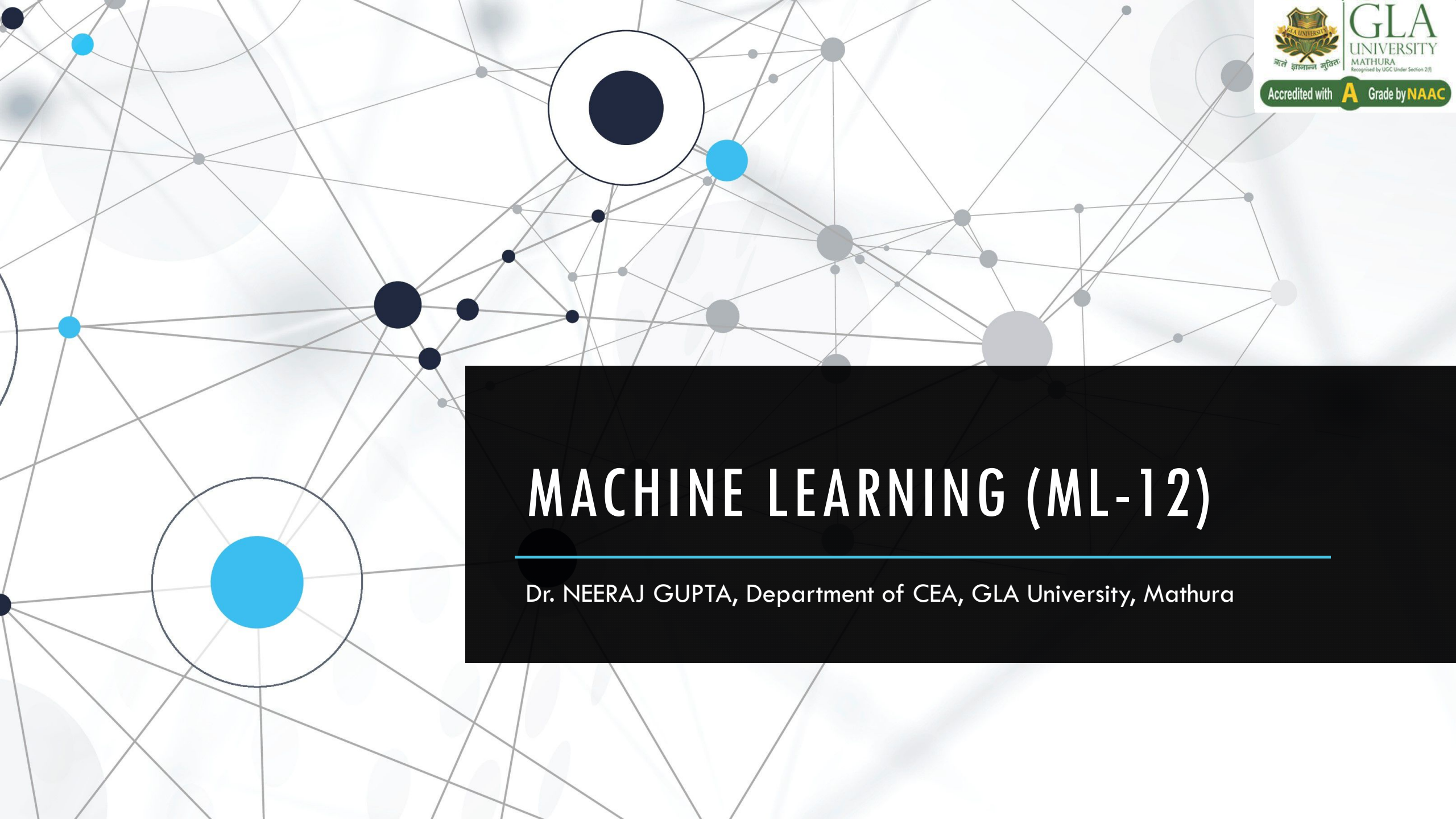


MACHINE LEARNING (ML-12)

Dr. NEERAJ GUPTA, Department of CEA, GLA University, Mathura

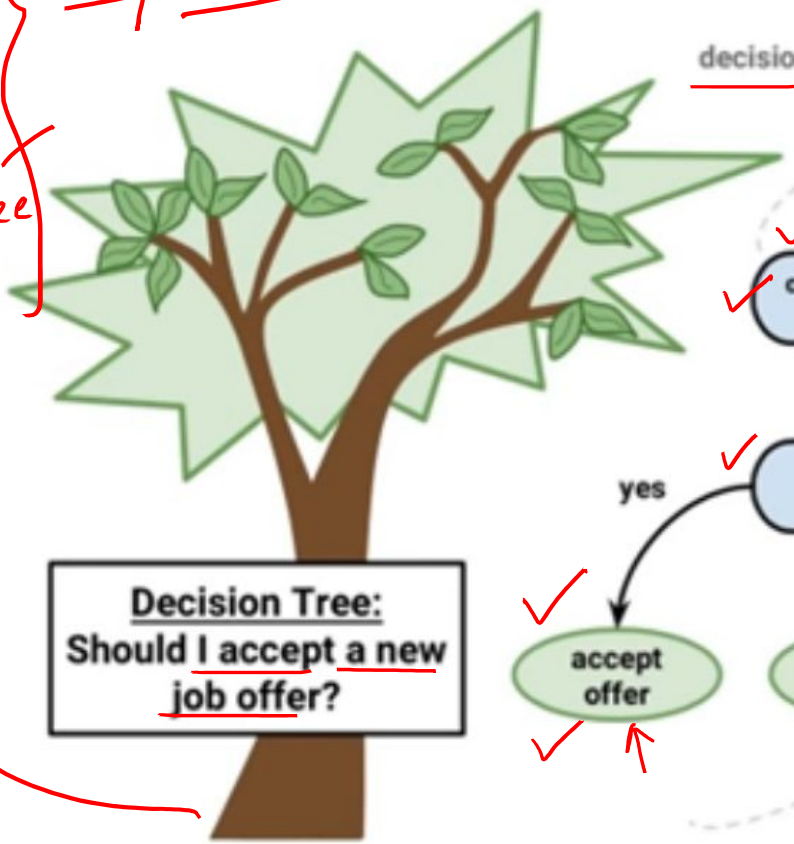
AGENDA

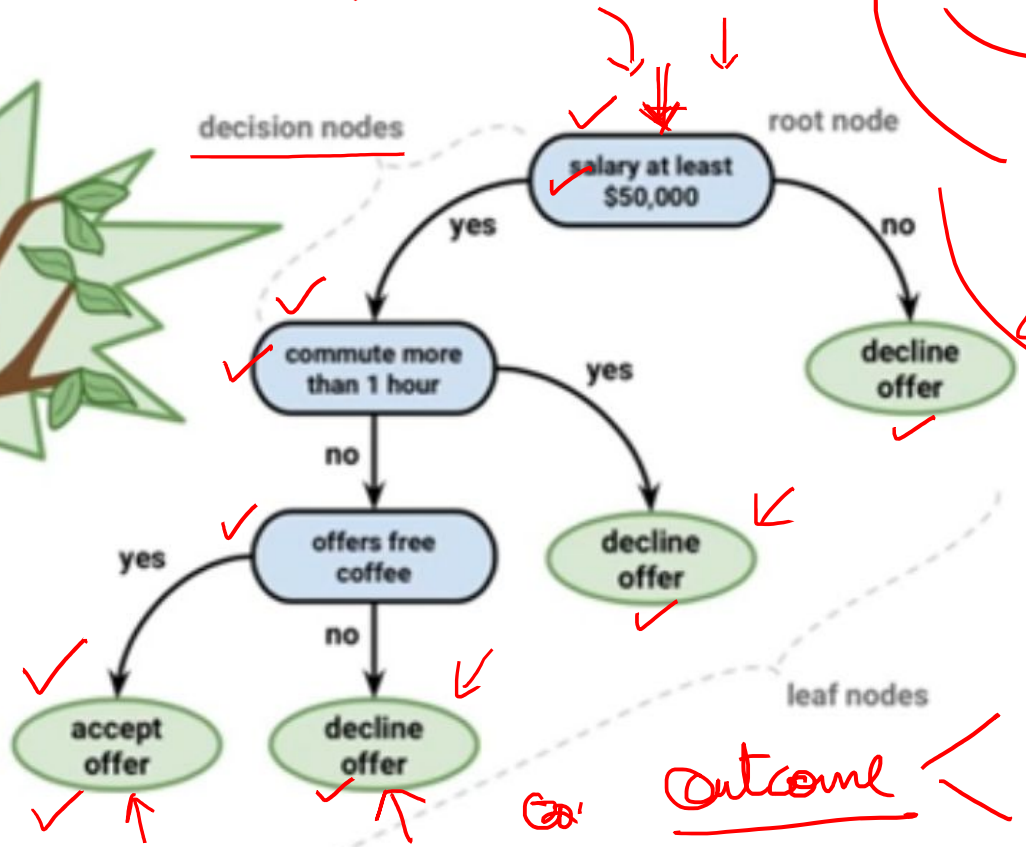
- Decision Tree

DECISION TREE

A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions.

DECISION TREE



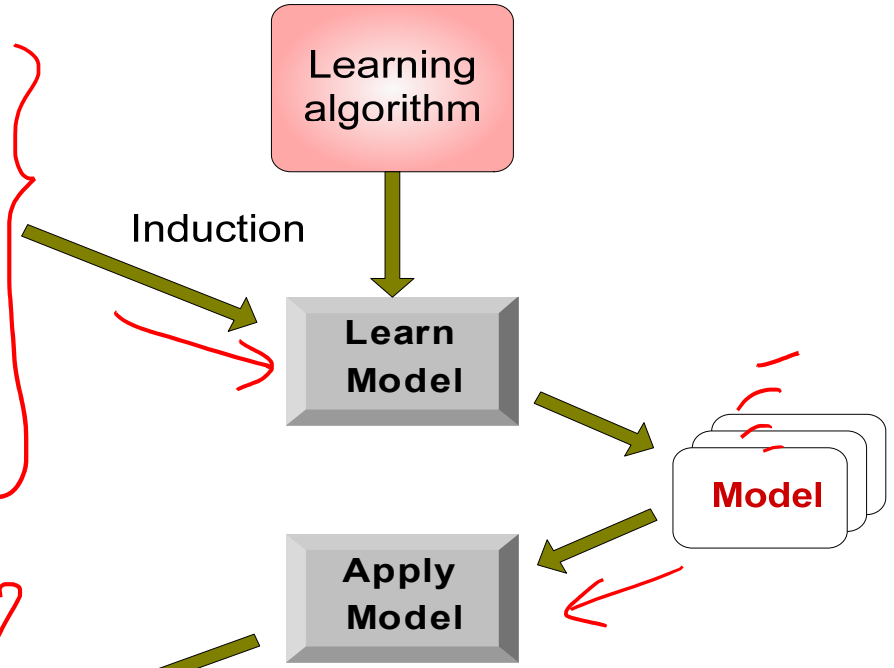


ILLUSTRATING CLASSIFICATION TASK

Dataset
Train
Test

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

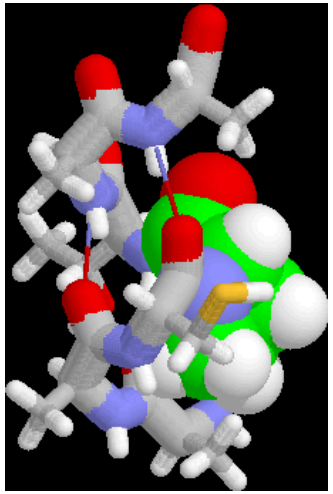
Test Set

EXAMPLES OF CLASSIFICATION TASK

Predicting tumor cells as benign or malignant

Classifying credit card transactions
as legitimate or fraudulent

Classifying secondary structures of protein
as alpha-helix, beta-sheet, or random
coil

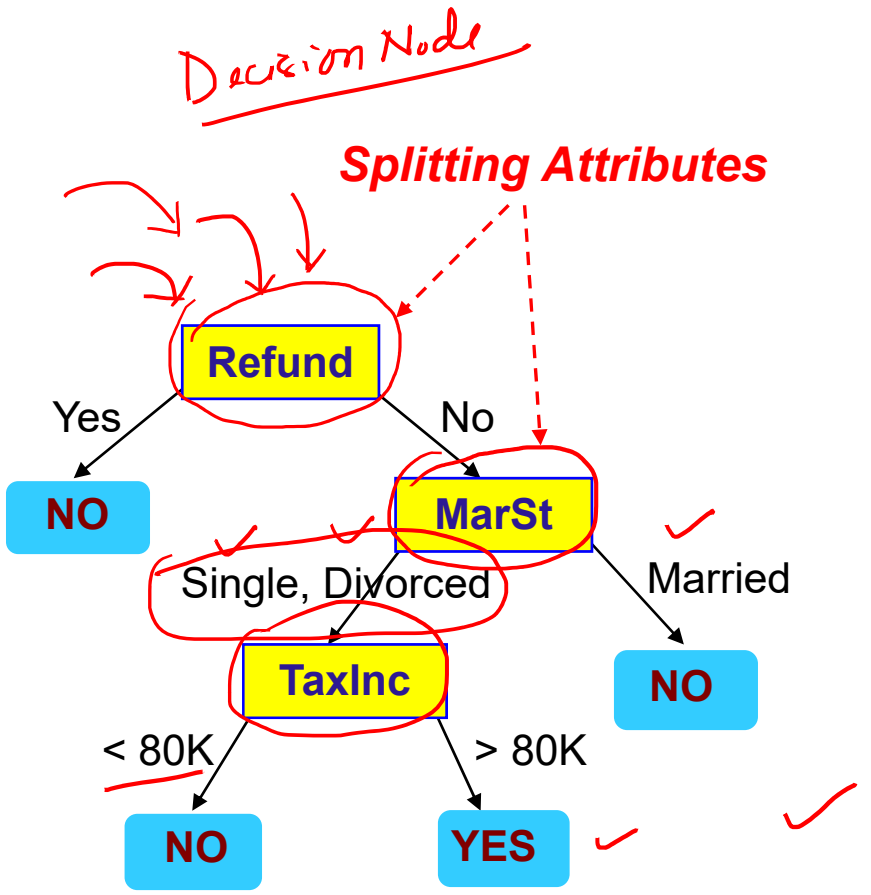


Categorizing news stories as finance,
weather, entertainment, sports, etc

EXAMPLE OF A DECISION TREE

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

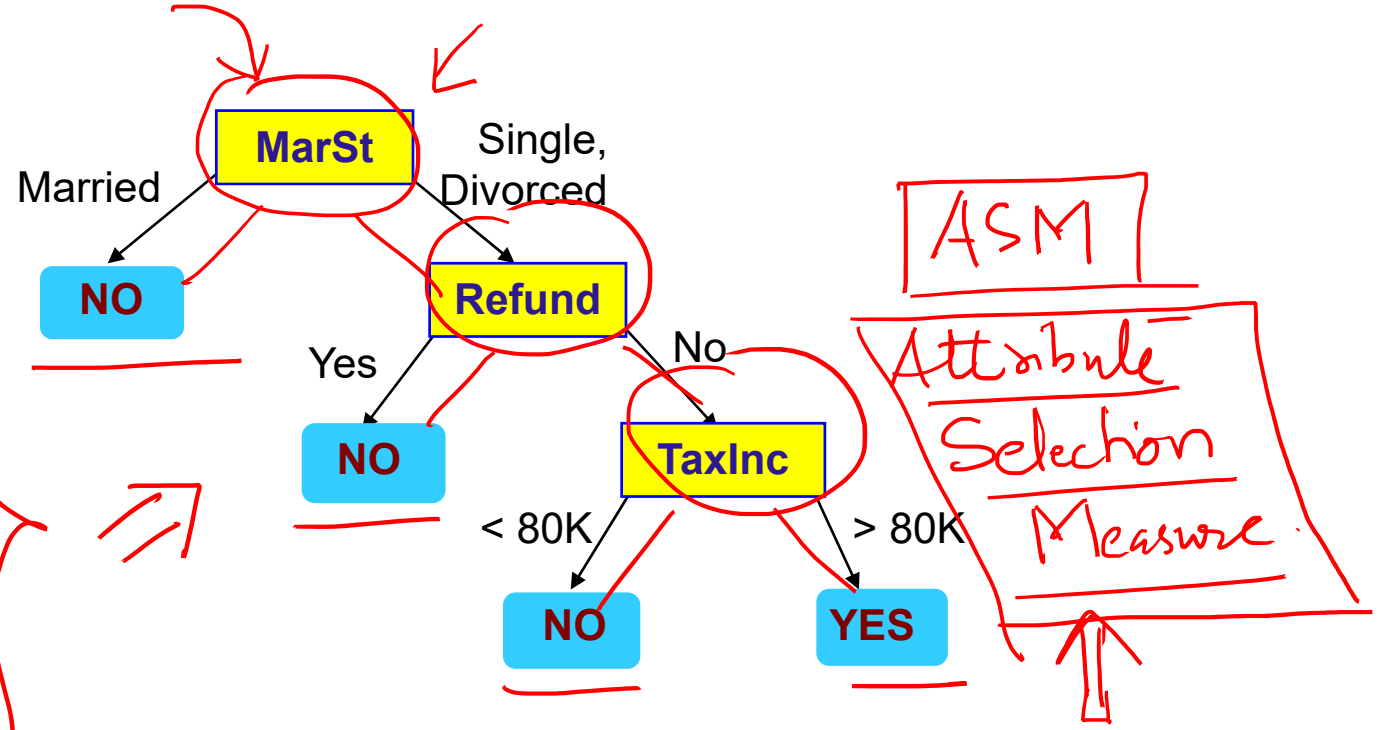
Training Data



Model: Decision Tree

ANOTHER EXAMPLE OF DECISION TREE

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



ASM
Attribute Selection Measure

There could be more than one tree that fits the same data!

