AdaBoost

By Krunal Gandhare

In Random forest we have number of trees formed by choosing random variable and the trees formed have branches and leaves

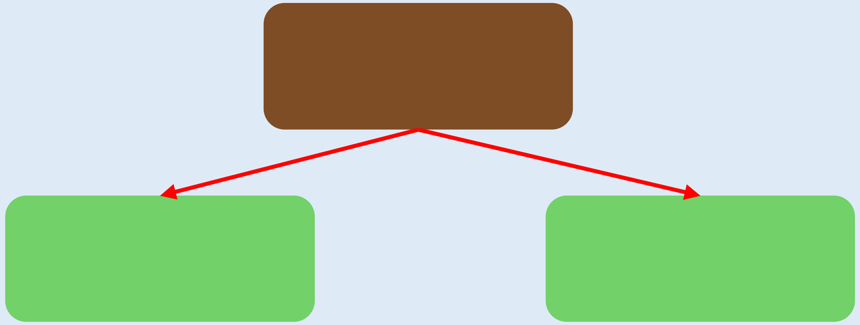
In AdaBoost, we don’t have full grown trees but root node and two leaves only

If there are root nodes and two leaves only, it will be called stump, a decision stump

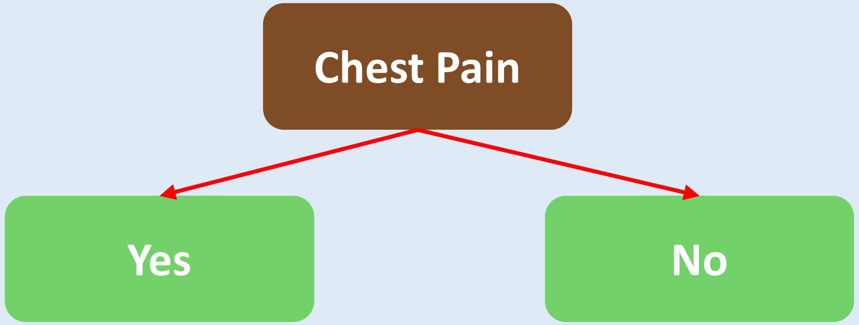
In Random forest we have number of trees formed by choosing random variable and the trees formed have branches and leaves

In AdaBoost, we don't have full grown trees but root node and two leaves only

If there are root nodes and two leaves only, it will be called stump, a decision stump



Chest Pain	Good Blood Circulation	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

