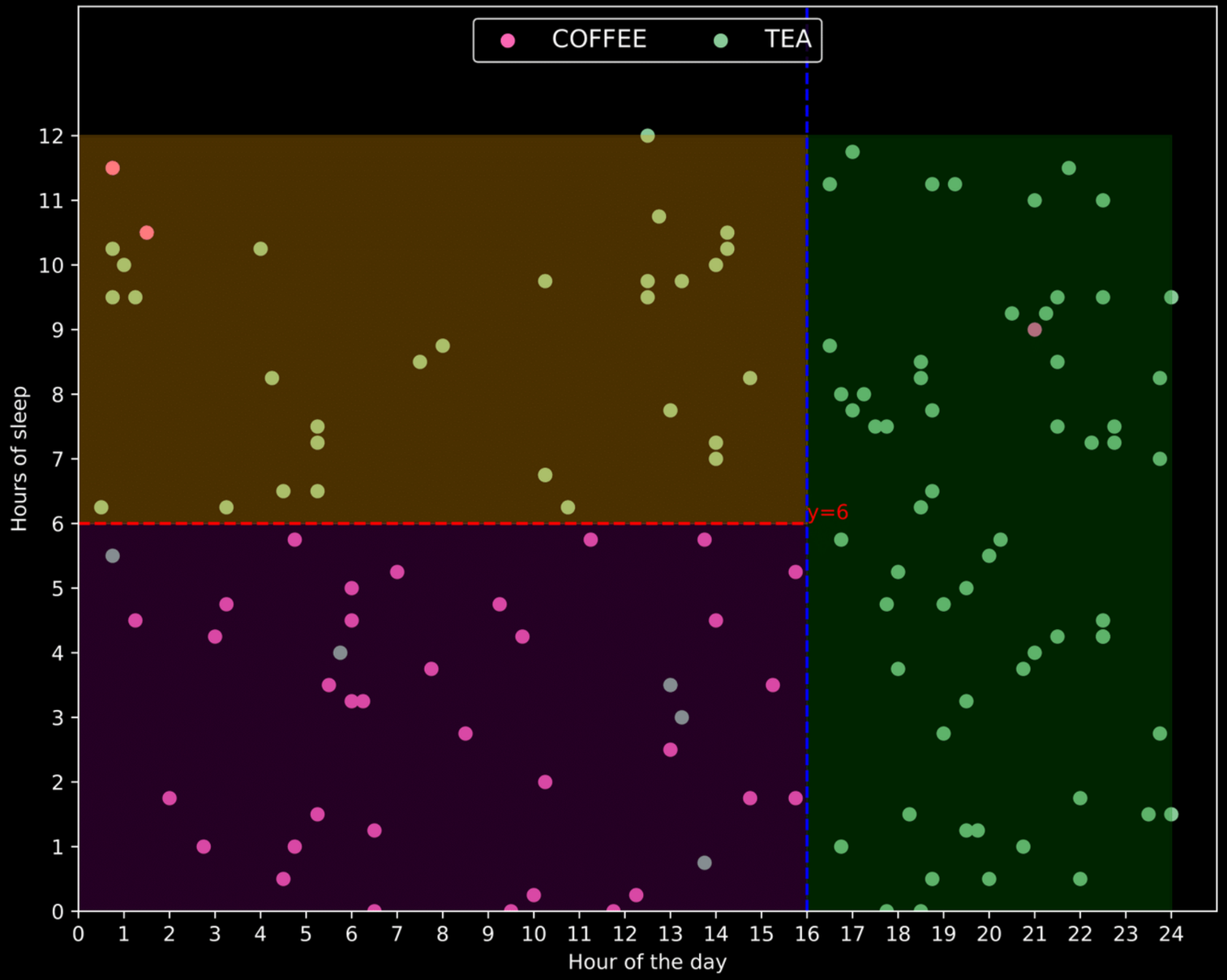

COFFEE
TEA
Hours of sleep
Hour of the day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


COFFEE
TEA
Hours of sleep
Hour of the day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

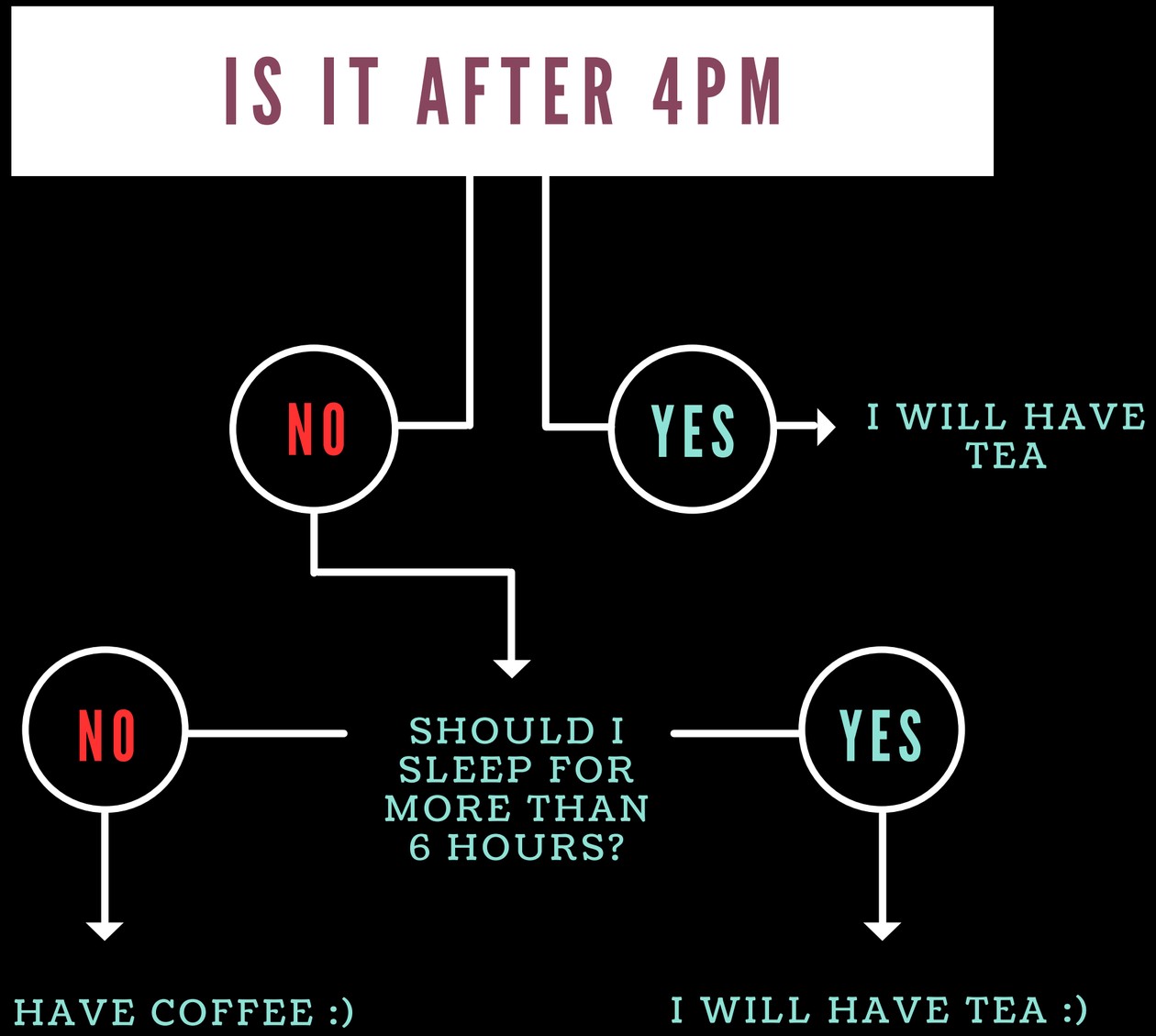

COFFEE
TEA
Hours of sleep
Hour of the day
y=6
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