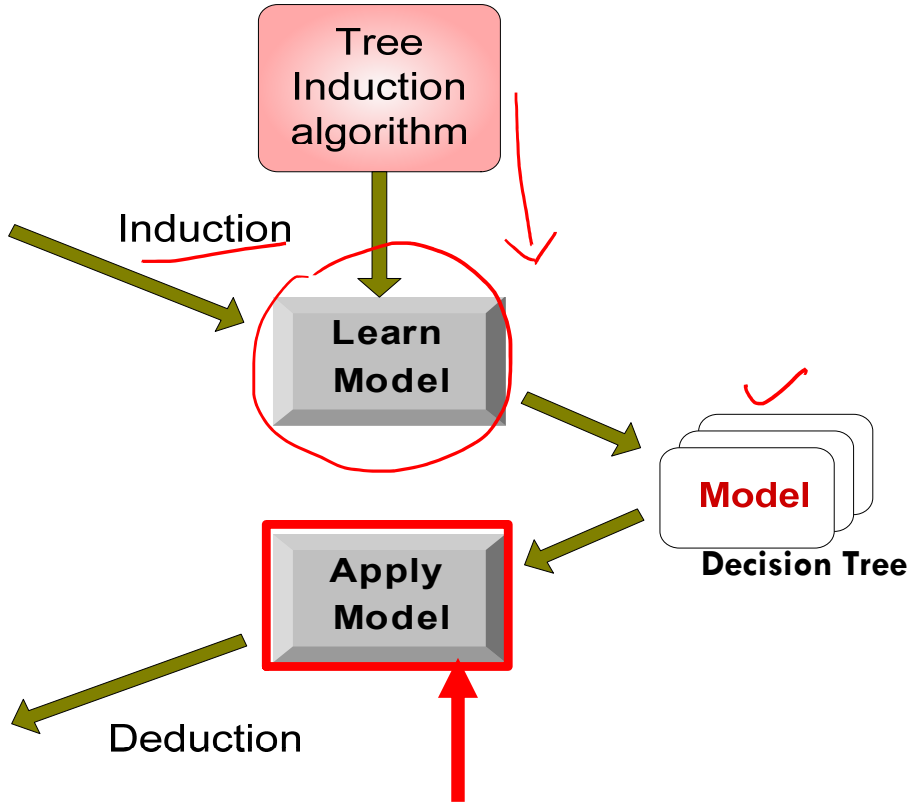
DECISION TREE CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

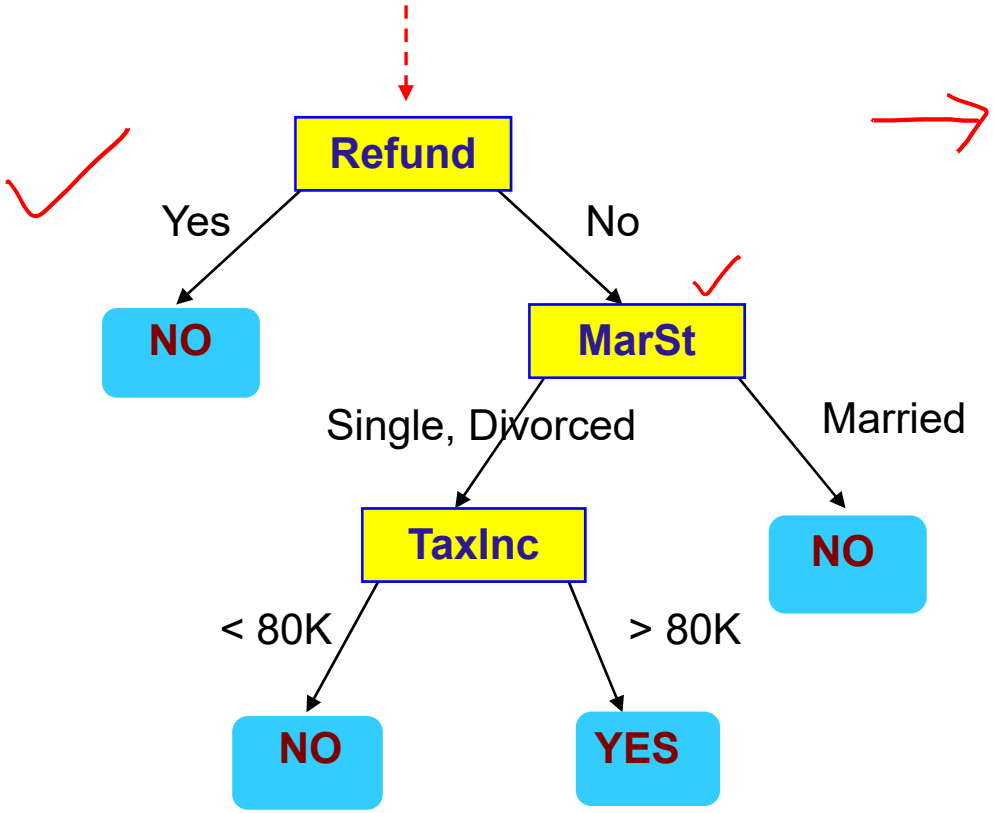
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



APPLY MODEL TO TEST DATA

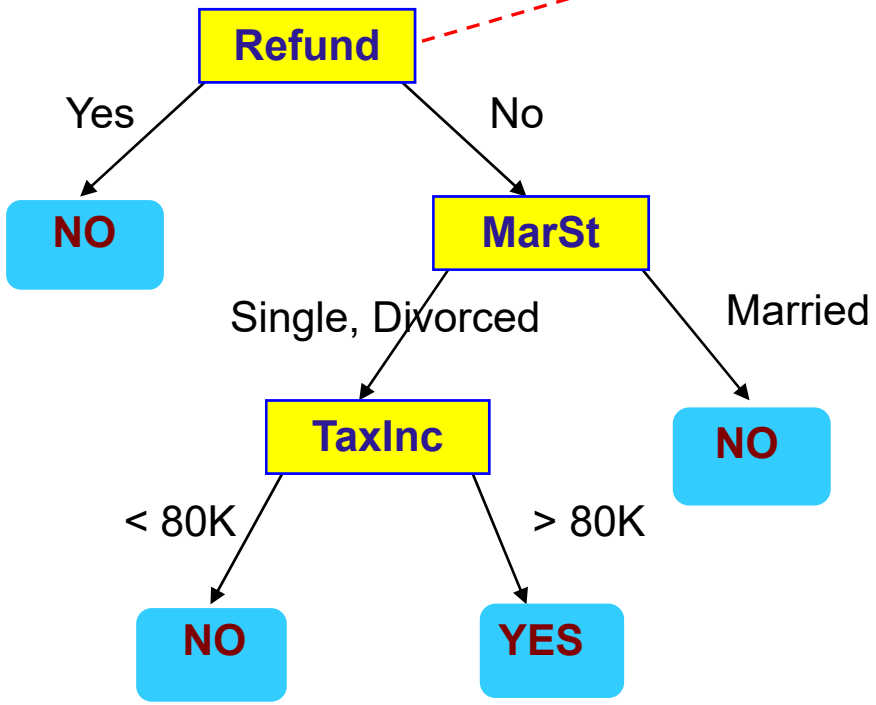
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

APPLY MODEL TO TEST DATA



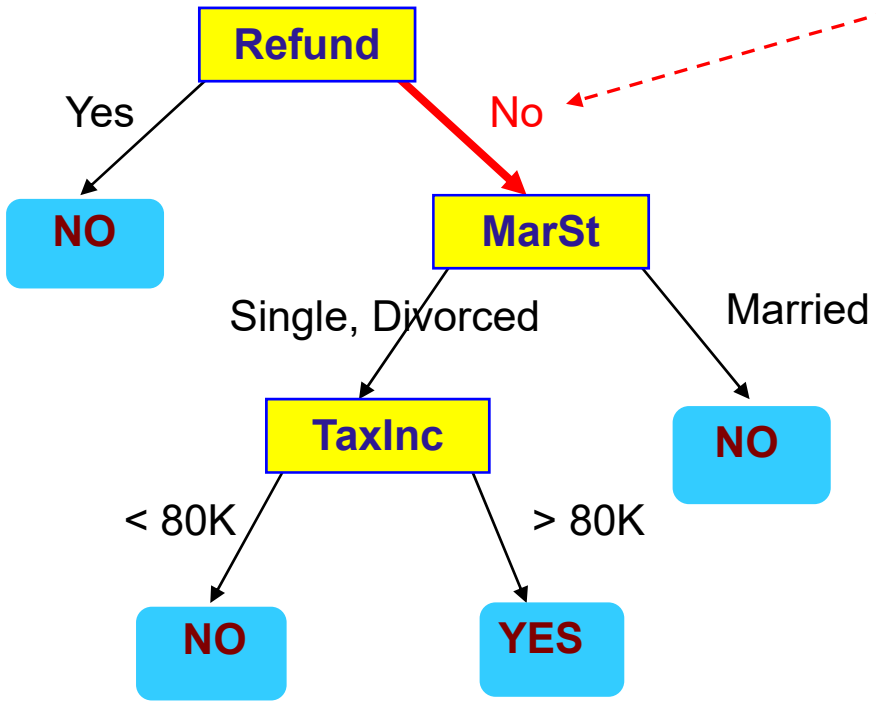
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

APPLY MODEL TO TEST DATA

Test Data

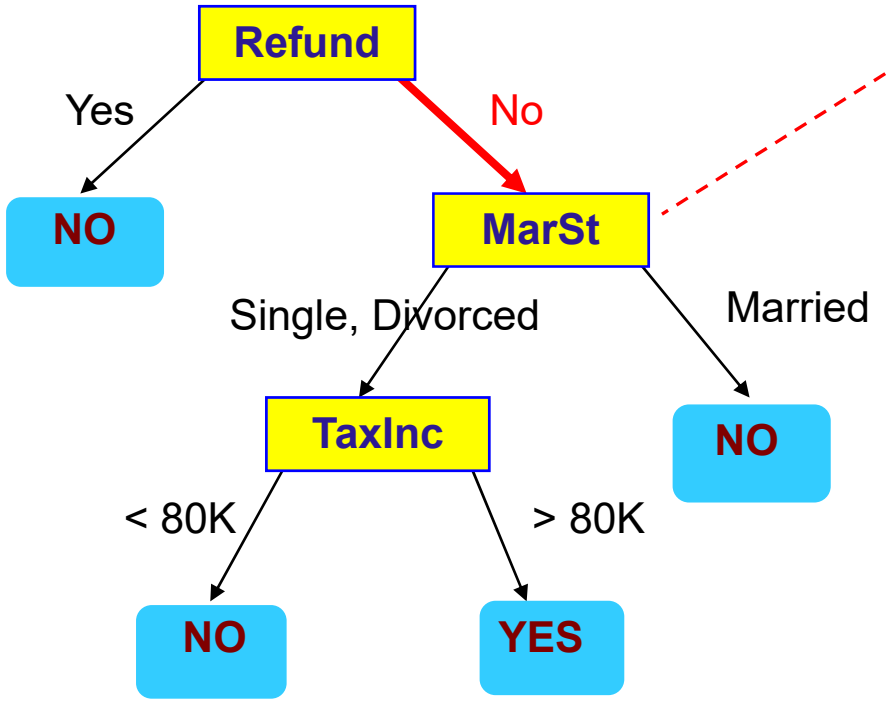
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APPLY MODEL TO TEST DATA

Test Data

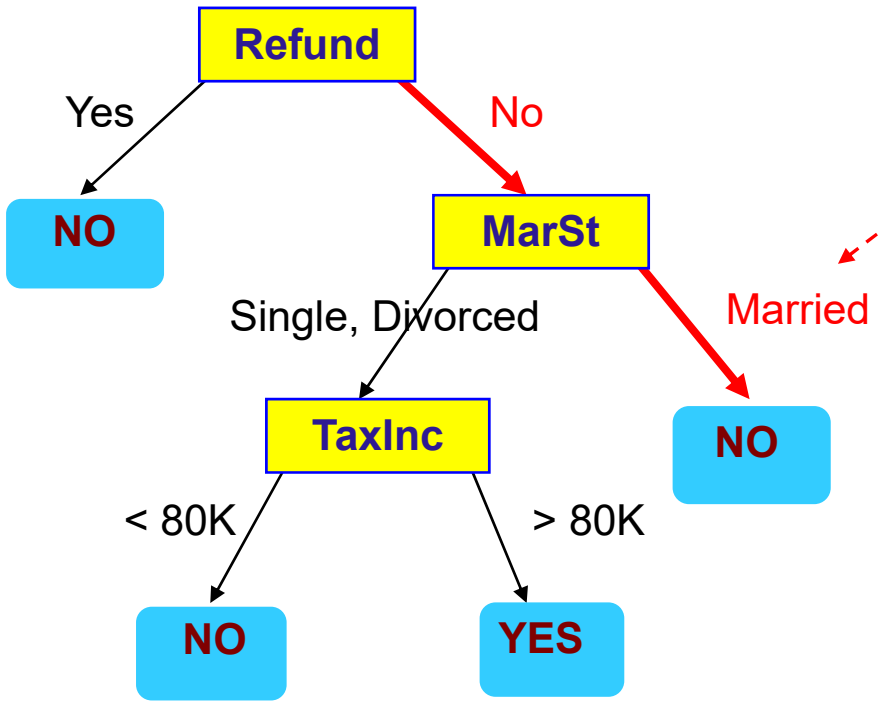
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