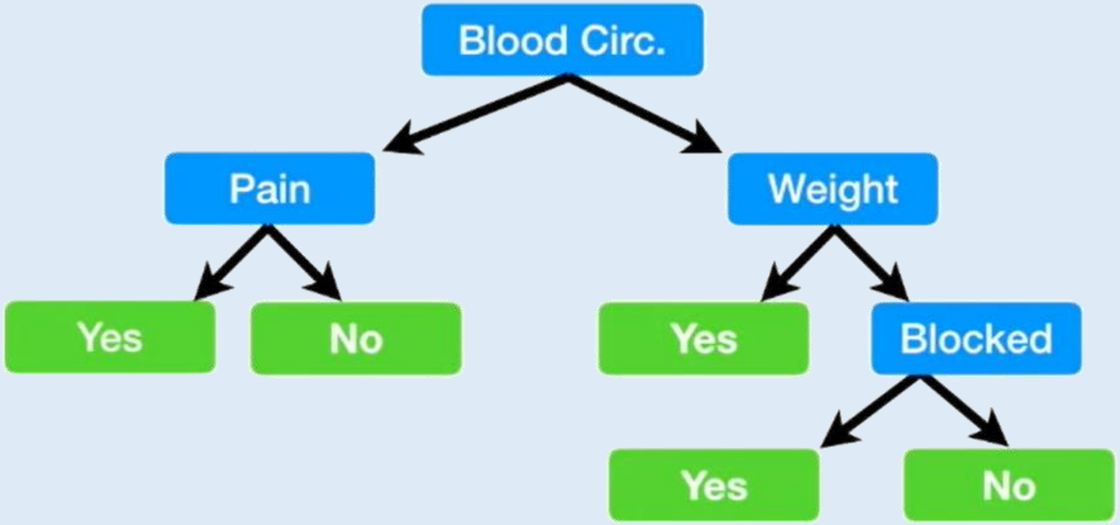


Chest Pain

Yes

No

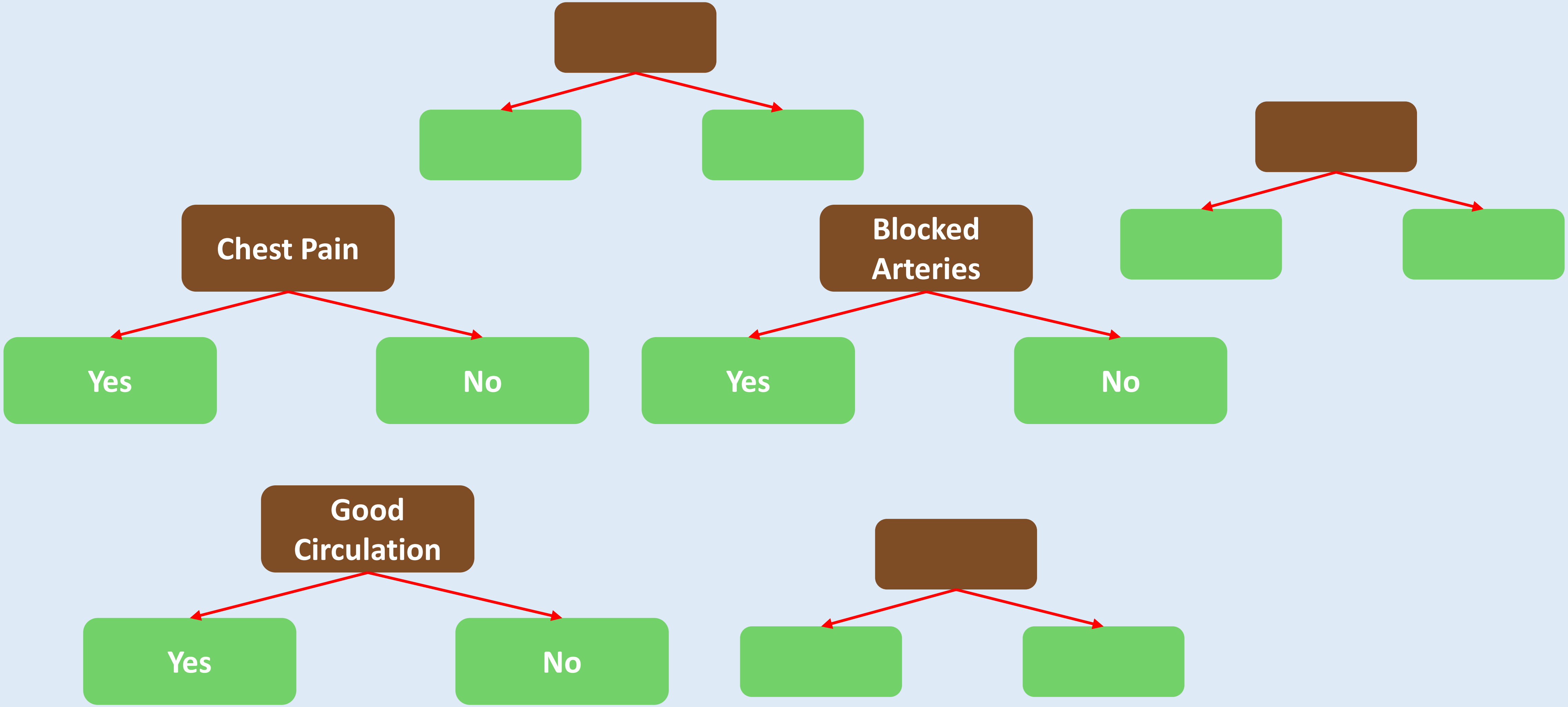
Full grown decision trees are likely to give right result