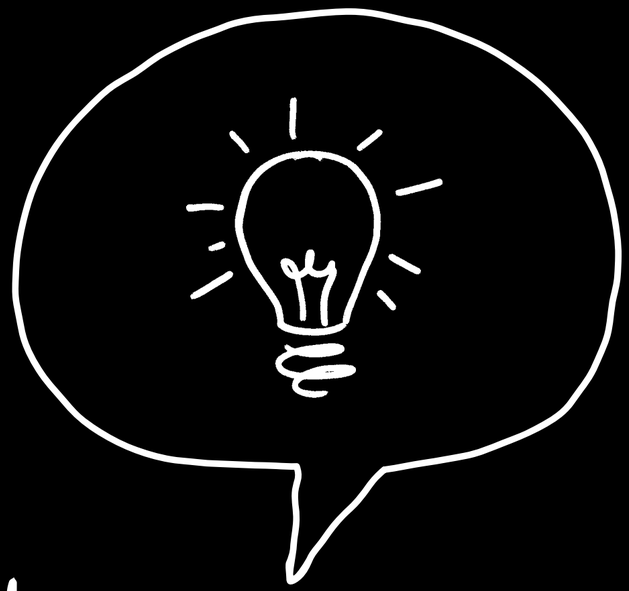

COFFEE
TEA
Hours of sleep
Hour of the day
y=6
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


COFFEE
TEA
y=6
Hours of sleep
Hour of the day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


COFFEE
TEA
Hours of sleep
Hour of the day
y=6
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

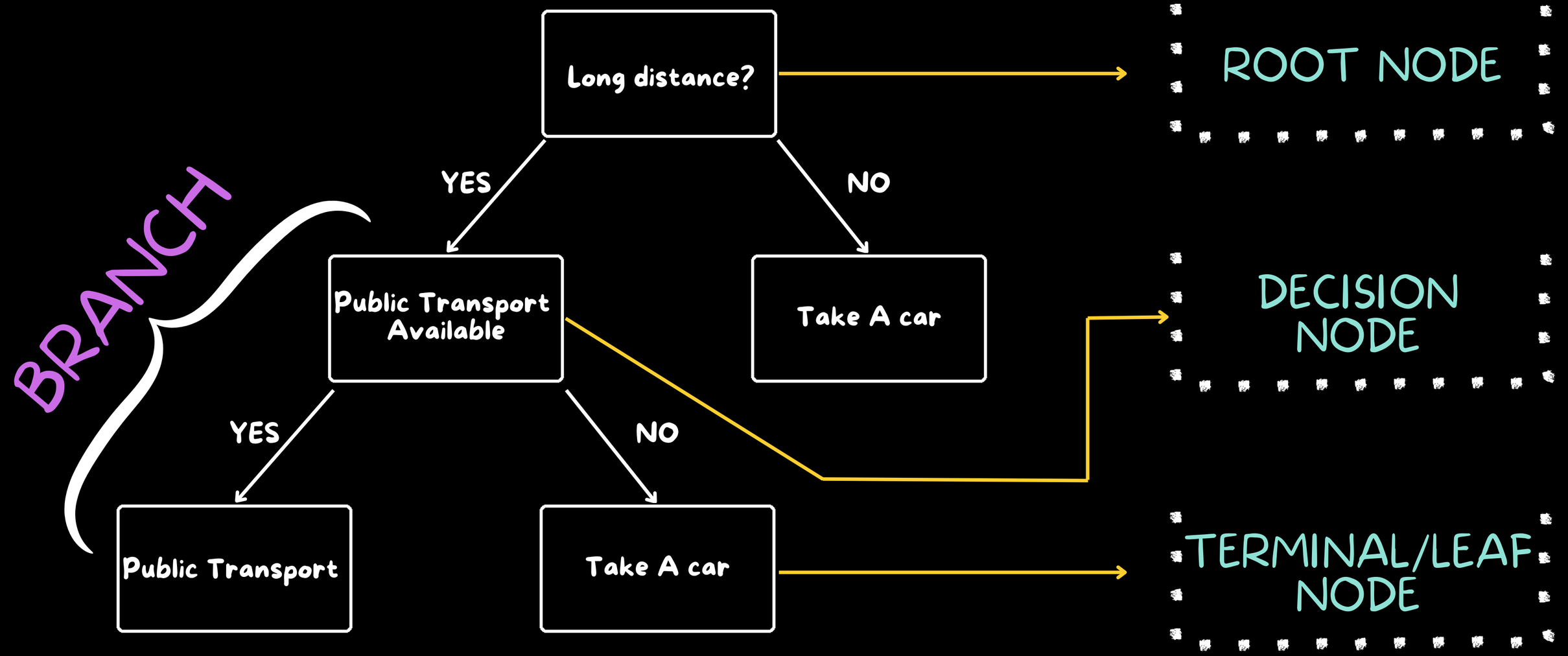
TEA OR COFFEE





How does a computer do all this?

TERMINOLOGIES



FEATURES

we use features to
classify things

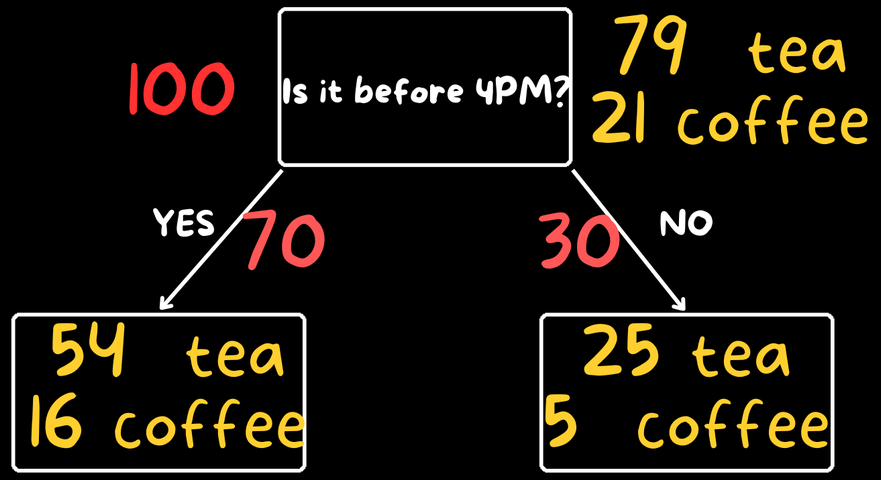
Examples:

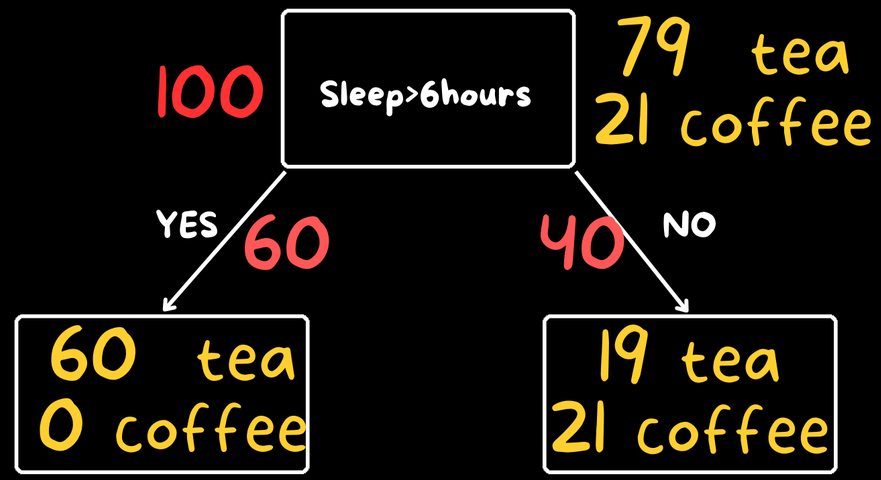
phase for ice and water

hours of sleep and time
of the day
for tea or coffee

ecg, pulse,
MRI scan, CT scan
for heart conditions

Which split is better?