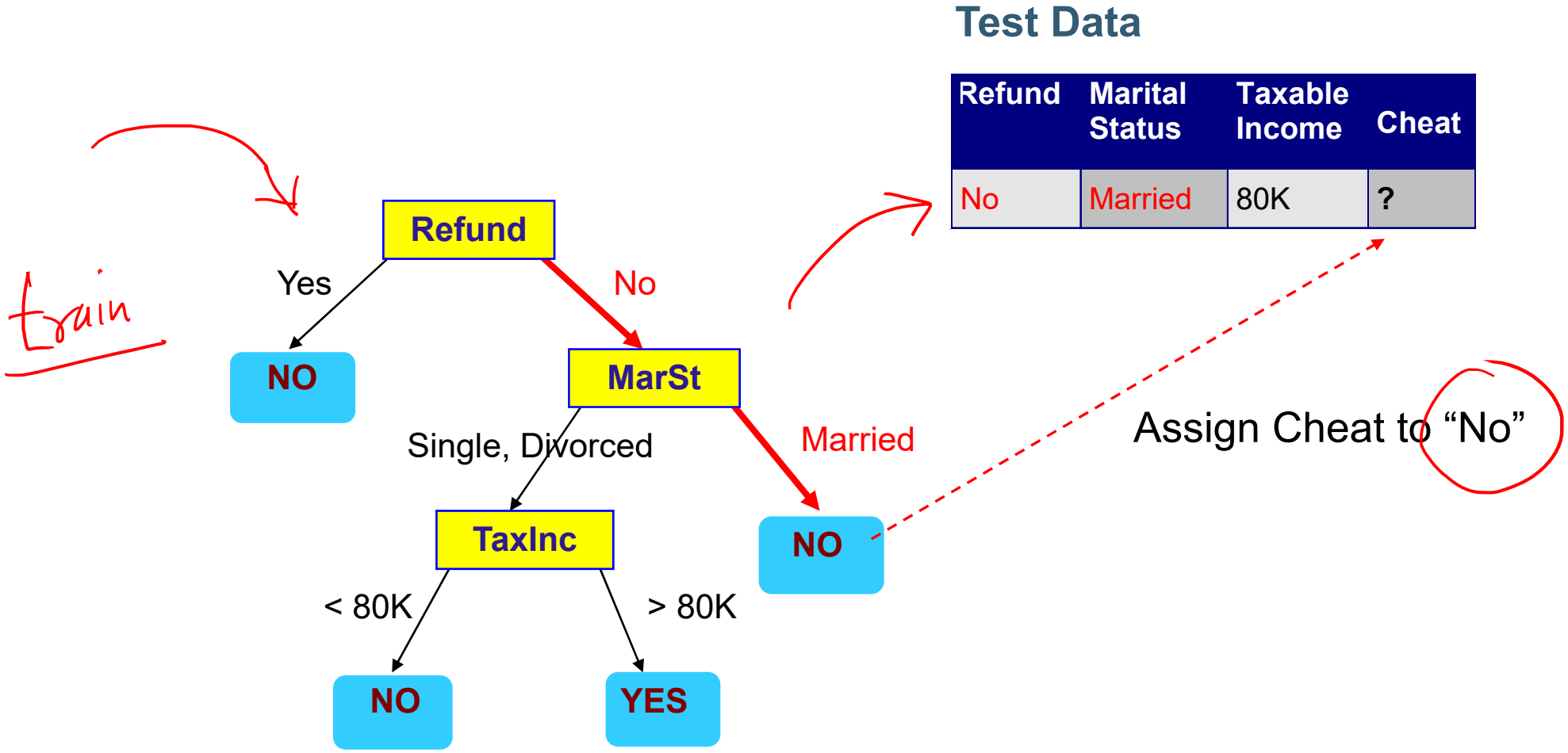
APPLY MODEL TO TEST DATA

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APPLY MODEL TO TEST DATA



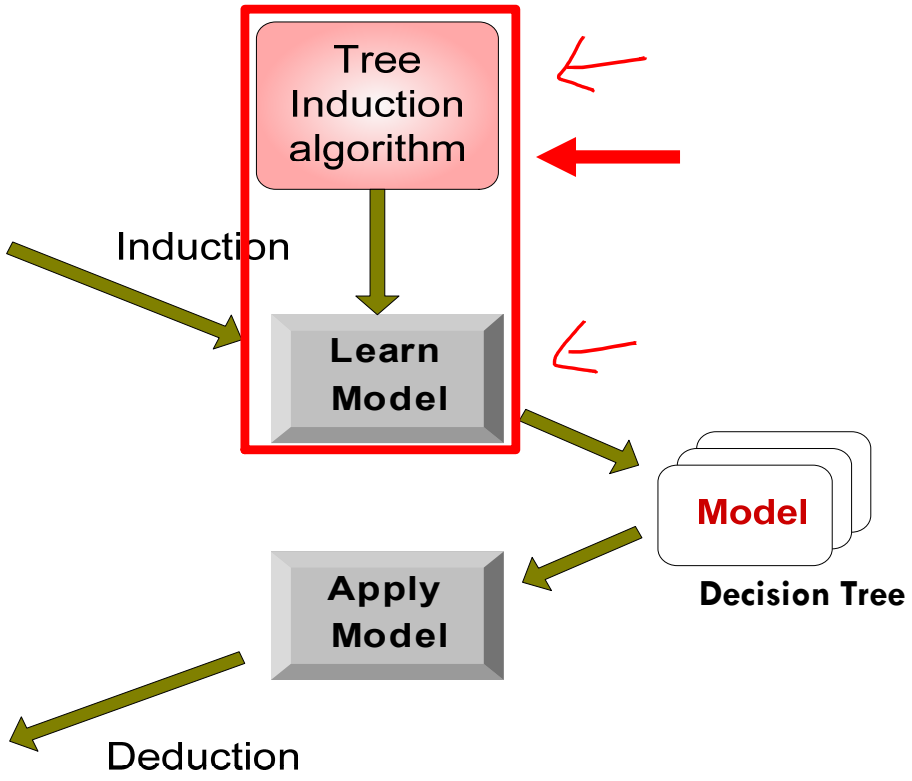
DECISION TREE CLASSIFICATION TASK

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



DECISION TREE INDUCTION

Many Algorithms:

- Hunt's Algorithm (one of the earliest)
- CART
- ID3, C4.5
- SLIQ, SPRINT

TREE INDUCTION

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

Issues

- Determine how to split the records
  - How to specify the attribute test condition?
  - How to determine the best split?
- Determine when to stop splitting

TREE INDUCTION

Greedy strategy.

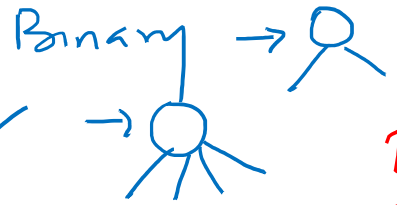
- Split the records based on an attribute test that optimizes certain criterion.

Issues

- Determine how to split the records
  - How to specify the attribute test condition?
  - How to determine the best split?
- Determine when to stop splitting

HOW TO SPECIFY TEST CONDITION?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous
- Depends on number of ways to split
  - 2-way split
  - Multi-way split



Quantitative Vs Qualitative data

Count

Discrete Vs Continuous

like 1, 2, 10 / whole no

Temp
32.543
31.639

Decimal place

like weight of bike
152.6 kg
148.9

Categorical

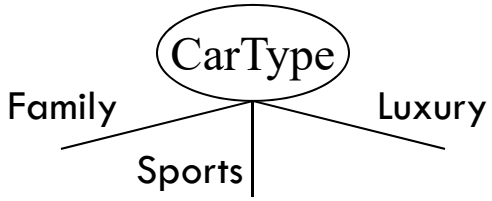
Nominal
to label variable
Gender (Male/Female)
Nationalities (Indian, American, African)

Speed

Ordinal (order)
likely
Neutral
Unlikely

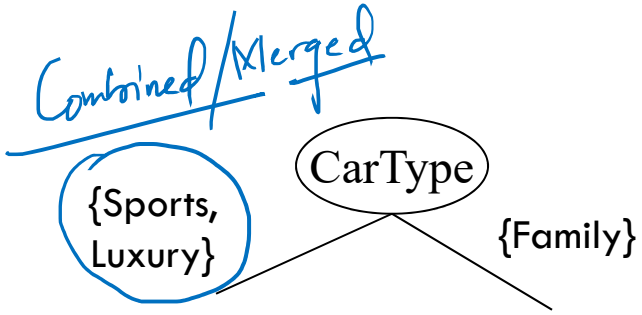
SPLITTING BASED ON NOMINAL ATTRIBUTES

Multi-way split: Use as many partitions as distinct values.

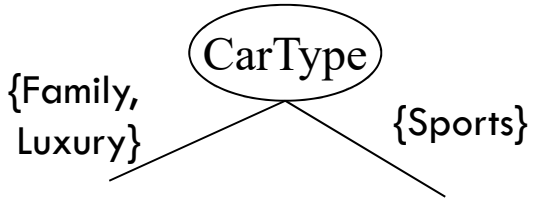


Binary split: Divides values into two subsets.
Need to find optimal partitioning.

Combined/Merged

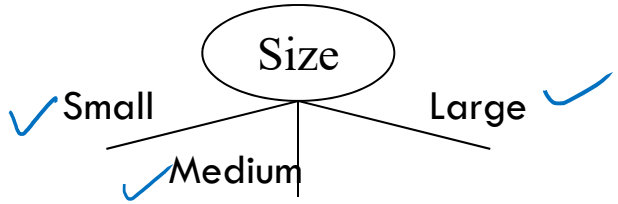


OR

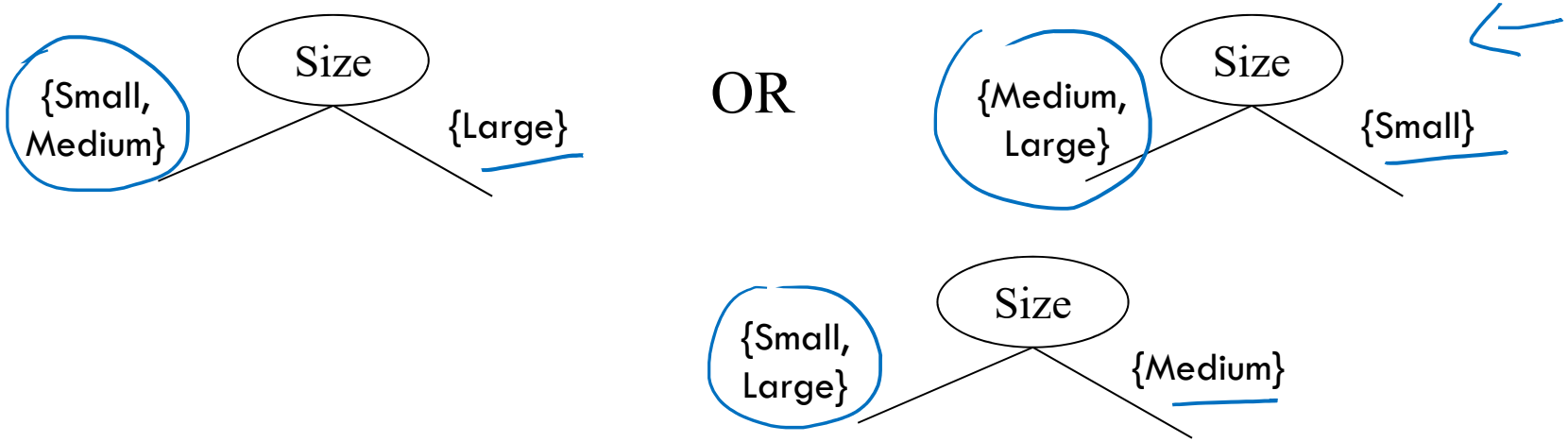


SPLITTING BASED ON ORDINAL ATTRIBUTES

Multi-way split: Use as many partitions as distinct values.



Binary split: Divides values into two subsets.
Need to find optimal partitioning.



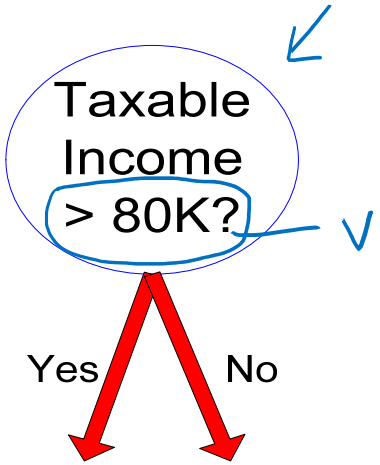
SPLITTING BASED ON CONTINUOUS ATTRIBUTES

Different ways of handling

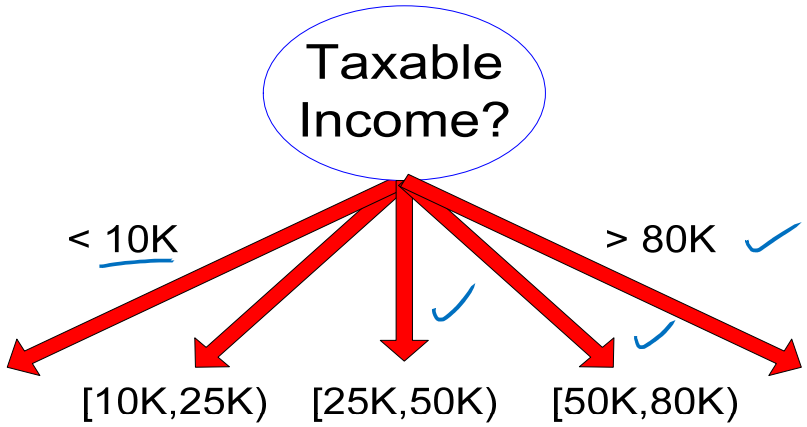
- Discretization to form an ordinal categorical attribute
  - Static – discretize once at the beginning
  - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.

- Binary Decision: $(A < v)$ or $(A \geq v)$ ← Two options.
  - consider all possible splits and finds the best cut
  - can be more compute intensive

SPLITTING BASED ON CONTINUOUS ATTRIBUTES



(i) Binary split



(ii) Multi-way split

TREE INDUCTION

Greedy strategy.

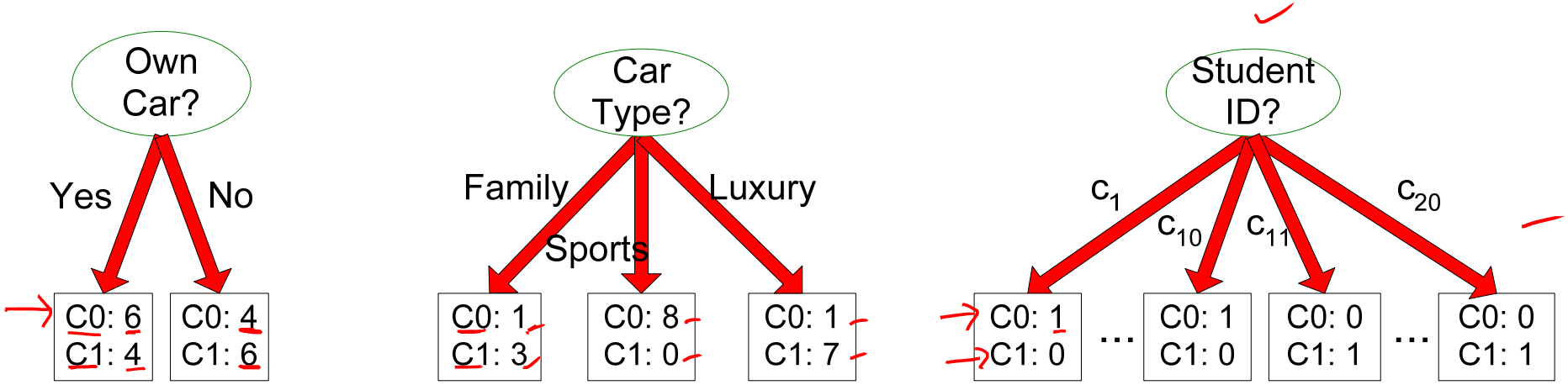
- Split the records based on an attribute test that optimizes certain criterion.

Issues

- Determine how to split the records
  - How to specify the attribute test condition?
  - How to determine the best split?
- Determine when to stop splitting

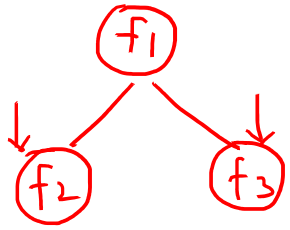
HOW TO DETERMINE THE BEST SPLIT

Before Splitting: 10 records of class 0,
10 records of class 1



Which test condition is the best?