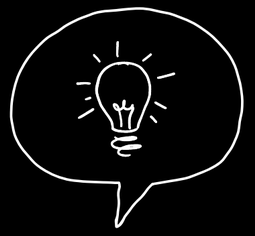


SPLITTING CONDITION

Decide whether to use hours of sleep or time of the day to split

1. GINI IMPURITY
2. ENTROPY & INFORMATION GAIN

1. GINI IMPURITY
2. ENTROPY & INFORMATION GAIN



GINI IMPURITY

60 tea
0 coffee
pure

20 tea
20 coffee
impure

How is it calculated?

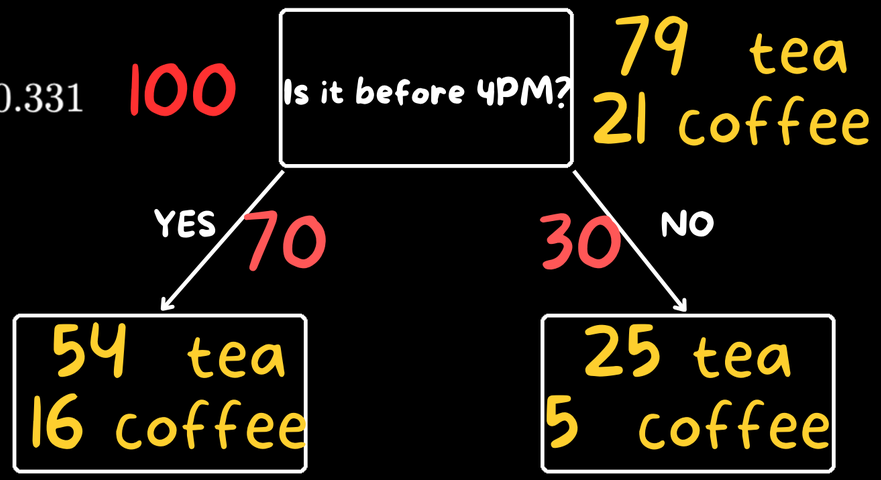
Understand intuitively
Why this works

Calculate for sub-nodes

Gini impurity of a subnode = 1 - (probability of tea)^2 - (probability of coffee)^2

$$1 - \left(\frac{79}{100}\right)^2 - \left(\frac{21}{100}\right)^2 = 0.331$$

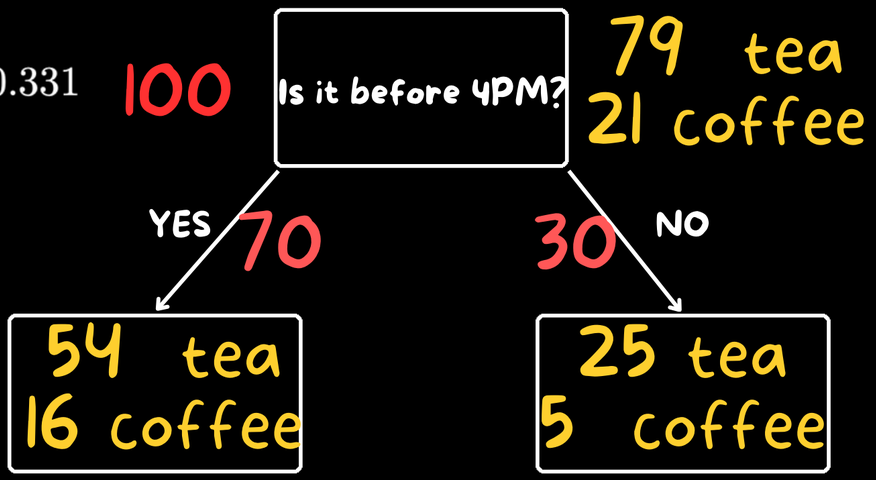
Calculating Gini impurity for a subnode



Calculate for sub-nodes

Gini impurity of a subnode = 1 - (probability of tea)^2 - (probability of coffee)^2

$$1 - \left(\frac{79}{100}\right)^2 - \left(\frac{21}{100}\right)^2 = 0.331$$



$$1 - \left(\frac{54}{70}\right)^2 - \left(\frac{16}{70}\right)^2 = 0.35$$

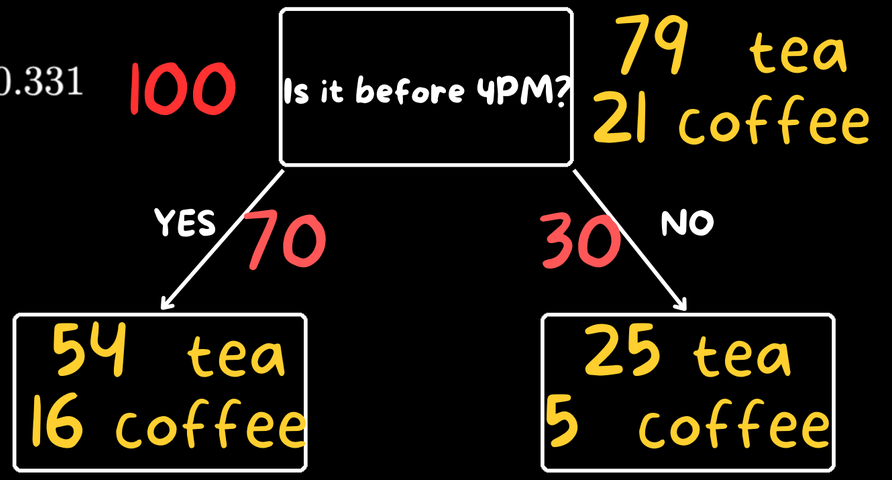
Calculating Gini impurity for a subnode

$$1 - \left(\frac{25}{30}\right)^2 - \left(\frac{5}{30}\right)^2 = 0.277$$

Calculate for split

Gini impurity of a subnode = 1 - (probability of tea)^2 - (probability of coffee)^2

$$1-\left(\frac{79}{100}\right)^2-\left(\frac{21}{100}\right)^2=0.331$$



$$1-\left(\frac{54}{70}\right)^2-\left(\frac{16}{70}\right)^2=0.35$$

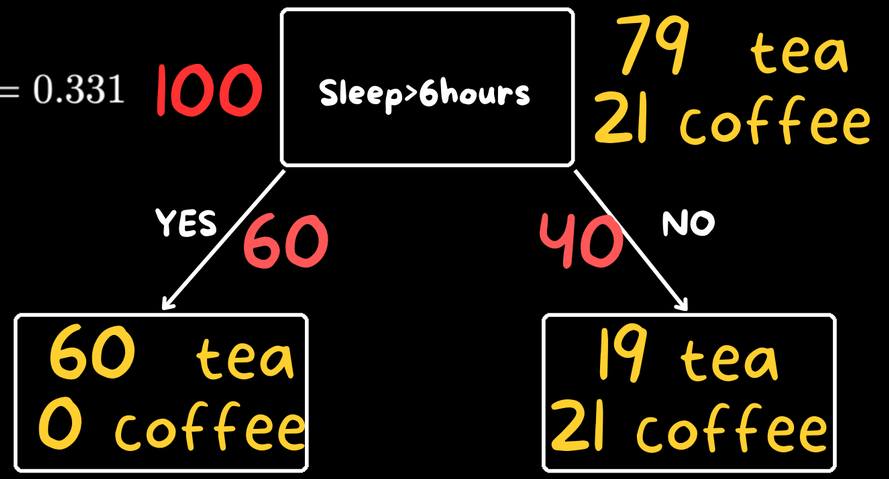
$$1-\left(\frac{25}{30}\right)^2-\left(\frac{5}{30}\right)^2=0.277$$

Gini impurity of a SPLIT = weighted average of the gini impurities of its sub branch

Gini impurity of split

0.35(70/100)+0.277(30/100)=0.328

$$1-\left(\frac{79}{100}\right)^2-\left(\frac{21}{100}\right)^2=0.331$$



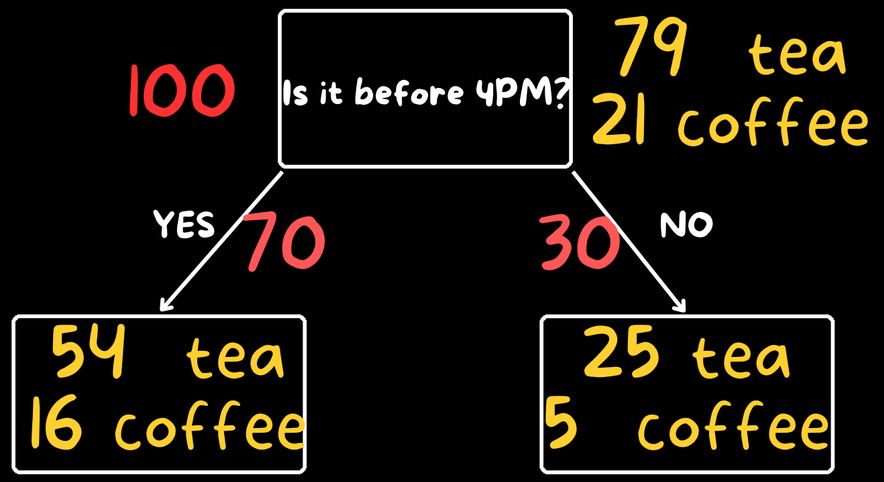
$$1-\left(\frac{60}{60}\right)^2-\left(\frac{0}{60}\right)^2=0$$

$$1-\left(\frac{21}{40}\right)^2-\left(\frac{19}{40}\right)^2=0.498$$

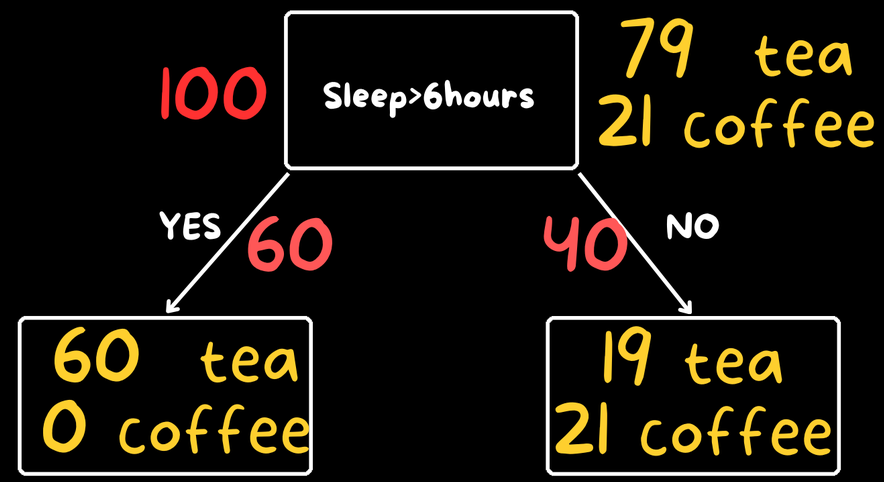
Gini impurity of a SPLIT = weighted average of the gini impurities of its sub branch