HOW TO DETERMINE THE BEST SPLIT



Greedy approach:

- Nodes with homogeneous class distribution are preferred

Need a measure of node impurity:

C0: 5
C1: 5

Non-homogeneous,

High degree of impurity

C0: 9
C1: 1

Homogeneous,

Low degree of impurity

MEASURES OF NODE IMPURITY

- Entropy
- Gini Index
- Misclassification error

ENTROPY

target

f_1	f_2	f_3	f_4	
outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

We are dealing with categorical variables (predictors) or features: **outlook, temperature, humidity, wind**

~ we want to predict whether to play golf or not (play is he **target variable**)

ENTROPY

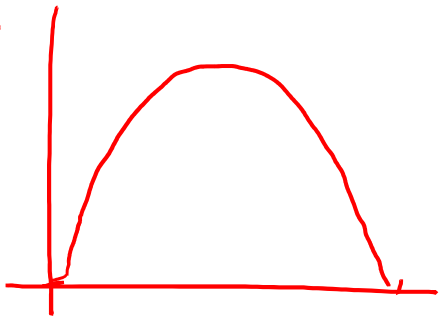
The **H(X)** Shannon-entropy of a dicrete random variable **X** with possible values $x_1 x_2 \dots x_n$ and probability mass function **P(X)** is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i)$$

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i)$$

Example: https://en.wikipedia.org/wiki/Entropy_(information_theory)

For completely homogoeneous dataset (all TRUE or all FALSE values): entropy is **0**
If the dataset is equally divided (same amount of TRUEs and FALSEs): entropy is **1**



https://en.wikipedia.org/wiki/Entropy_(information_theory)

ENTROPY

outlook	temperature	humidity	wind	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

yes - 9
No - 5
} Tot - 14

$P(yes) = \frac{9}{14}$

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

→ **PLAYING GOLF**
→ **9** times **YES** ✓
→ **5** times **NO** ✓

We just have to use the Shannon-entropy formula to calculate the **H(x)** values

H(PlayingGolf) = H(9,5) =

= -(0.64 $\log_2$0.64) – (0.36 $\log_2$0.36) = 0.94

$-(0.64 \log_2 0.64) - (0.36 \log_2 0.36)$

$H(9,5) = 0.94$

$$H(9,5) = -\sum_{i=1}^{2} P_i(x_i) \log_2 P(x_i)$$

$$= -P(yes)\log_2 P(yes) - P(No)\log_2 P(No)$$

$$= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

ENTROPY

Information Gain

outlook	temperature	humidity	wind	play
sunny	hot	high	false	yes
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	no
rainy	cold	normal	false	yes
rainy	cold	normal	true	no
overcast	cold	normal	true	yes
sunny	mild	high	false	no
sunny	cold	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

$$E(T,X) = \sum_{x} P(x)\, E(x)$$

We have to calculate the entropy with respect to a given predictor/feature in order to be able to calculate information gain

		PLAY GOLF: YES	PLAY GOLF: NO
OUTLOOK	sunny	3	2
	Overcast	4	0
	rainy	2	3

$$E(3,2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$E(PlayGolf, Outlook) = P(sunny)E(3,2) + P(overcast)E(4,0) + P(rainy)E(2,3)$$

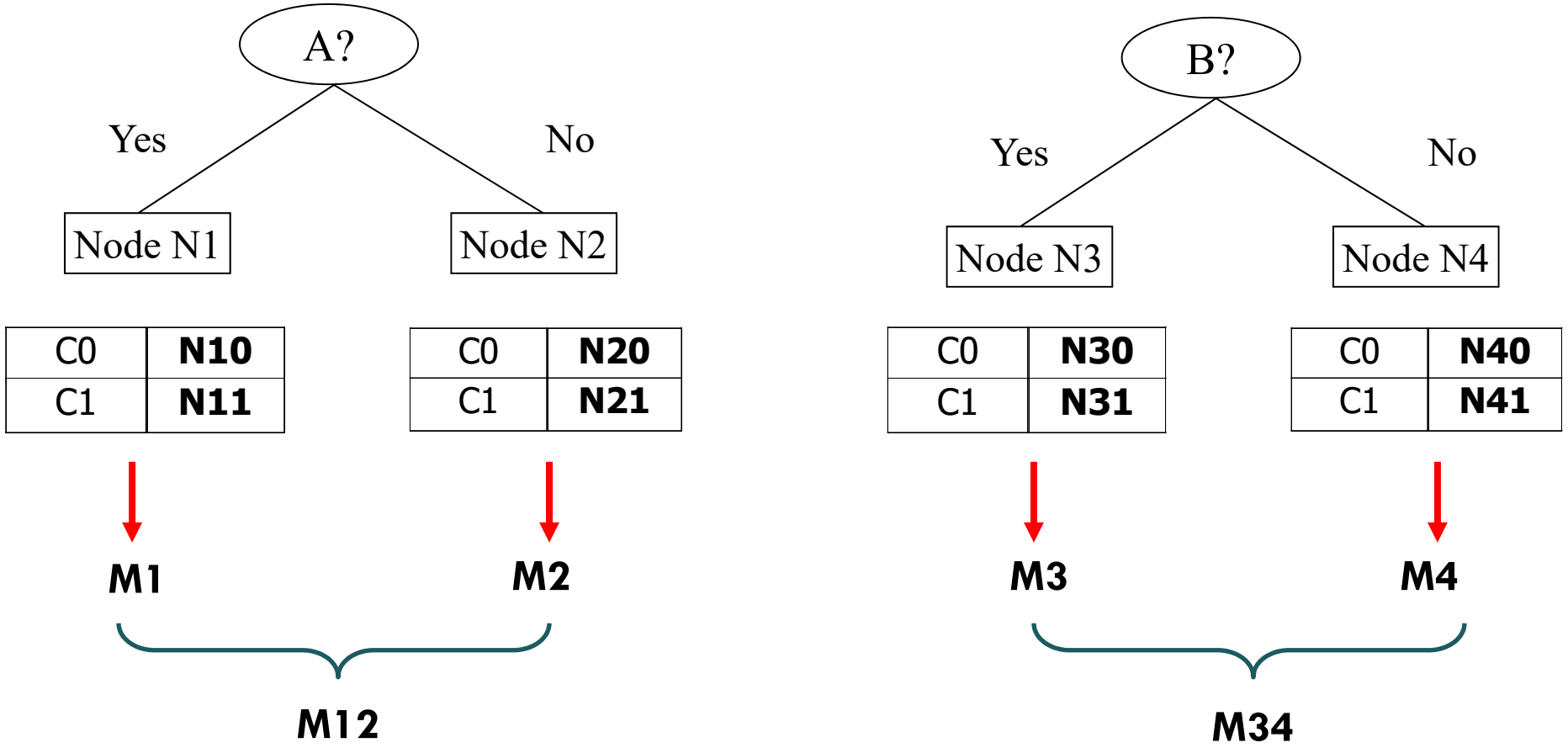
$$\frac{5}{14}\,0.971 + \frac{5}{14}\,0 + \frac{5}{14}\,0.971 = 0.693$$

HOW TO FIND THE BEST SPLIT

Before Splitting:

C0	**N00**
C1	**N01**

→ **M0**



Gain = M0 – M12 vs M0 – M34

MEASURE OF IMPURITY: GINI

Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

Homogeneity is decreasing

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Homo

Gini Index value is increasing

EXAMPLES FOR COMPUTING GINI

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

j - 2

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

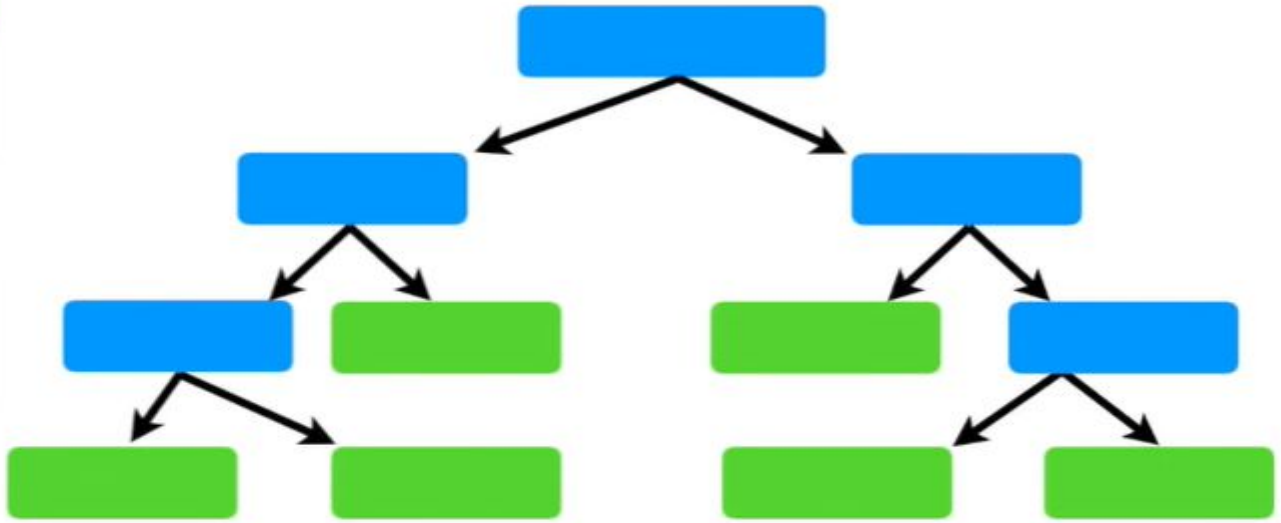
0 - 1

- Entropy
- Gini Index

EXAMPLE

In this example, we want to create a tree that uses **chest pain**, **good blood circulation** and **blocked artery status** to predict…

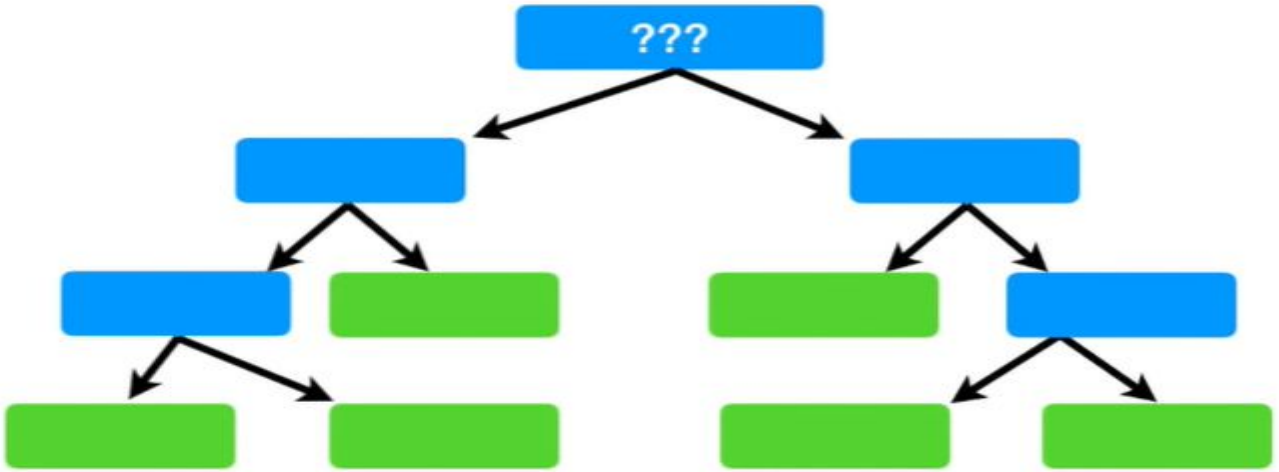
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



EXAMPLE

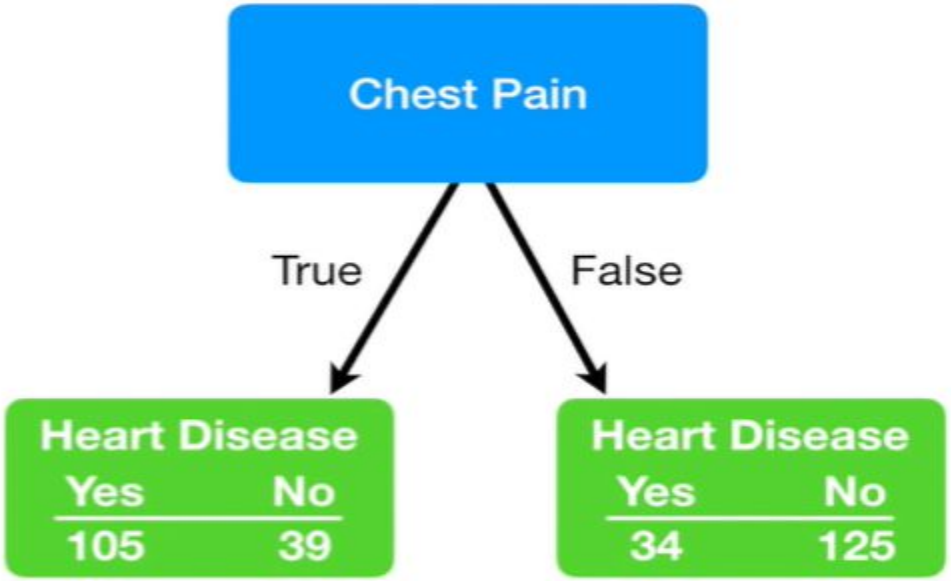
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.



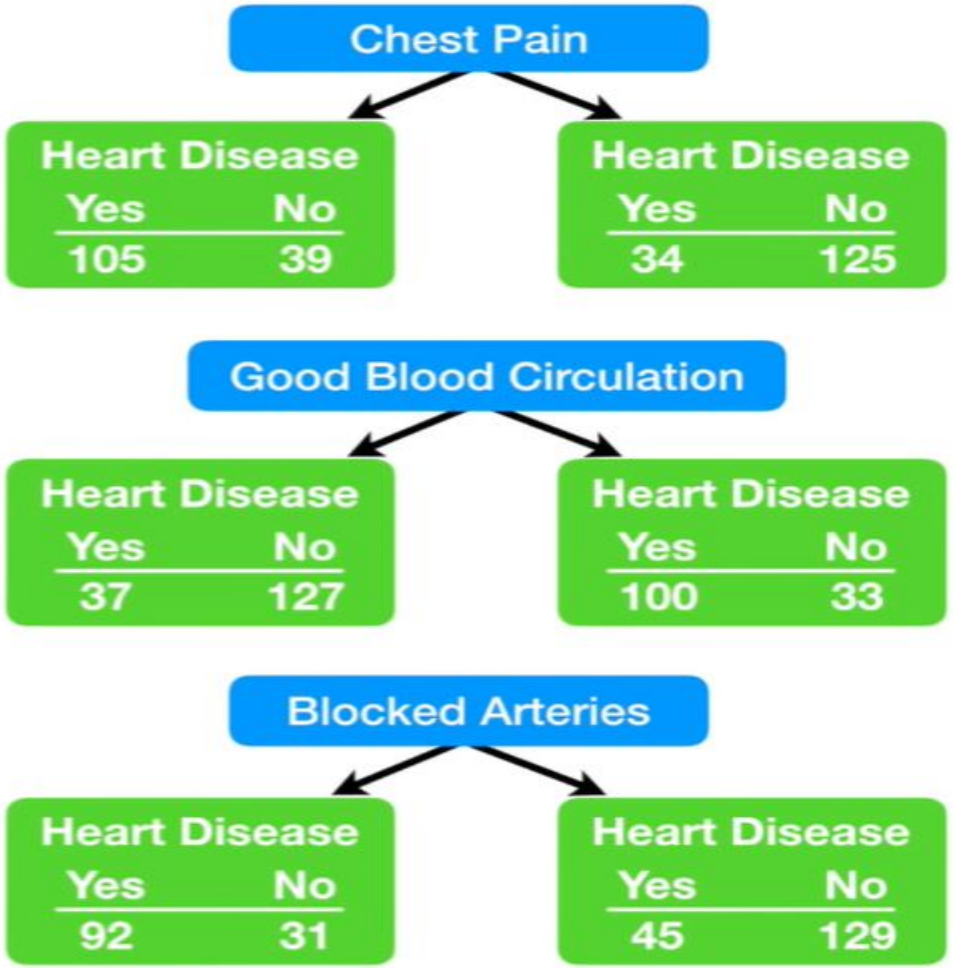
EXAMPLE

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



Ultimately, we look at chest pain and heart disease for all 303 patients in this study.

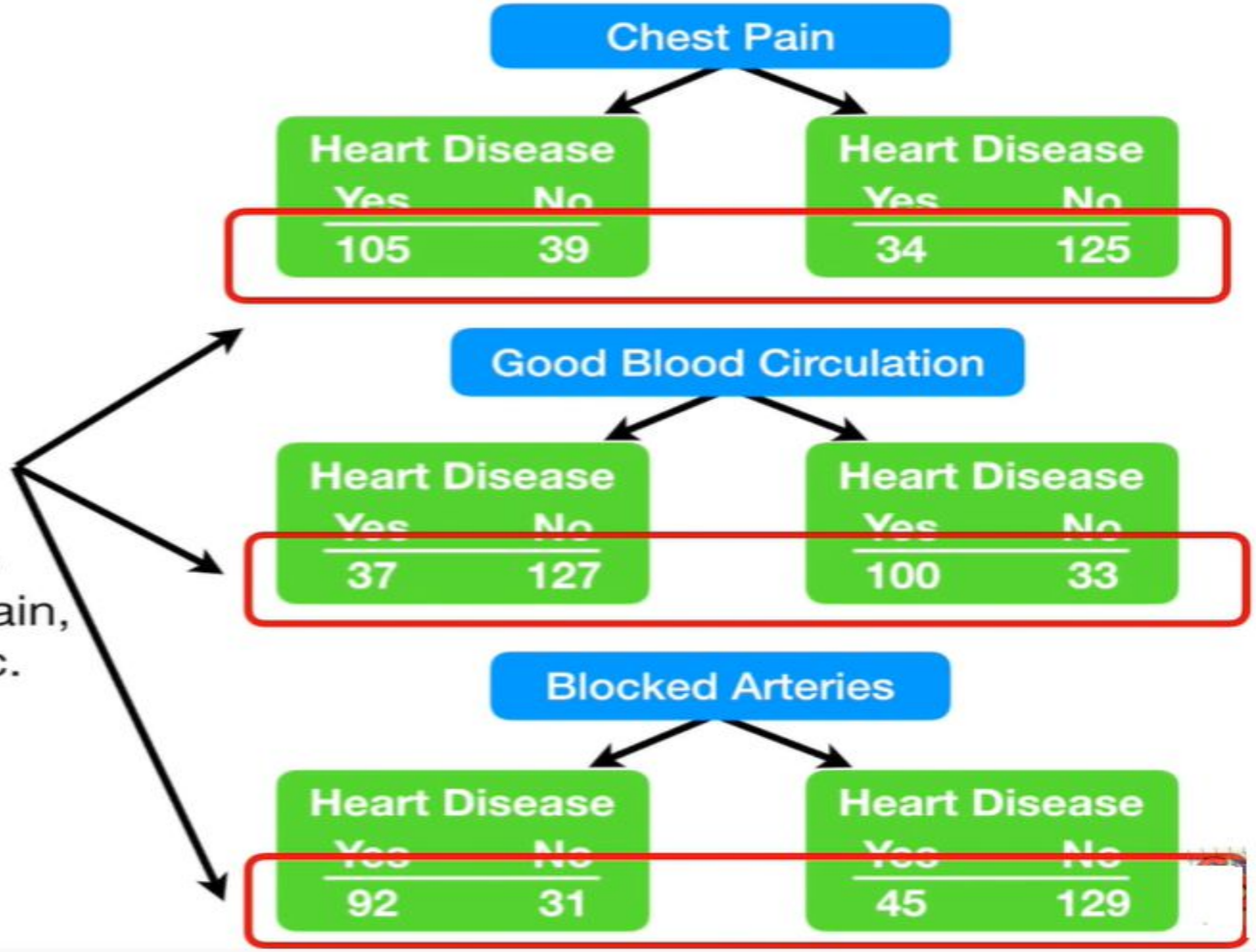
EXAMPLE



Lastly, we looked at how well **Blocked Arteries** separated patients with and without heart disease.