Gini impurity of split

0(60/100)+0.498(40/100)=0.199



Gini impurity of split
0.328

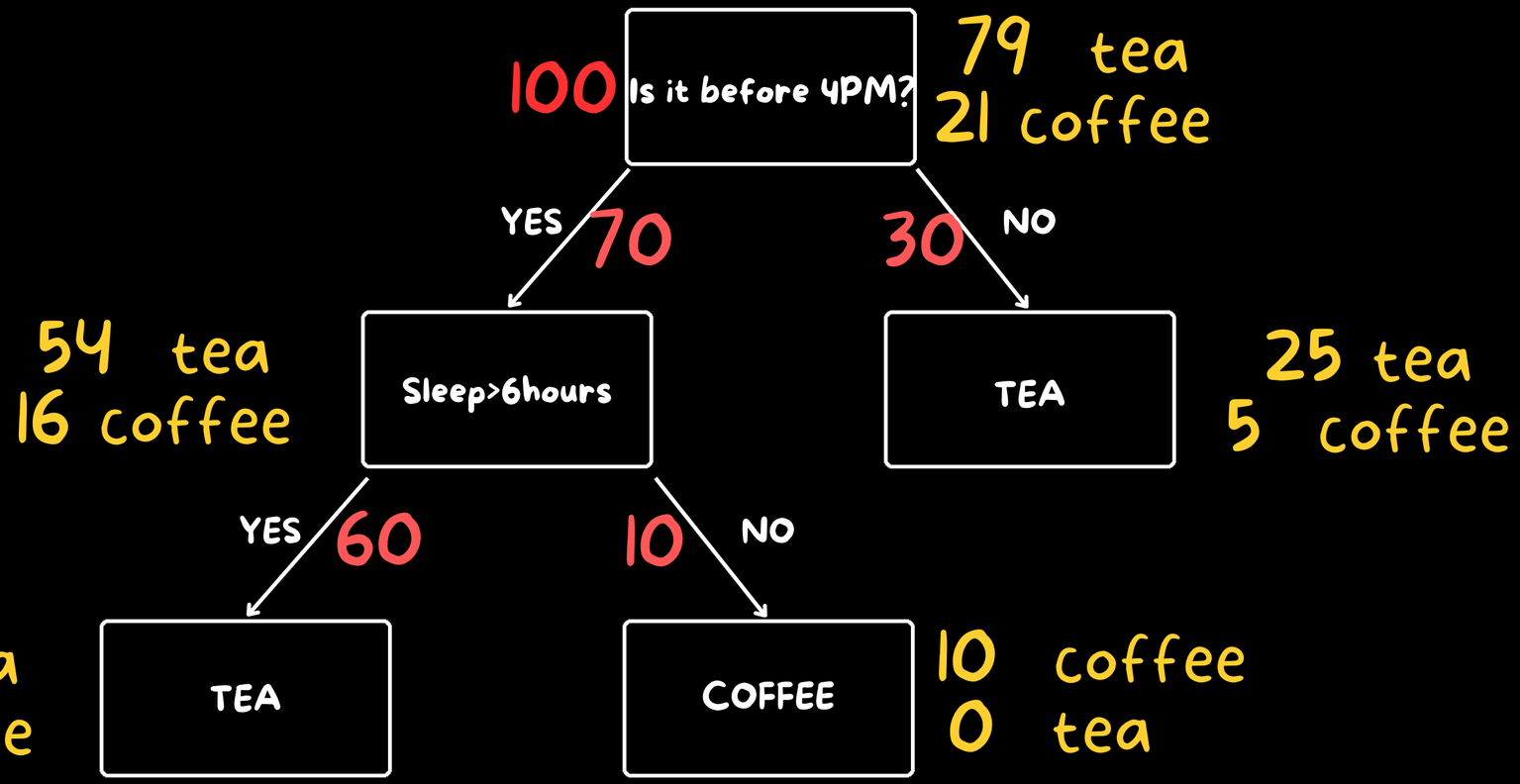


Gini impurity of split
0.199

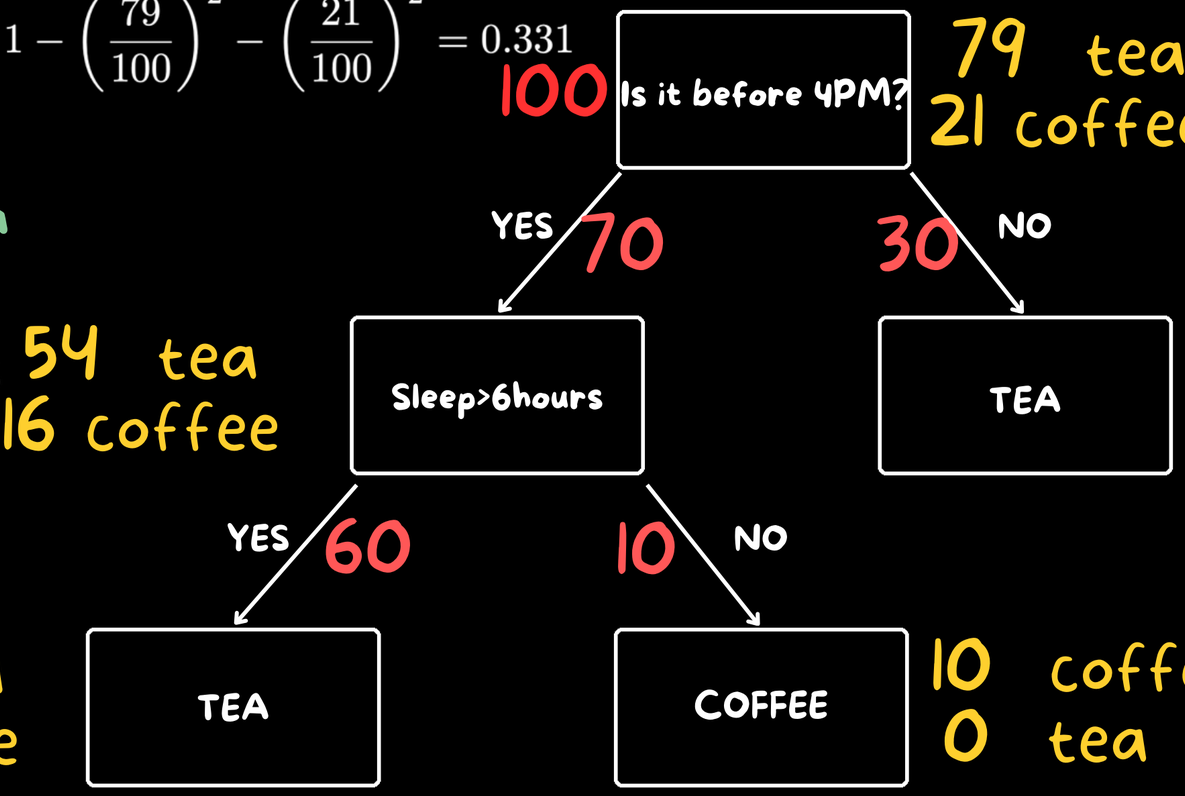
What are the gini impurities of the leaf nodes

Gini impurity of a subnode = 1 - (probability of tea)^2 - (probability of coffee)^2

Scan QR code and lock your answers



Gini impurity of a subnode = 1 - (probability of tea)^2 - (probability of coffee)^2



$$1 - \left(\frac{79}{100}\right)^2 - \left(\frac{21}{100}\right)^2 = 0.331$$

Calculating Gini impurity for a subnode

$$1 - \left(\frac{54}{70}\right)^2 - \left(\frac{16}{70}\right)^2 = 0.35$$

Calculating Gini impurity for a subnode

$$1 - \left(\frac{25}{30}\right)^2 - \left(\frac{5}{30}\right)^2 = 0.277$$

Calculating Gini impurity for a subnode

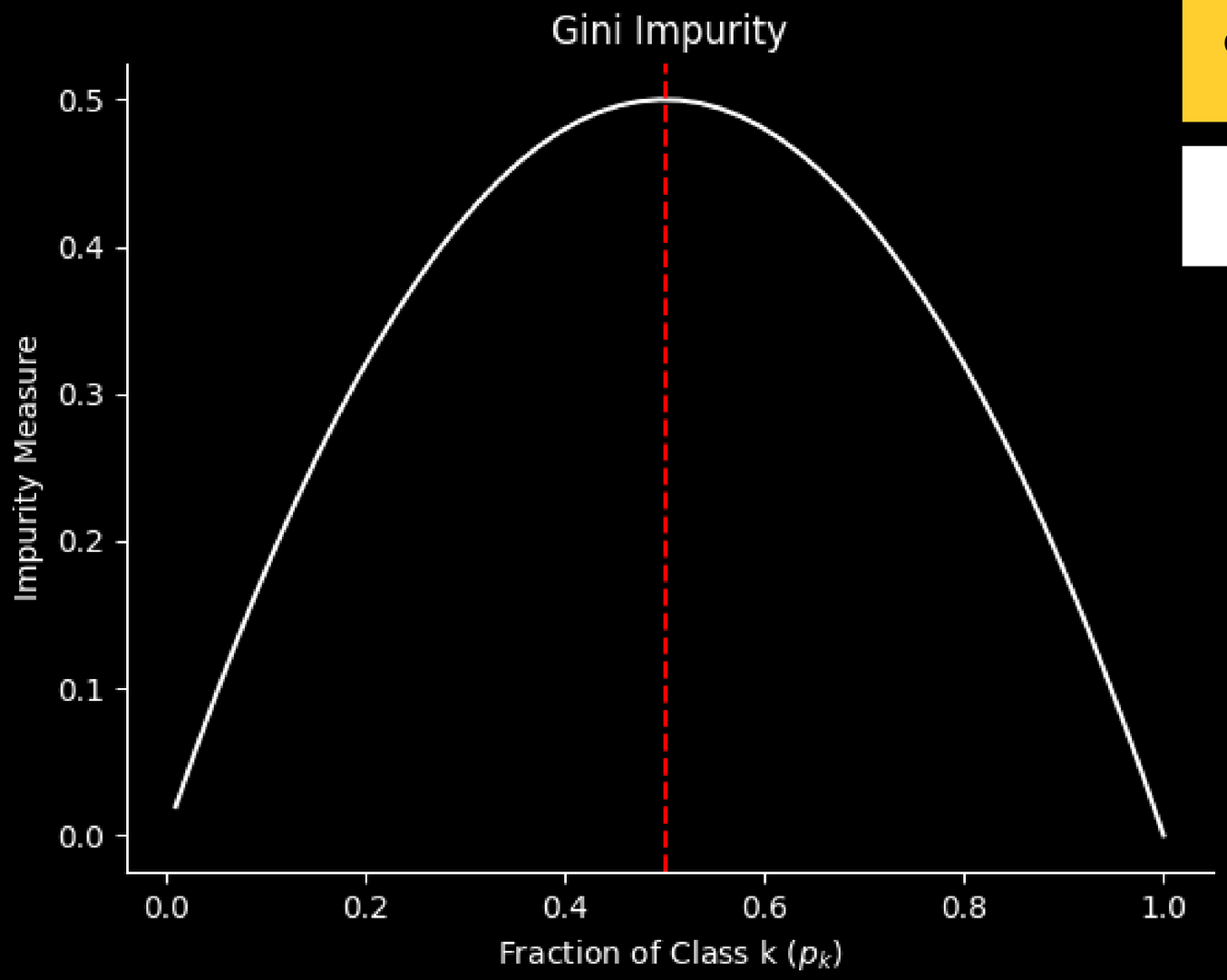
$$1 - \left(\frac{54}{60}\right)^2 - \left(\frac{6}{60}\right)^2 = 0.18$$

Calculating Gini impurity for a subnode

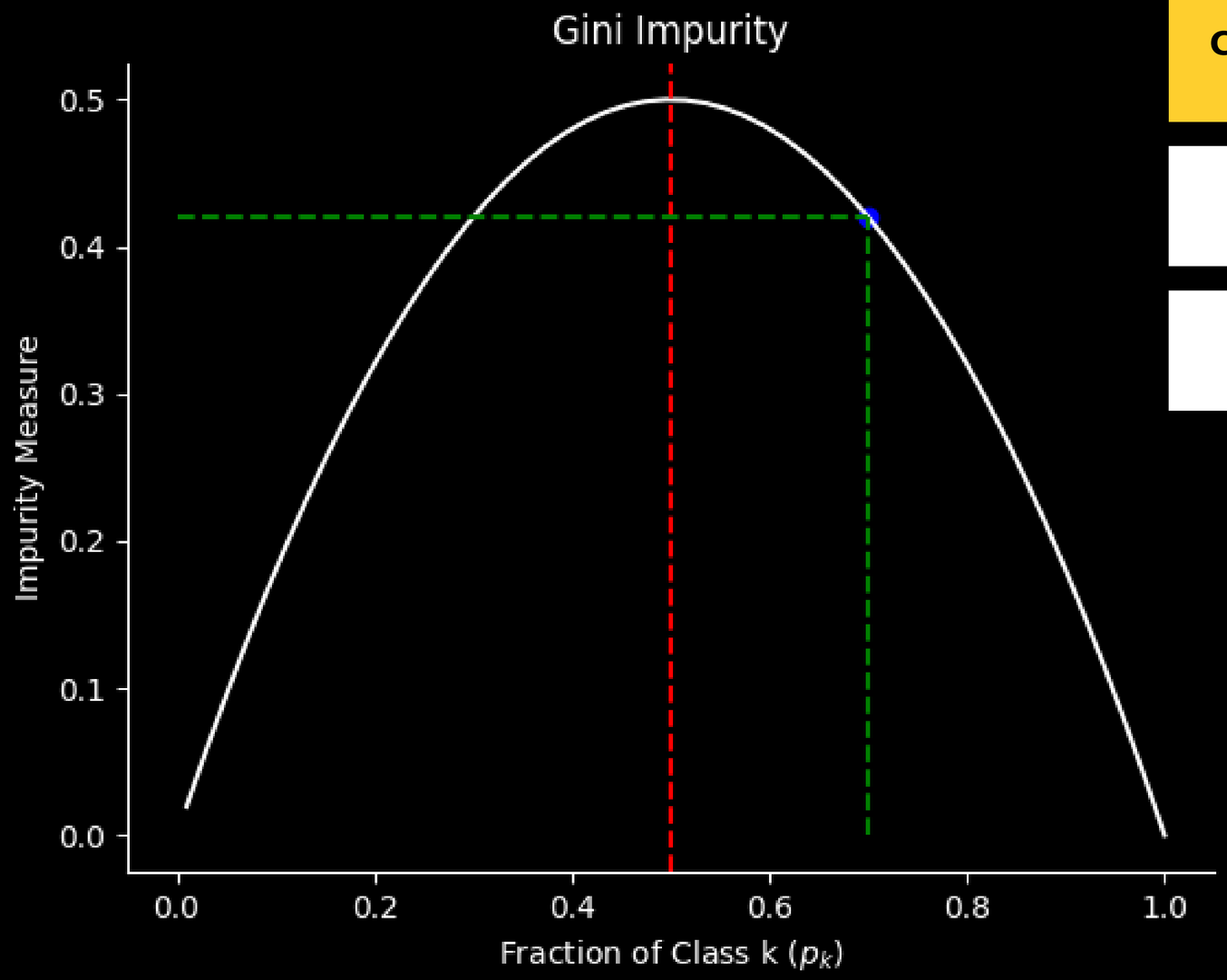
$$1 - \left(\frac{10}{10}\right)^2 - \left(\frac{0}{10}\right)^2 = 0$$

Why does this work?

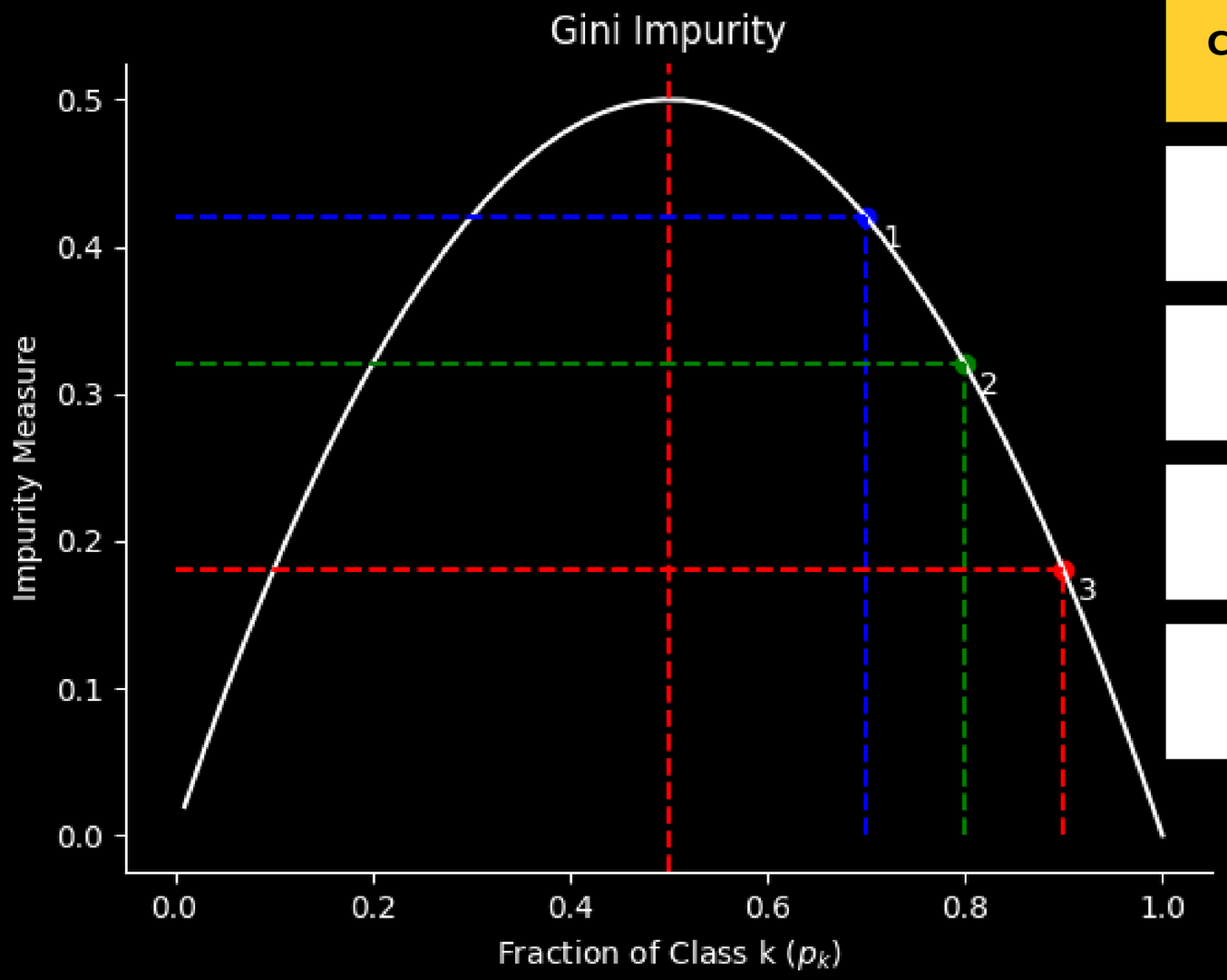
Max of gini impurity is 0.5
Min of gini impurity is when probability of tea or coffee is 1