EXAMPLE

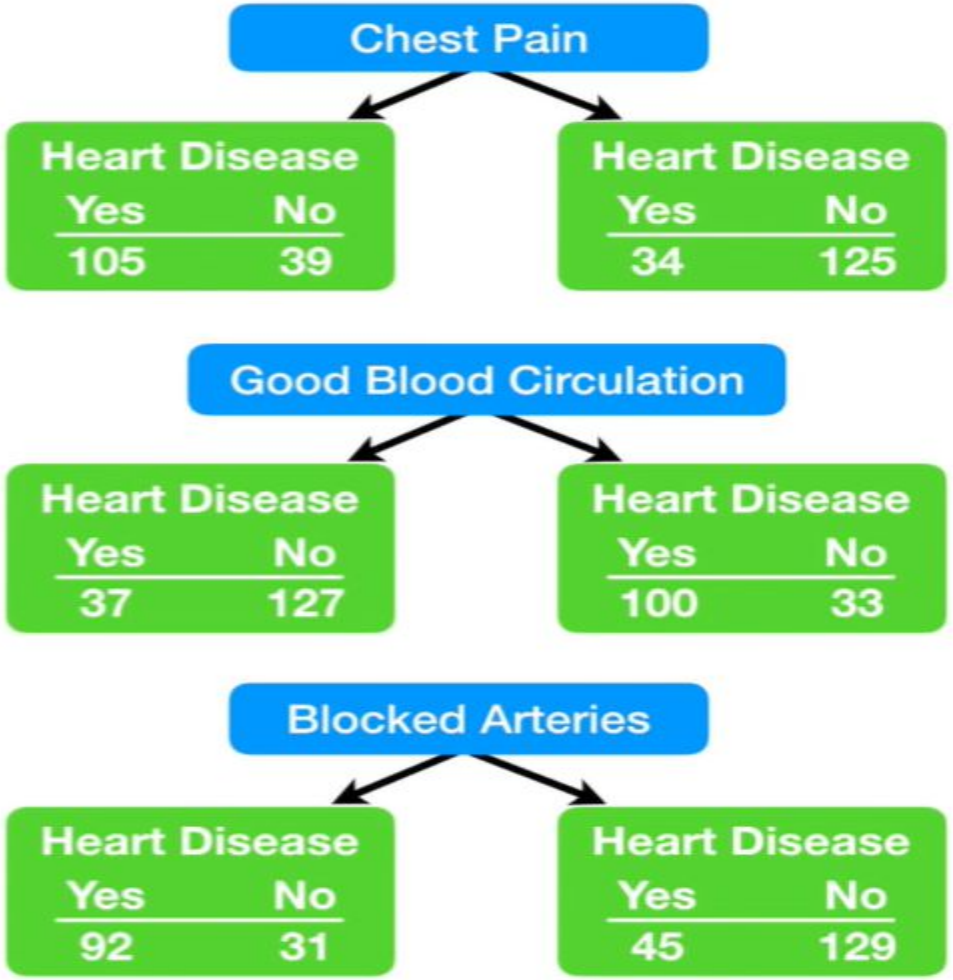
NOTE: The total number of patients with heart disease is different for Chest Pain, Good Blood Circulation and Blocked Arteries because some patients had measurements for Chest Pain, but not for Blocked Arteries, etc.



EXAMPLE

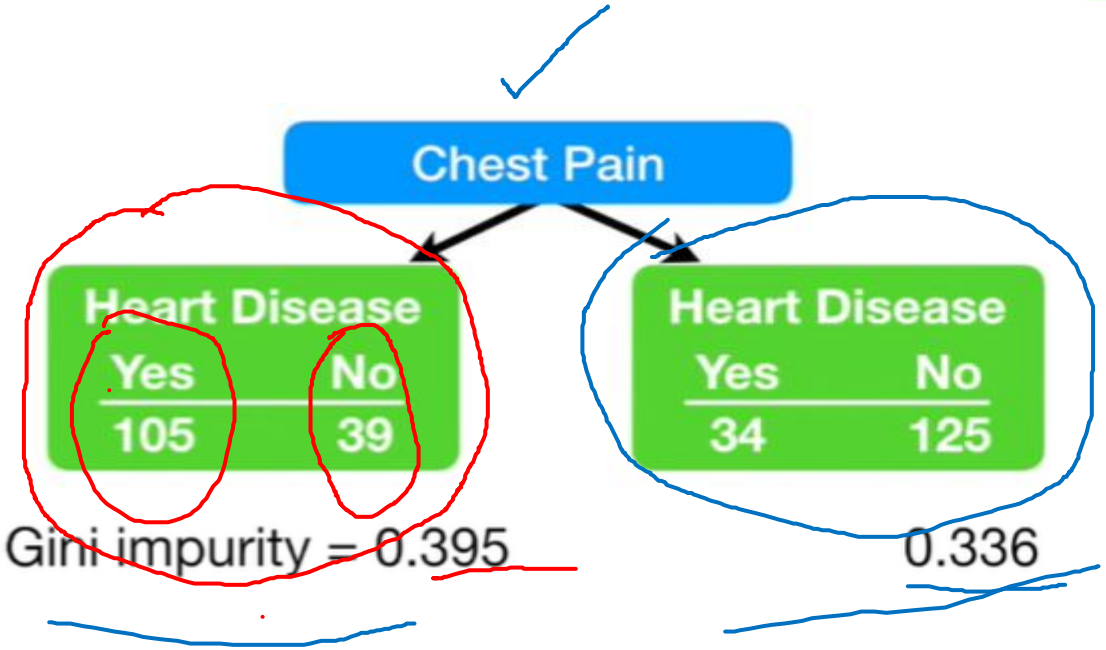
Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

To determine which separation is best, we need a way to measure and compare “**impurity**”.



EXAMPLE

Let's start by calculating Gini impurity for Chest Pain…



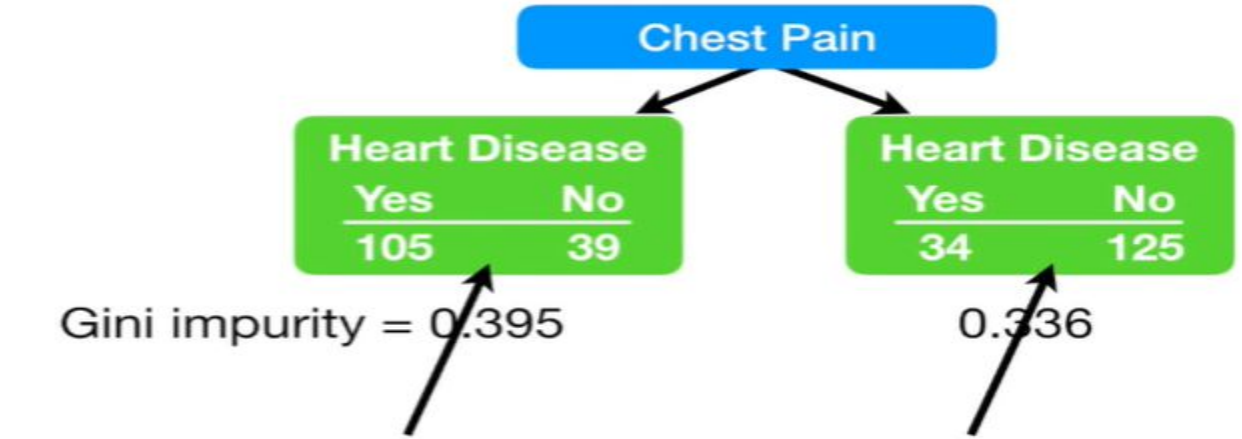
For this leaf, the Gini impurity =

=

= 0.395

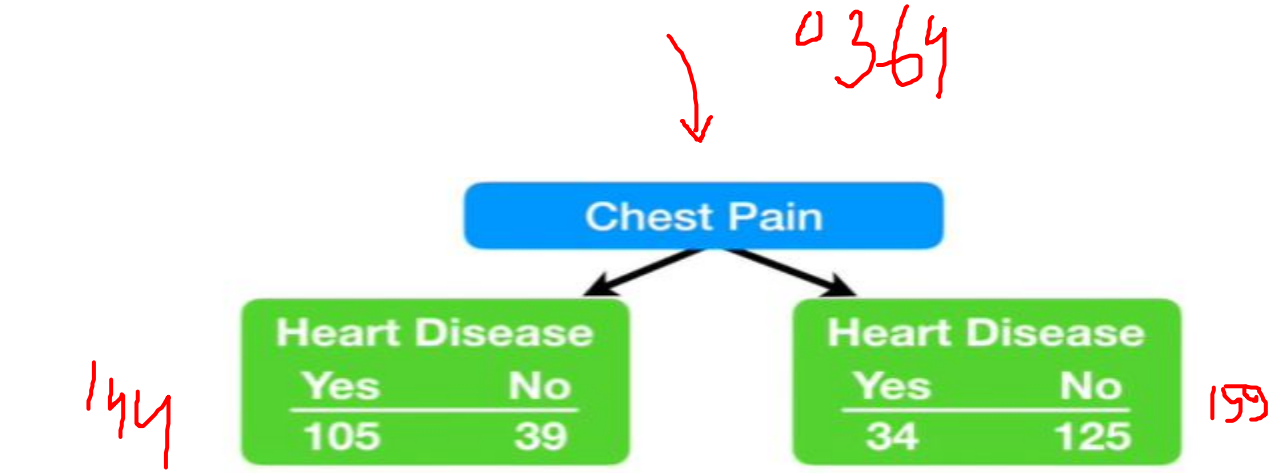
Now that we have measured the Gini impurity for both leaf nodes, we can calculate the total Gini impurity for using Chest Pain to separate patients with and without heart disease.

EXAMPLE



Thus, the total Gini impurity for using Chest Pain to separate patients with and without heart disease is the **weighted average of the leaf node impurities**.

EXAMPLE



Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

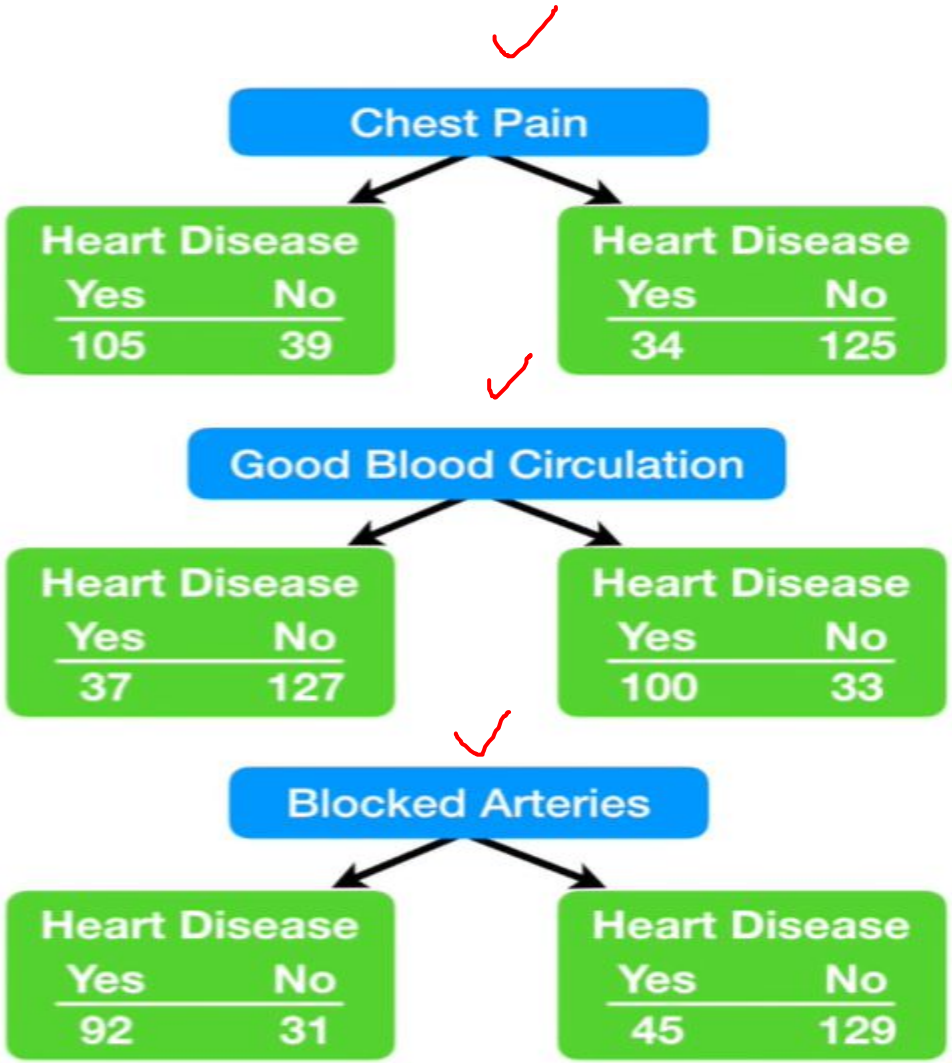
$$= 0.364$$

EXAMPLE

Gini impurity for Chest Pain = 0.364

Gini impurity for Good Blood Circulation = 0.360

Gini impurity for Blocked Arteries = 0.381



EXAMPLE

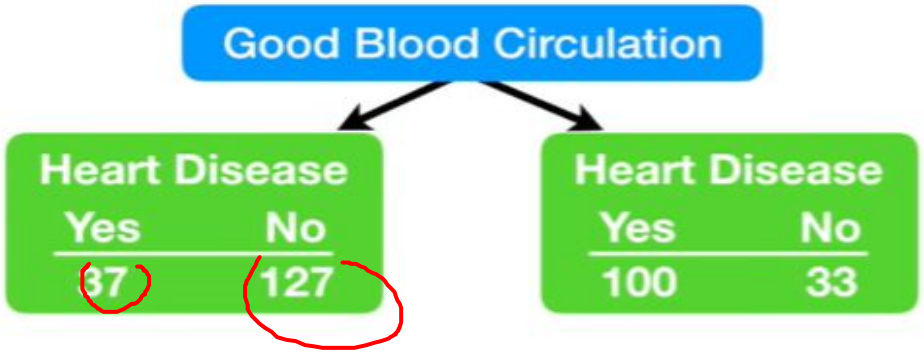
Gini impurity for Chest Pain = 0.364

Gini impurity for Good Blood Circulation = 0.360

Good Blood Circulation has the lowest impurity (it separates patients with and without heart disease the best)…

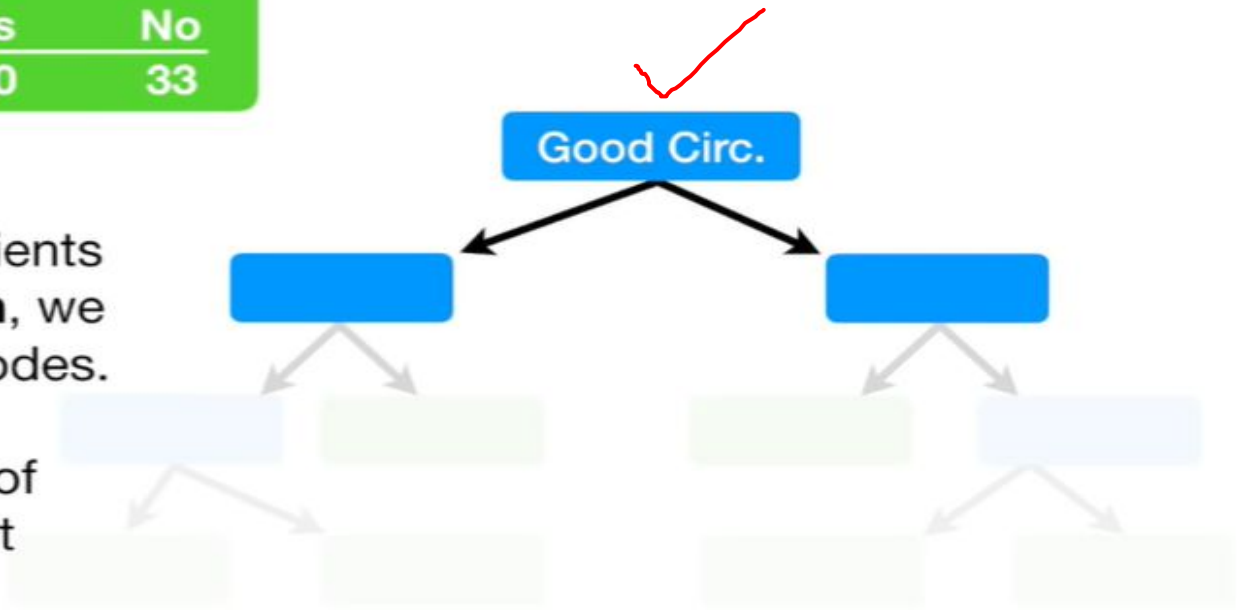
Gini impurity for Blocked Arteries = 0.381