Coffee class	Tea class	GINI IMPURITY
0.5	0.5	0.5

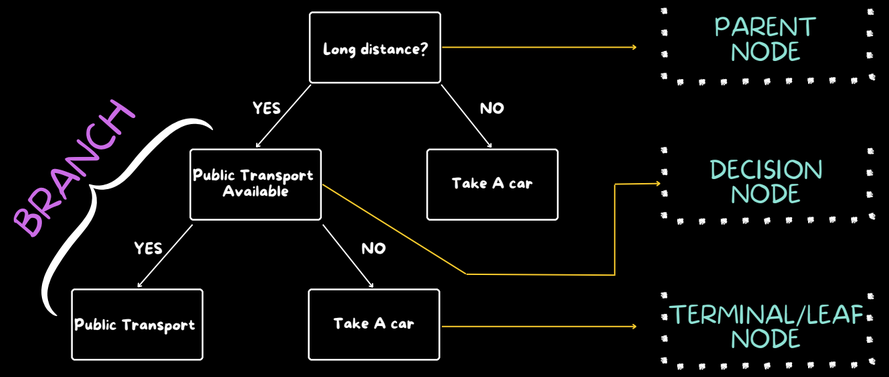


Coffee class	Tea class	GINI IMPURITY
0.5	0.5	0.5
0.7	0.3	0.42



Coffee class	Tea class	GINI IMPURITY
0.5	0.5	**0.5**
0.7	0.3	**0.42**
0.8	0.2	**0.32**
0.9	0.1	**0.18**

Lets Recap



ENTROPY

SPLITTING CONDITION
1. GINI IMPURITY
2. ENTROPY & INFORMATION GAIN

P1= probability of tea in any part of a branch

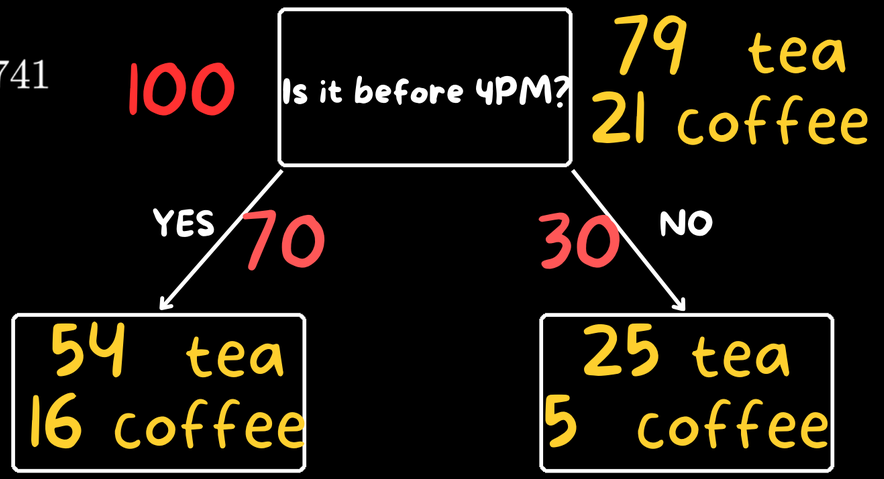
P2 = probability of coffee in any part of a branch

Calculate entropy of parent node

$$-\left(P1 \log_2 P1 + P2 \log_2 P2\right)$$

Calculate for sub-nodes

$$-\frac{79}{100}\log_2\frac{79}{100}-\frac{21}{100}\log_2\frac{21}{100}=0.741$$

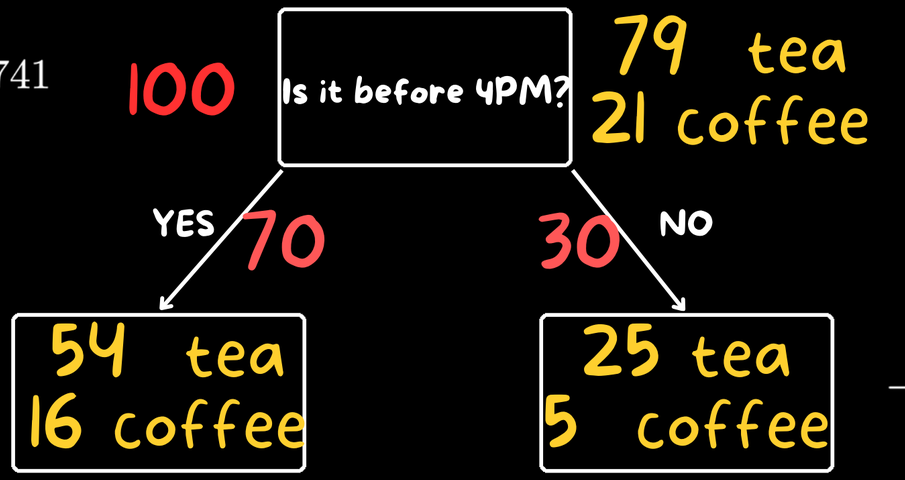


$$-\frac{54}{70}\log_2\frac{54}{70}-\frac{16}{70}\log_2\frac{16}{70}=0.775$$

$$-\frac{25}{30}\log_2\frac{25}{30}-\frac{25}{30}\log_2\frac{25}{30}=0.438$$

Calculate for split

$$-\frac{79}{100}\log_2\frac{79}{100}-\frac{21}{100}\log_2\frac{21}{100}=0.741$$

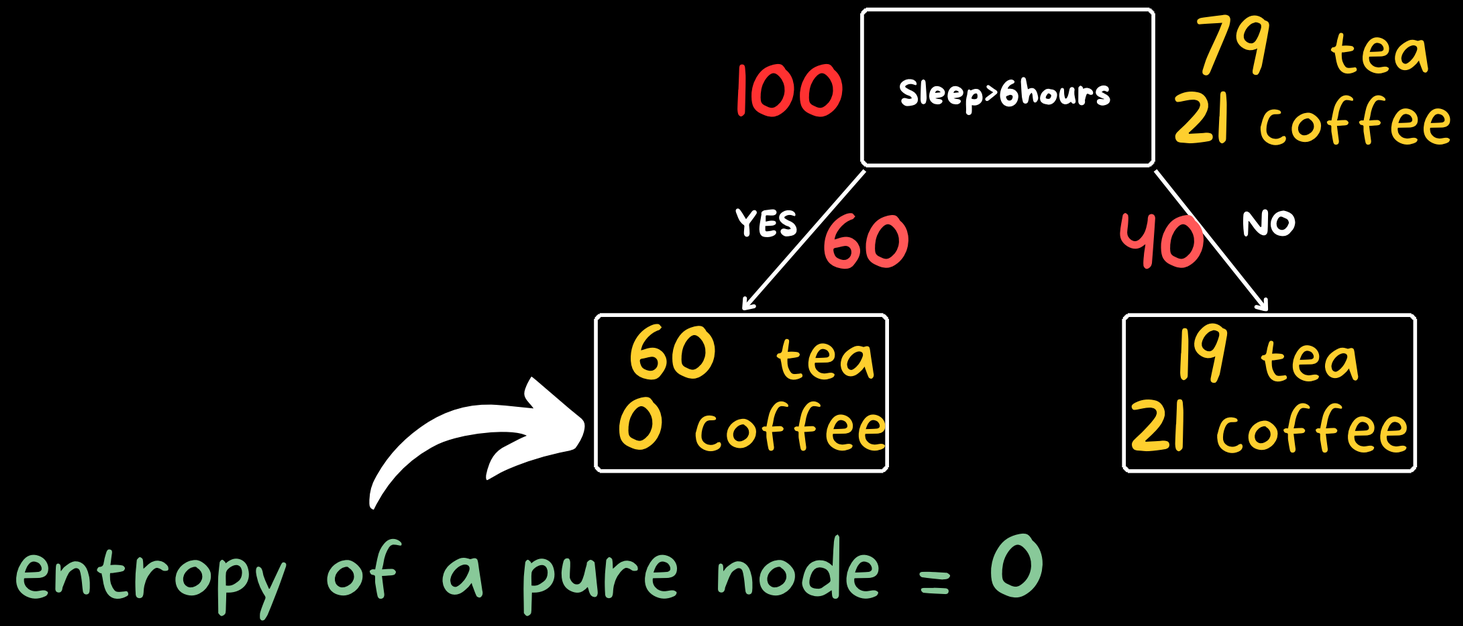


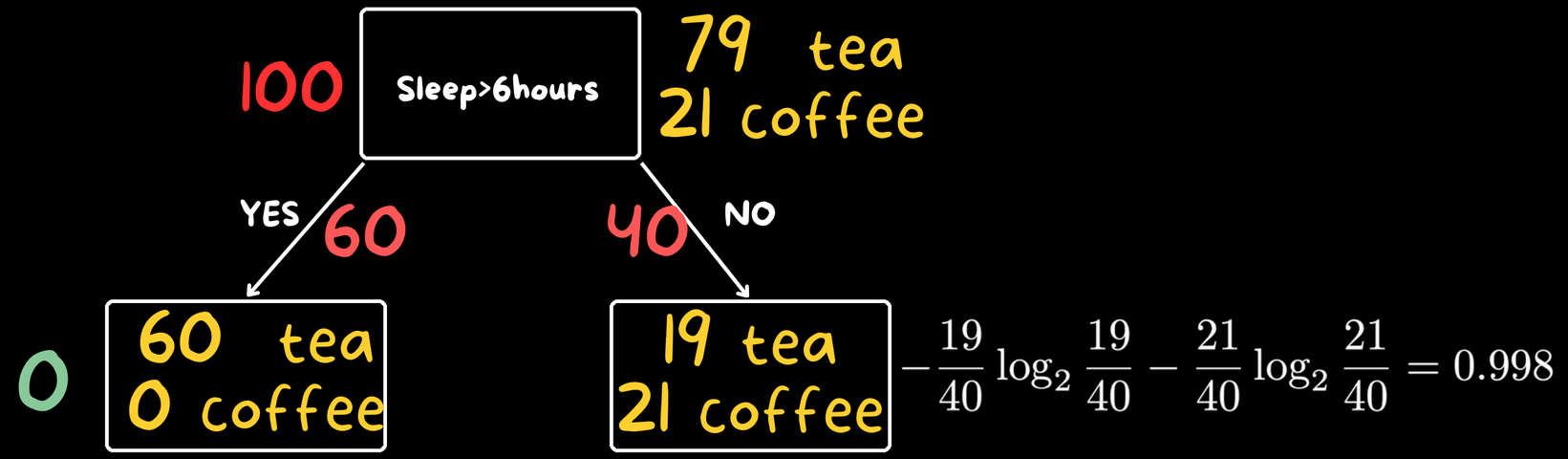
$$-\frac{54}{70}\log_2\frac{54}{70}-\frac{16}{70}\log_2\frac{16}{70}=0.775$$

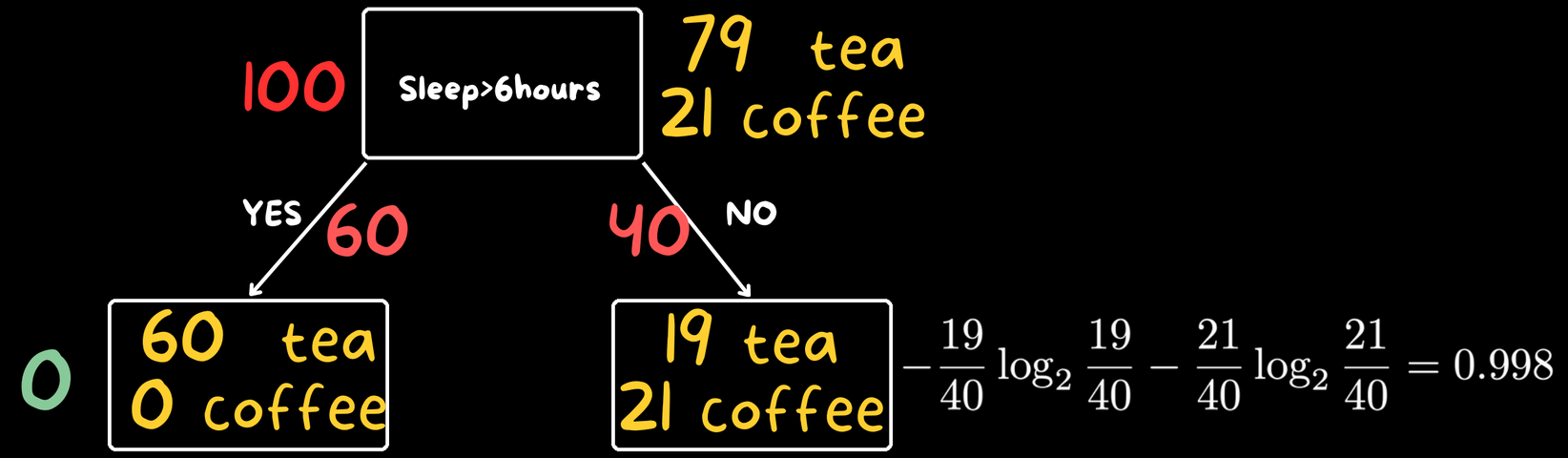
$$-\frac{25}{30}\log_2\frac{25}{30}-\frac{25}{30}\log_2\frac{25}{30}=0.438$$

entropy of split

0.775(70/100)+0.438(30/100)=0.673


100
Sleep>6hours
79 tea
21 coffee
YES
60
40
NO
60 tea
0 coffee
19 tea
21 coffee
entropy of a pure node = 0


100
Sleep>6hours
79 tea
21 coffee
YES
60
40
NO
0
60 tea
0 coffee
19 tea
21 coffee
$-\frac{19}{40}\log_2\frac{19}{40}-\frac{21}{40}\log_2\frac{21}{40}=0.998$



entropy of split

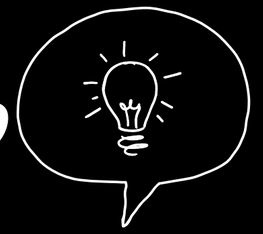
0(60/100)+0.998(40/100)=0.399

entropy of previous split

0.775(70/100)+0.438(30/100)=0.673

INFORMATION GAIN

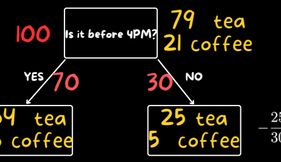
How do we think of this?



Information gain for a split= Entropy before the split- entropy of the split

Information gain for a 4pm split
= Entropy before the split- entropy of the split
= 0.741 - 0.673 = 0.068

$$-\frac{79}{100}\log_2\frac{79}{100}-\frac{21}{100}\log_2\frac{21}{100}=0.741$$



$$-\frac{54}{70}\log_2\frac{54}{70}-\frac{16}{70}\log_2\frac{16}{70}=0.775$$

$$-\frac{25}{30}\log_2\frac{25}{30}-\frac{25}{30}\log_2\frac{25}{30}=0.438$$

entropy of split
0.775(70/100)+0.438(30/100)=0.673

PRUNING

PrePruning

Setting a maximum depth to a tree

PostPruning

Reduces the size of decision trees by removing sections of
the tree that are non-critical and redundant to classify instances

We first make the decision tree to a large depth.

Suppose a split is giving us a gain of say -10 (loss of 10) and then the next split on that gives us a gain of 20.