EXAMPLE



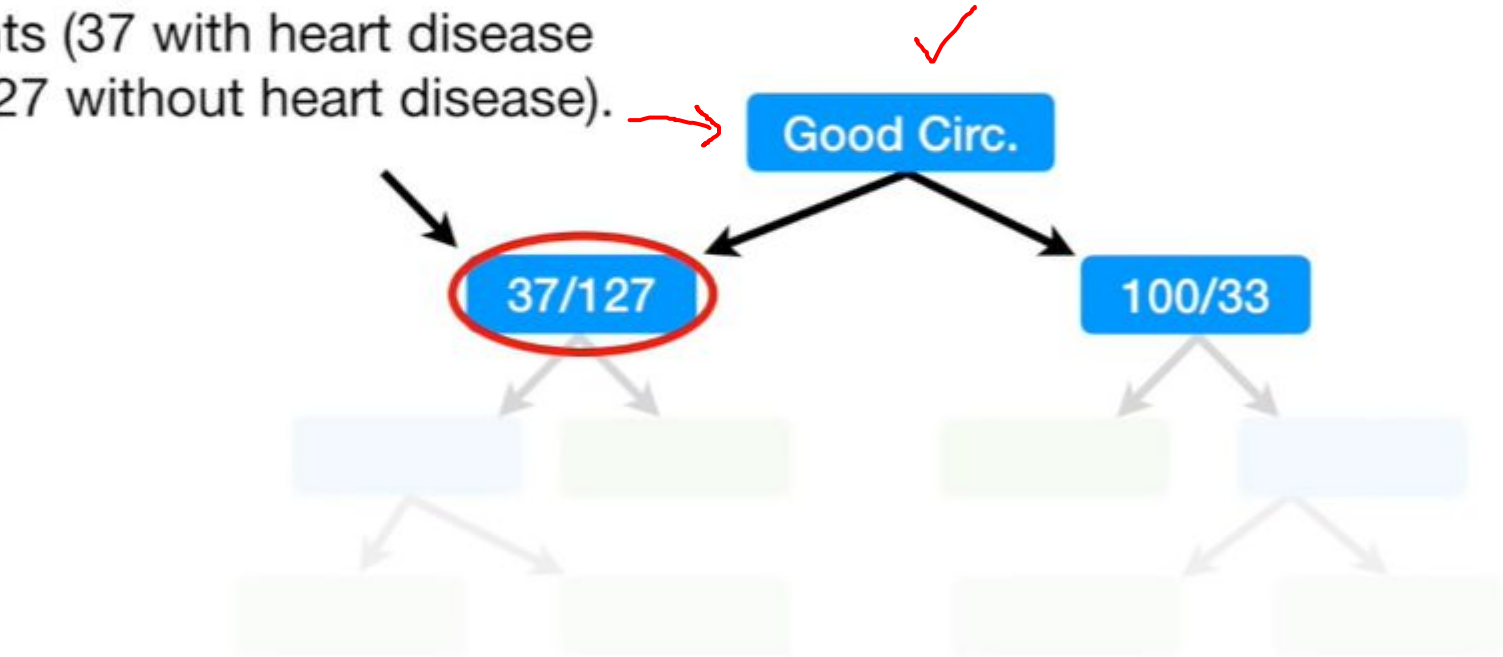
When we divided all of the patients using **Good Blood Circulation**, we ended up with "impure" leaf nodes.

Each leaf contained a mixture of patients with and without Heart Disease.

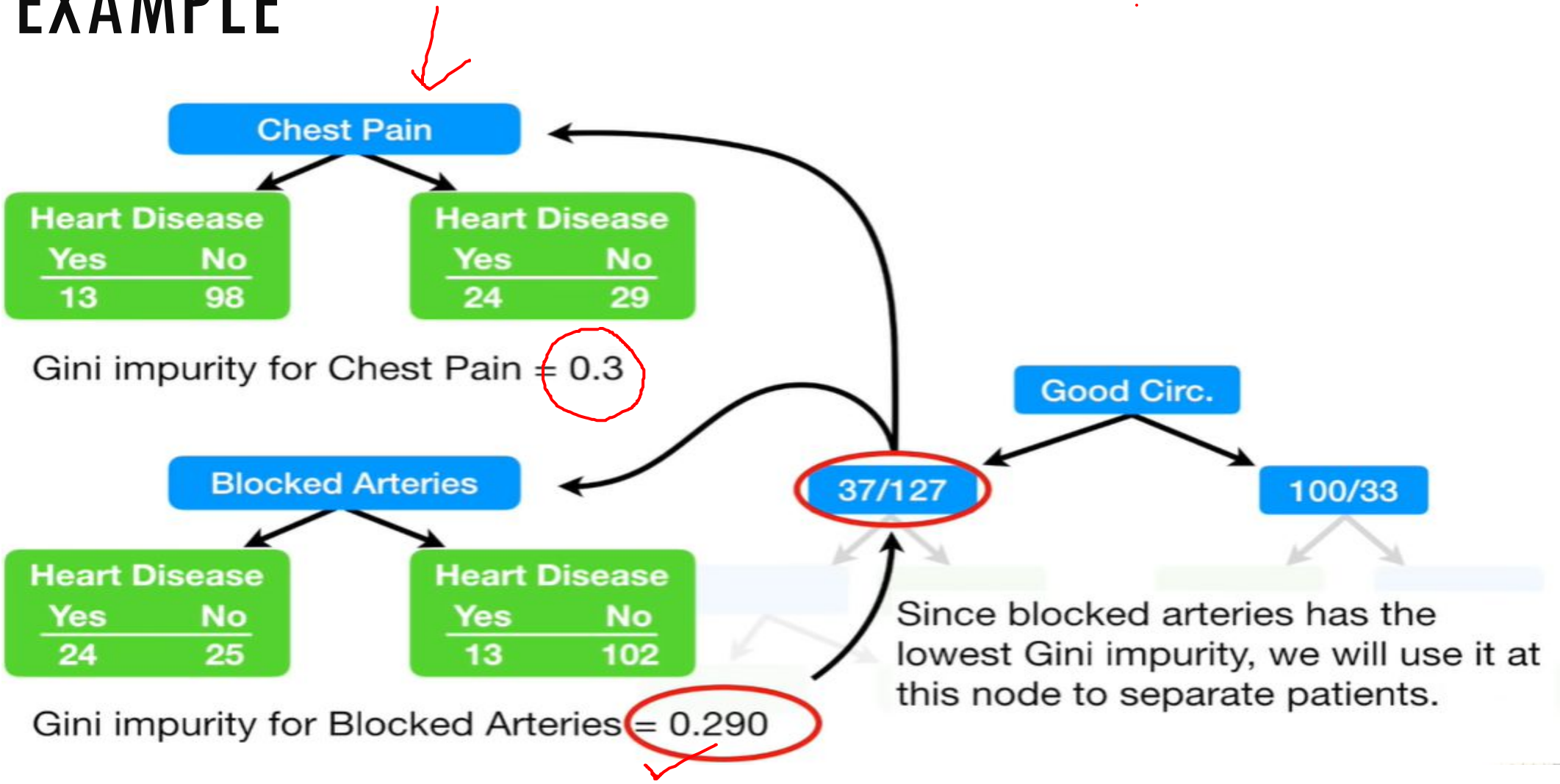


EXAMPLE

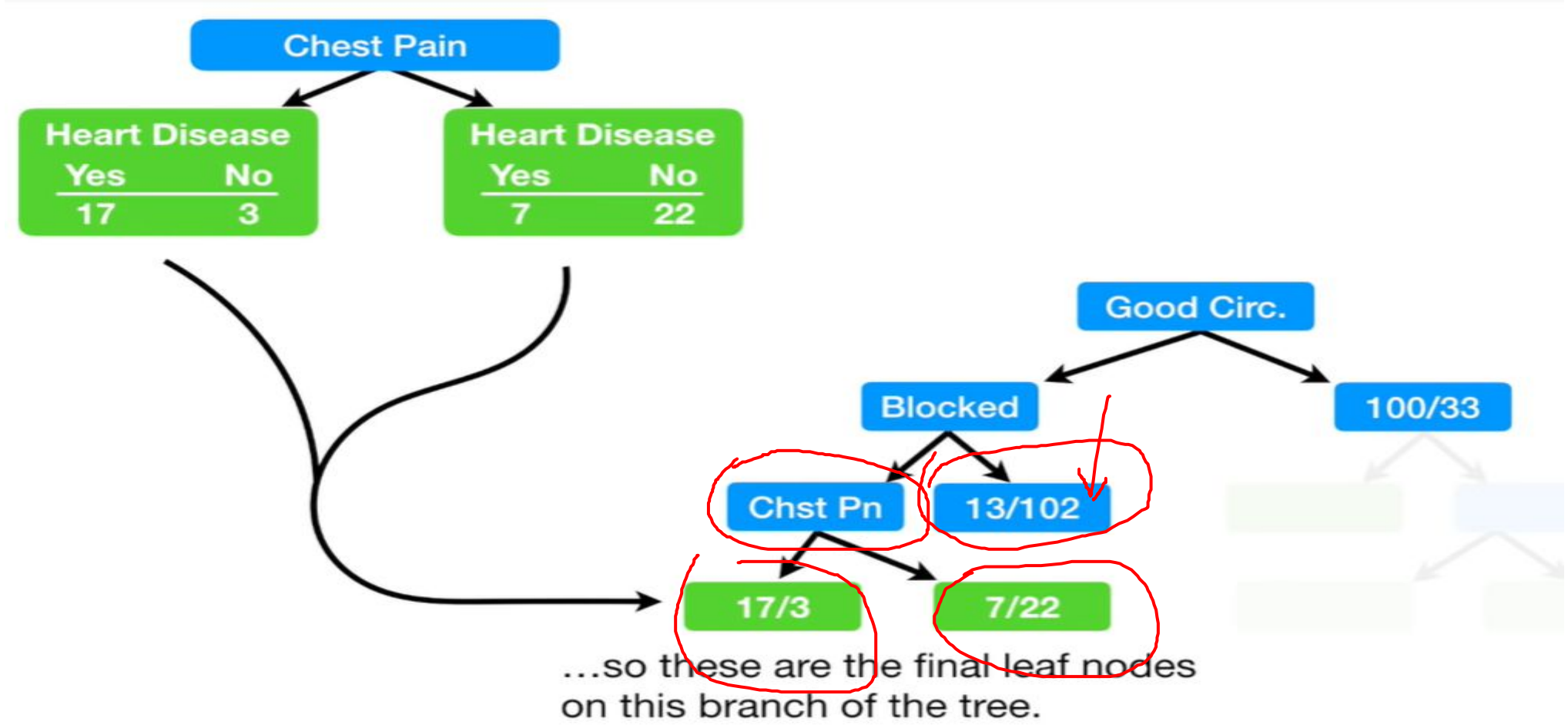
Now we need to figure how well **chest pain** and **blocked arteries** separate these 164 patients (37 with heart disease and 127 without heart disease).



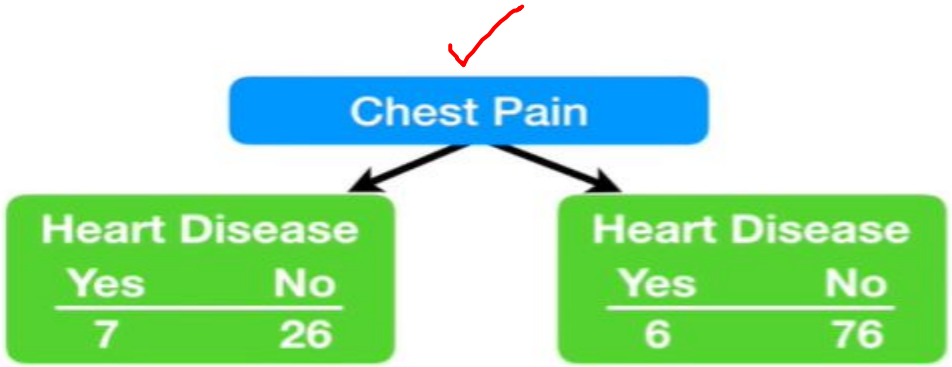
EXAMPLE


Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Gini impurity for Blocked Arteries = 0.290
Good Circ.
37/127
100/33
Since blocked arteries has the lowest Gini impurity, we will use it at this node to separate patients.

EXAMPLE


Chest Pain
Heart Disease
Yes No
17 3
Heart Disease
Yes No
7 22
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22
...so these are the final leaf nodes
on this branch of the tree.

EXAMPLE



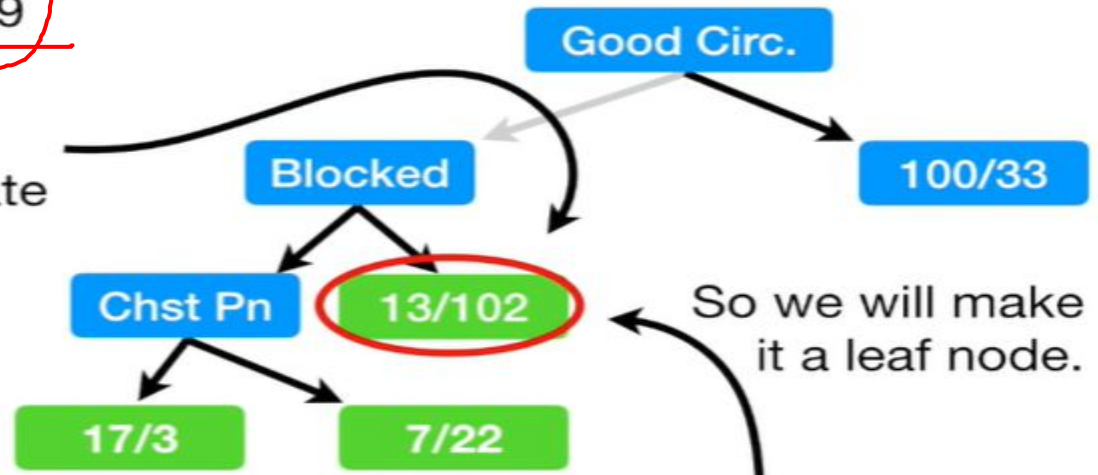
Gini impurity for Chest Pain = 0.29

The Gini impurity for this node, before using chest pain to separate patients is...

= 1 - (the probability of "yes")^2 - (the probability of "no")^2

$$= 1 - \left(\frac{13}{13+102}\right)^2 - \left(\frac{102}{13+102}\right)^2$$

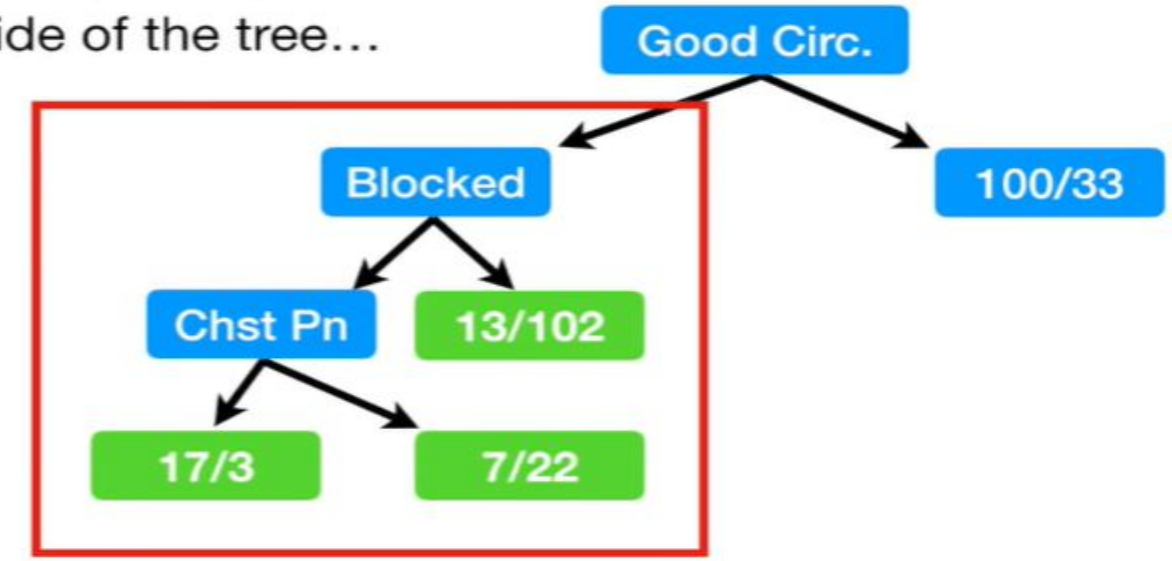
= 0.2



So we will make it a leaf node.

EXAMPLE

OK, at this point we've worked out the entire left side of the tree…



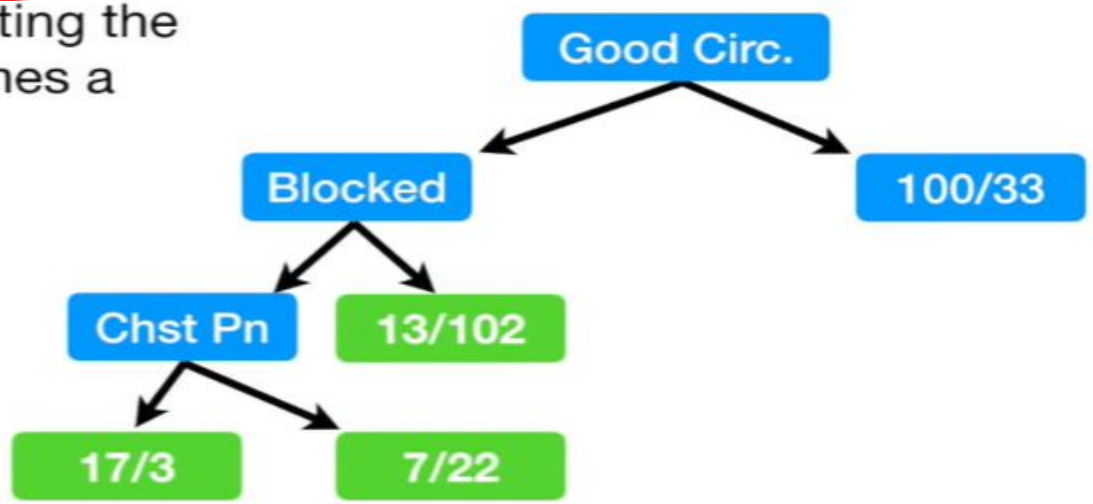
EXAMPLE

The good news is that we follow the exact same steps as we did on the left side:

1) Calculate all of the Gini impurity scores.

2) If the node itself has the lowest score, than there is no point in separating the patients any more and it becomes a leaf node.

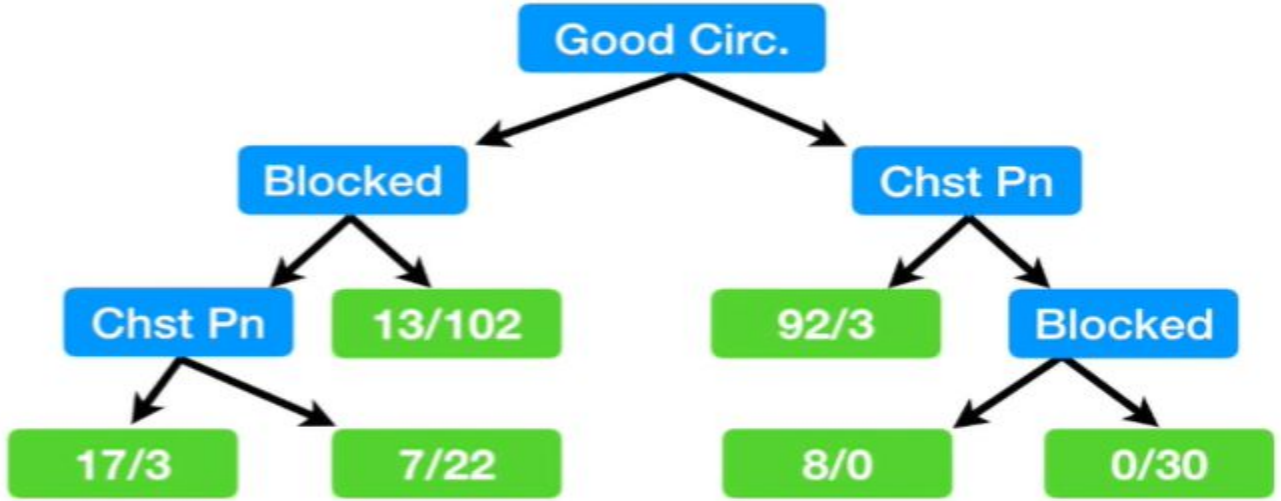
3) If separating the data results in an improvement, than pick the separation with the lowest impurity value.



EXAMPLE

So far we've seen how to build a tree with "yes/no" questions at each step…

…but what if we have numeric data, like patient weight?



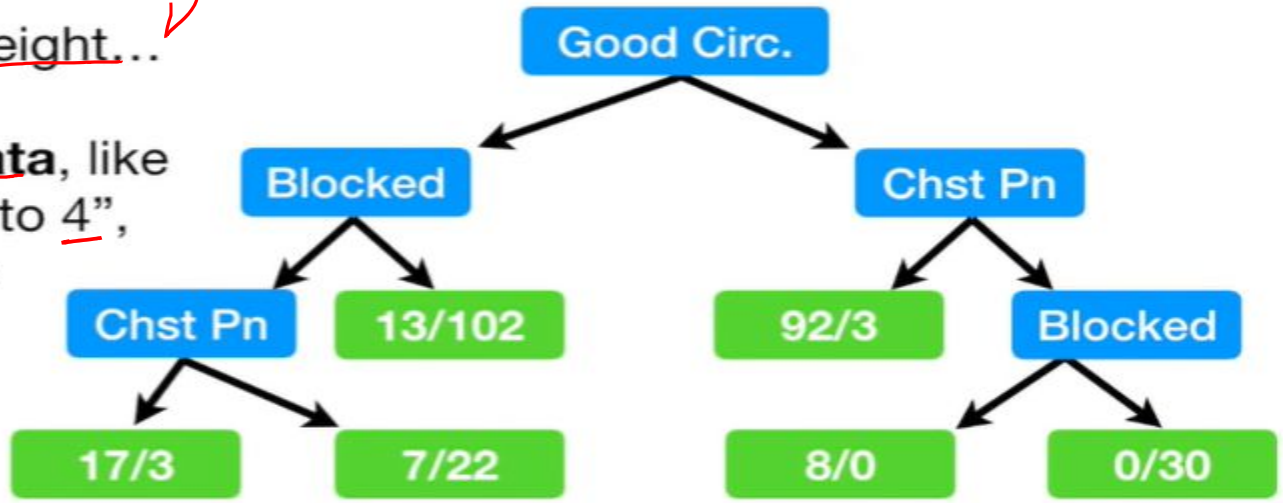
EXAMPLE

Now we've seen how to build a tree with…

1) "yes/no" questions at each step…

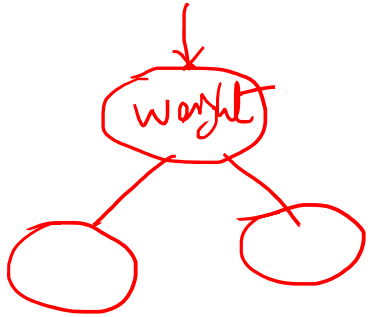
2) Numeric data, like patient weight…

Now let's talk about **ranked data**, like "rank my jokes on a scale of 1 to 4", and **multiple choice data**, like "which color do you like, red, blue or green?"



EXAMPLE

Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No



How do we determine what's the best weight to use to divide the patients?

EXAMPLE

	Weight	Heart Disease
Lowest	155	No
	180	Yes
	190	No
	220	Yes
Highest	225	Yes

Step 1) Sort the patients by weight, lowest to highest.

EXAMPLE

Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

(155+180)/2

Step 2) Calculate the average weight for all adjacent patients.

EXAMPLE

Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

Step 3) Calculate the impurity values for each average weight.

167.5 → Gini impurity = ?

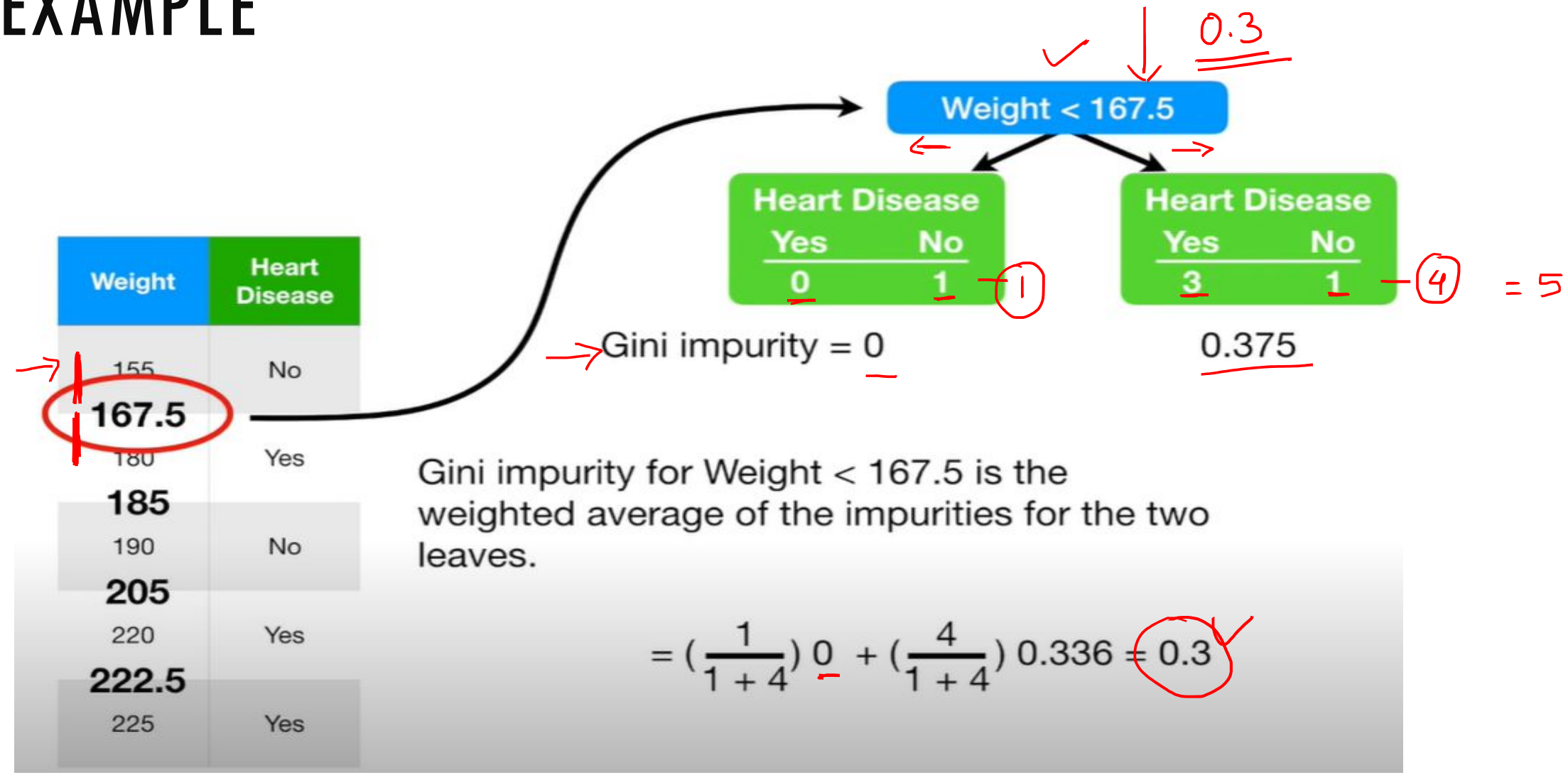
185 → Gini impurity = ?

205 → Gini impurity = ?

222.5 → Gini impurity = ?

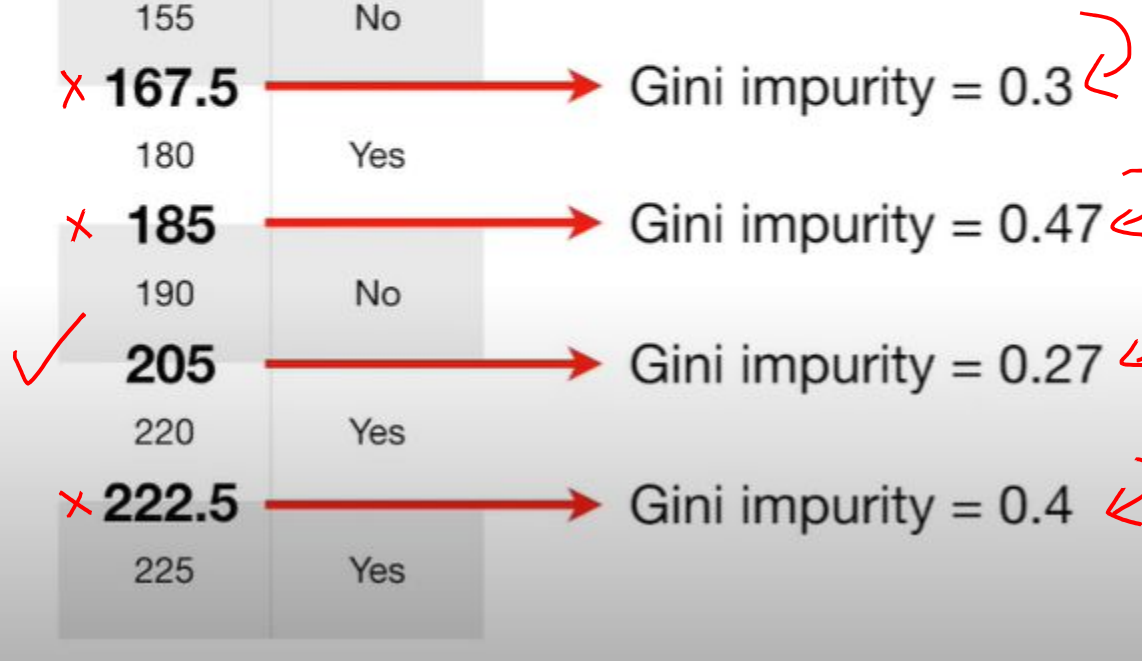
EXAMPLE

Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes



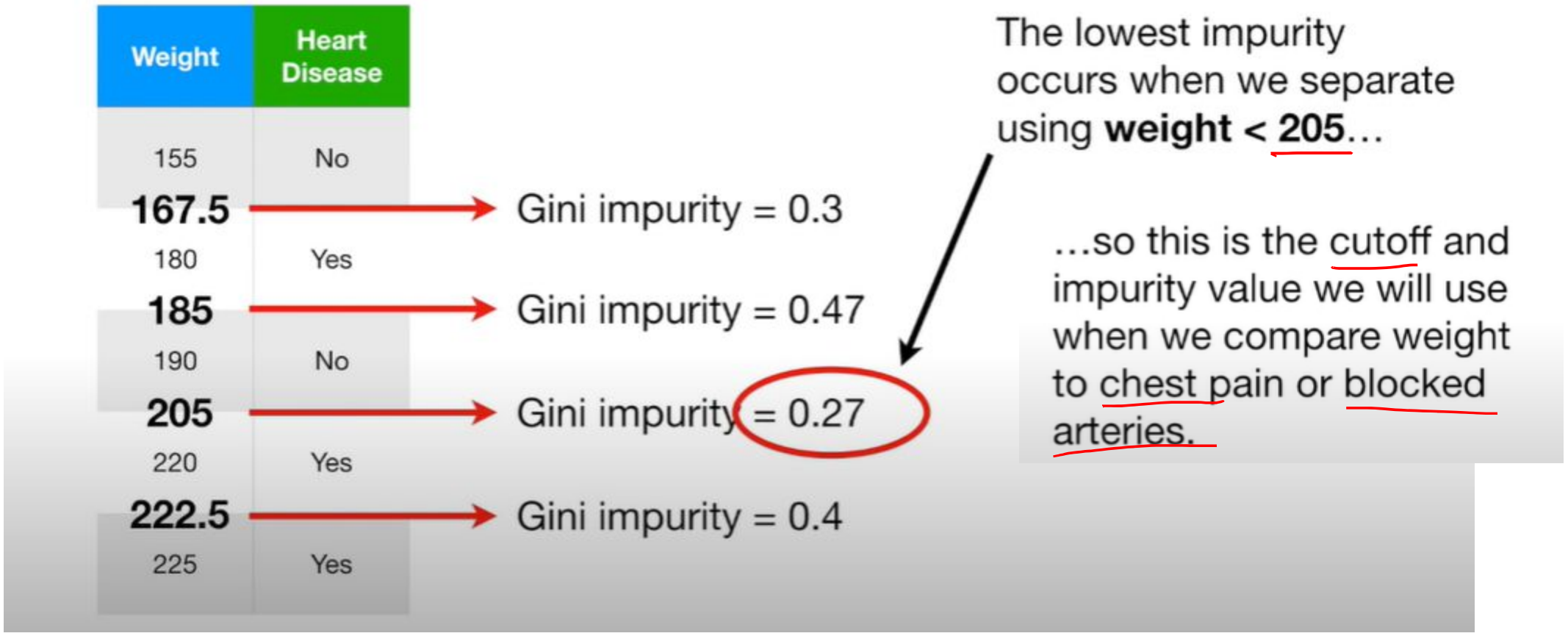
EXAMPLE

Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes



Best → least Gini values

EXAMPLE


Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes
The lowest impurity occurs when we separate using weight < 205…
…so this is the cutoff and impurity value we will use when we compare weight to chest pain or blocked arteries.

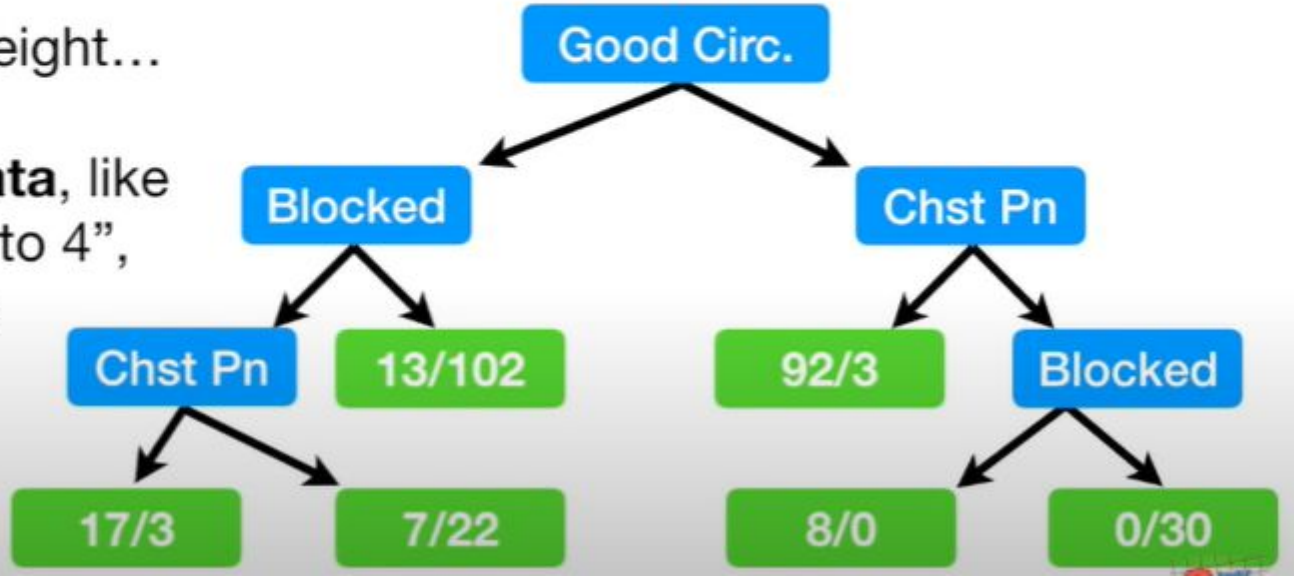
EXAMPLE

Now we've seen how to build a tree with…

1) "yes/no" questions at each step…

2) Numeric data, like patient weight…

Now let's talk about **ranked data**, like "rank my jokes on a scale of 1 to 4", and **multiple choice data**, like "which color do you like, red, blue or green?"



EXAMPLE

Rank my jokes...	Likes StatQuest
1	Yes
1	No
3	Yes
1	Yes
etc...	etc...

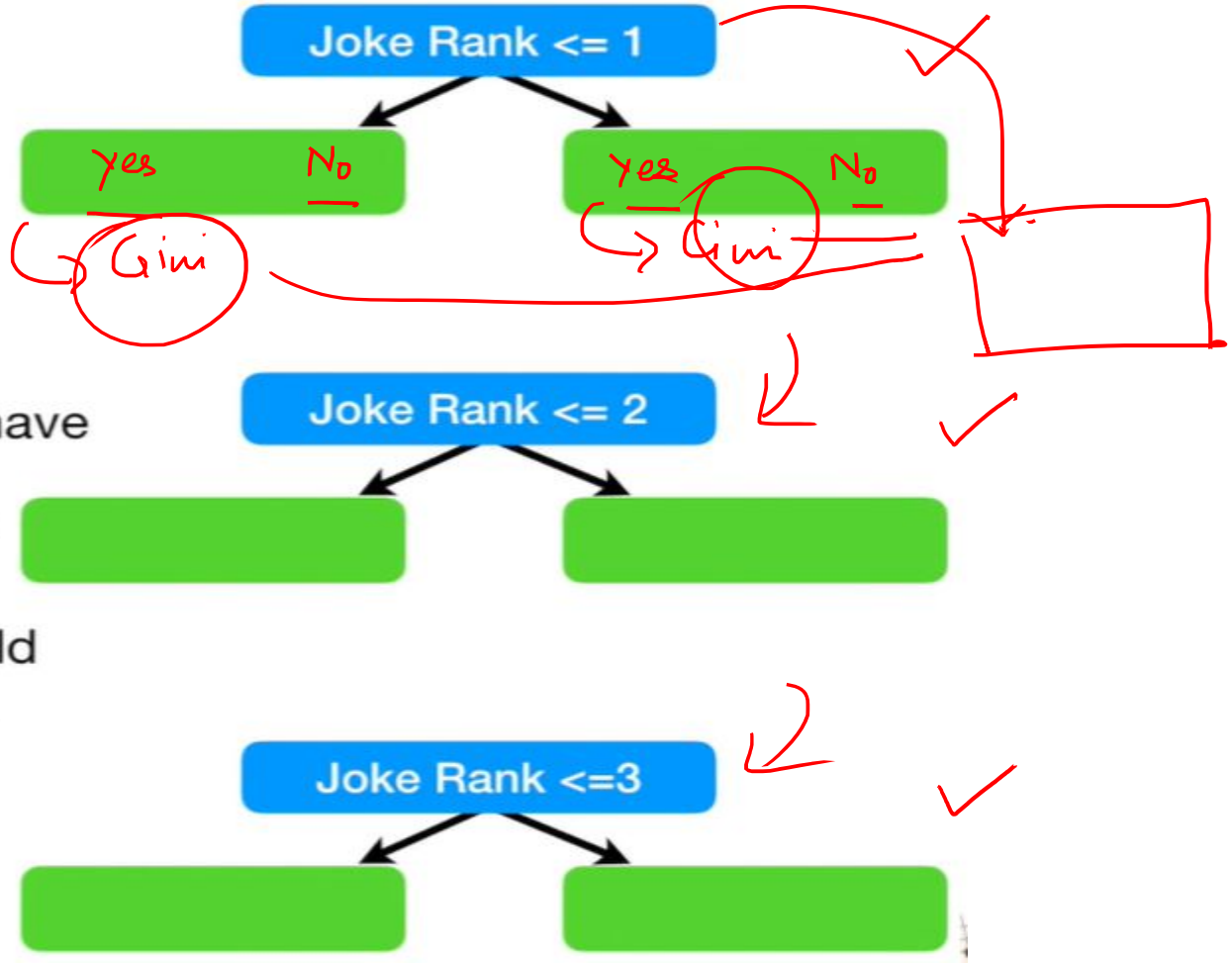
Ranked data is similar to numeric data, except instead now we calculate impurity scores for all of the possible ranks.

So if people could rank my jokes from 1 to 4 (4 being the funniest), we could calculate the following impurity scores...

EXAMPLE

Rank my jokes...	Likes StatQuest
1	Yes
1	No
3	Yes
1	Yes
etc...	etc...

NOTE: We don't have to calculate an impurity score for Joke Rank <= 4 because that would include everyone.



EXAMPLE

Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes
etc...	etc...

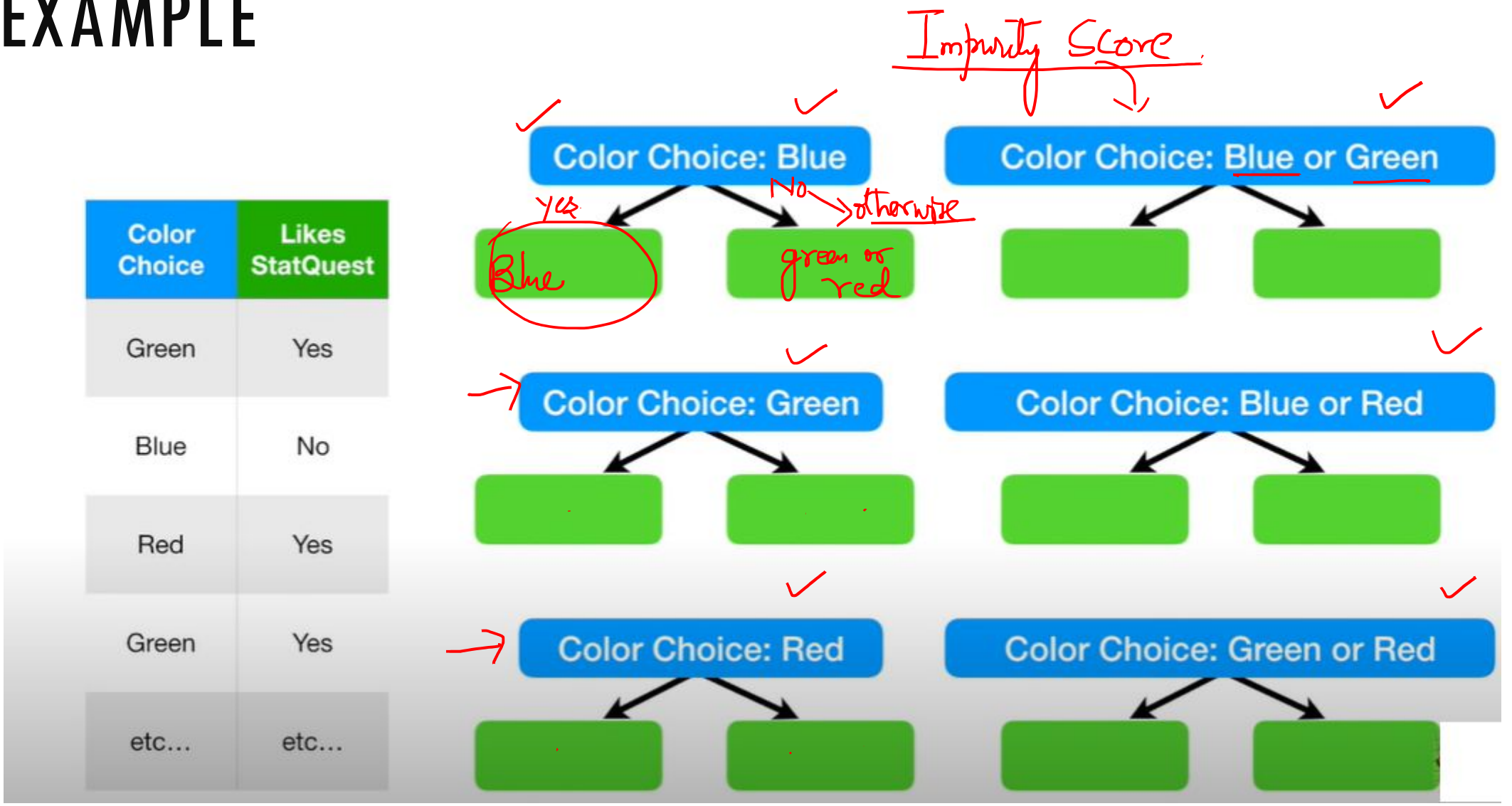
When there are **multiple choices**, like "**color choice can be blue, green or red**", you calculate an impurity score for each one as well as each possible combination.

Blue —
Green —
Red —

(Blue, Green) (Red)
(Green, Red) (Blue)
(Red, Blue) (Green)

EXAMPLE

Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes
etc...	etc...



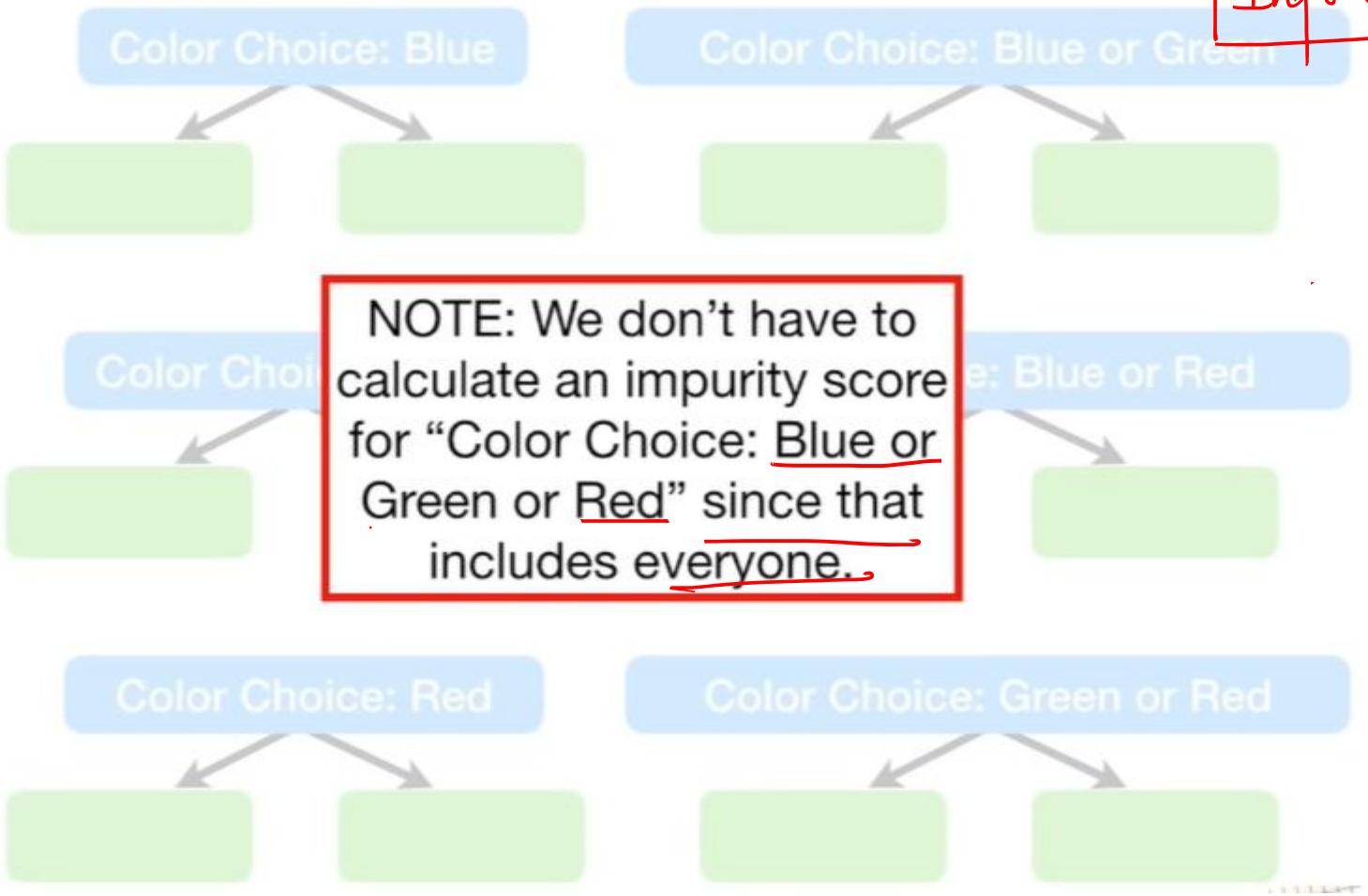
EXAMPLE

Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes
etc…	etc…

ASM

Gini Index
Entropy

Info Gain



THANKS


Keep Learning
Keep Growing


Dr. Neeraj Gupta
Assistant Professor, Dept. of CEA
neeraj.gupta@gla.ac.in