But stumps fail to give right output

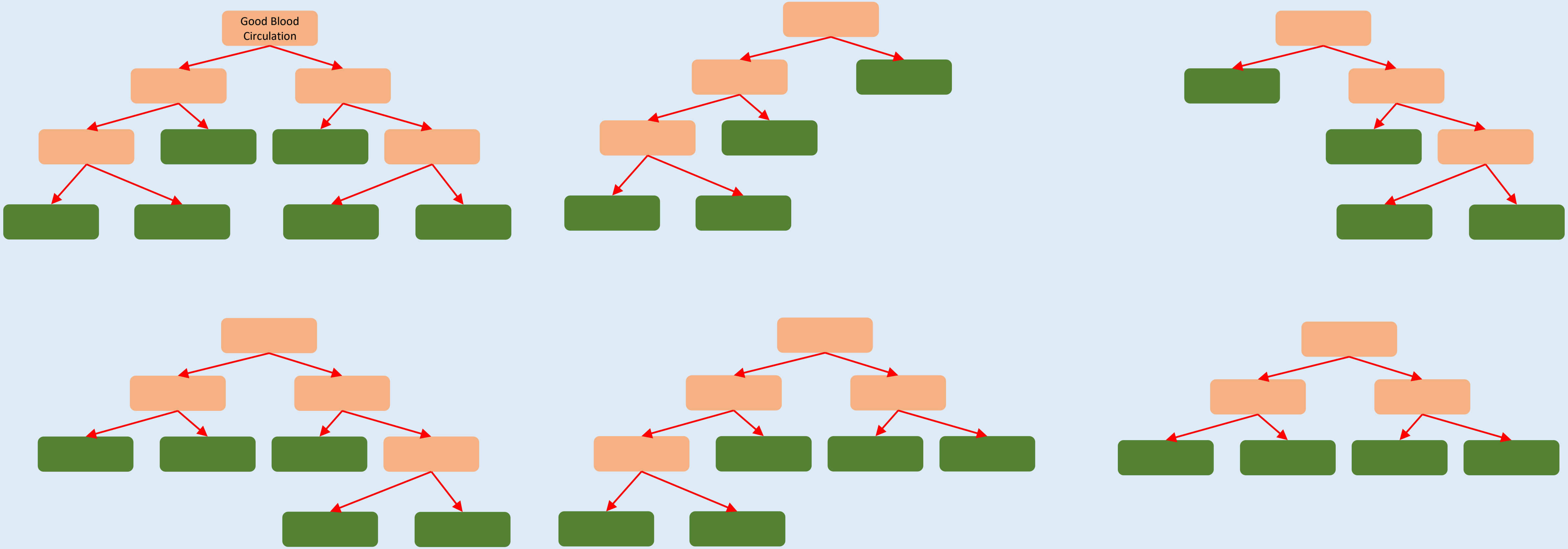
Hence stumps are called weak learners

AdaBoost combines such many stumps i.e. many weak learners to give right output

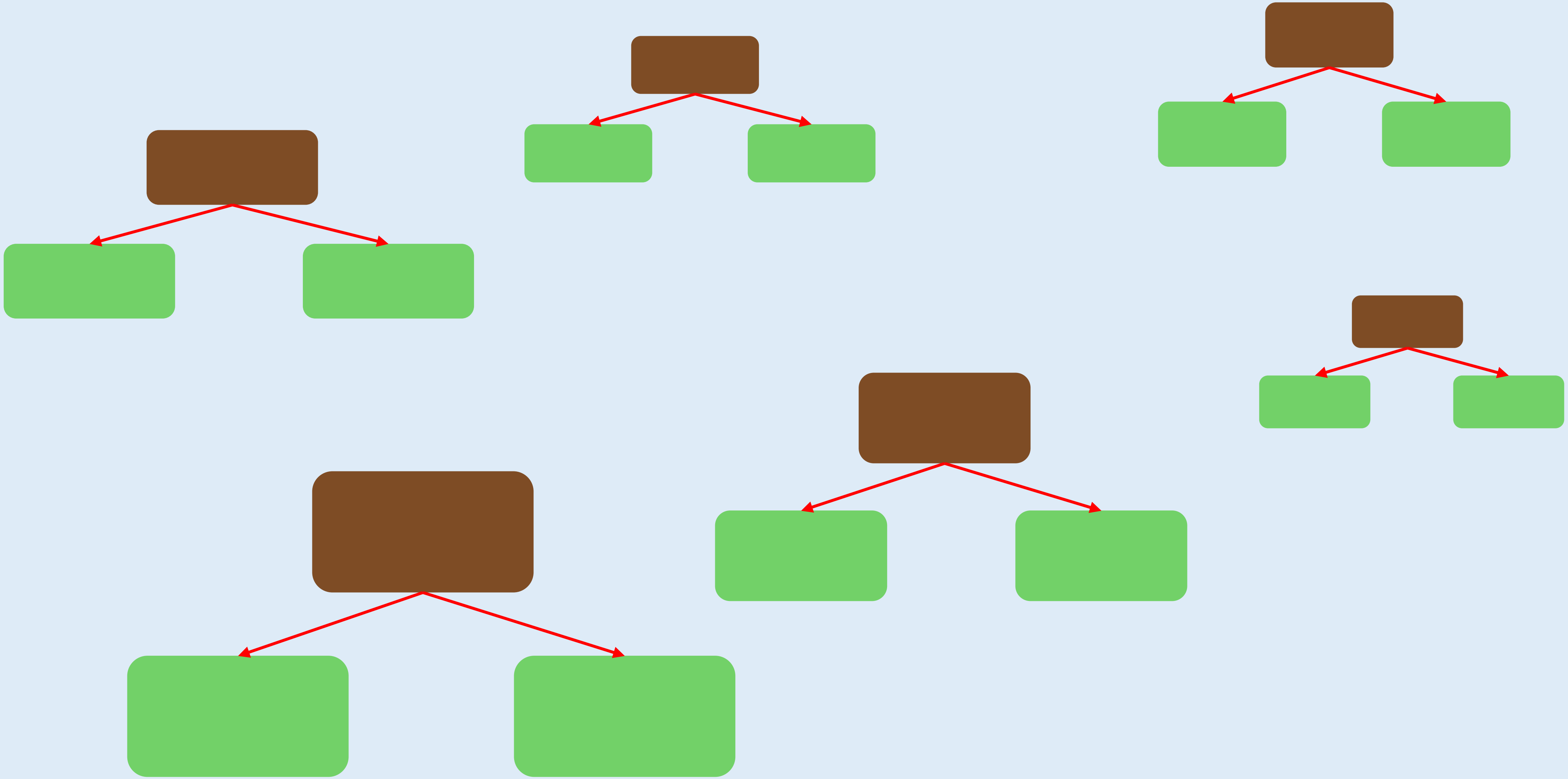


In random forest, for a classification problem each tree in random forest has equal weightage

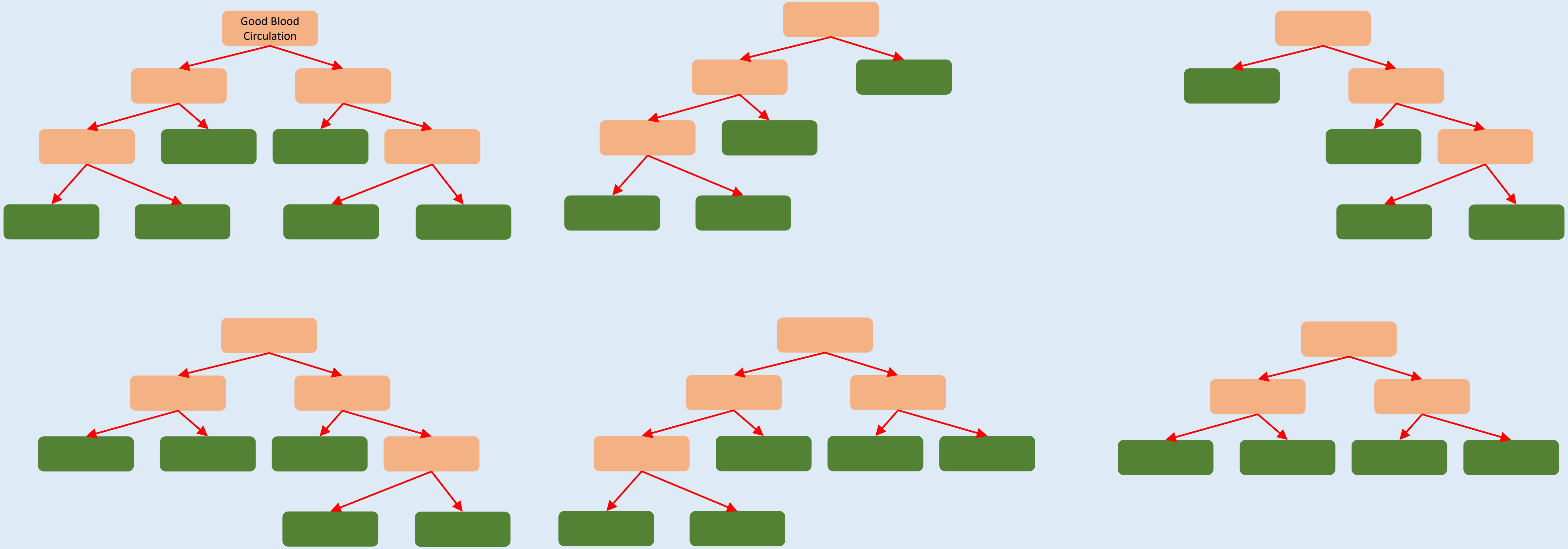
But in case of AdaBoost, some stumps have higher weightage as compered to others for solving classification problems



But in case of AdaBoost, some stumps have higher weightage as compered to others for solving classification problems



In random forest it does not matter which tree was made when, that is order of formation of tree is irrelevant

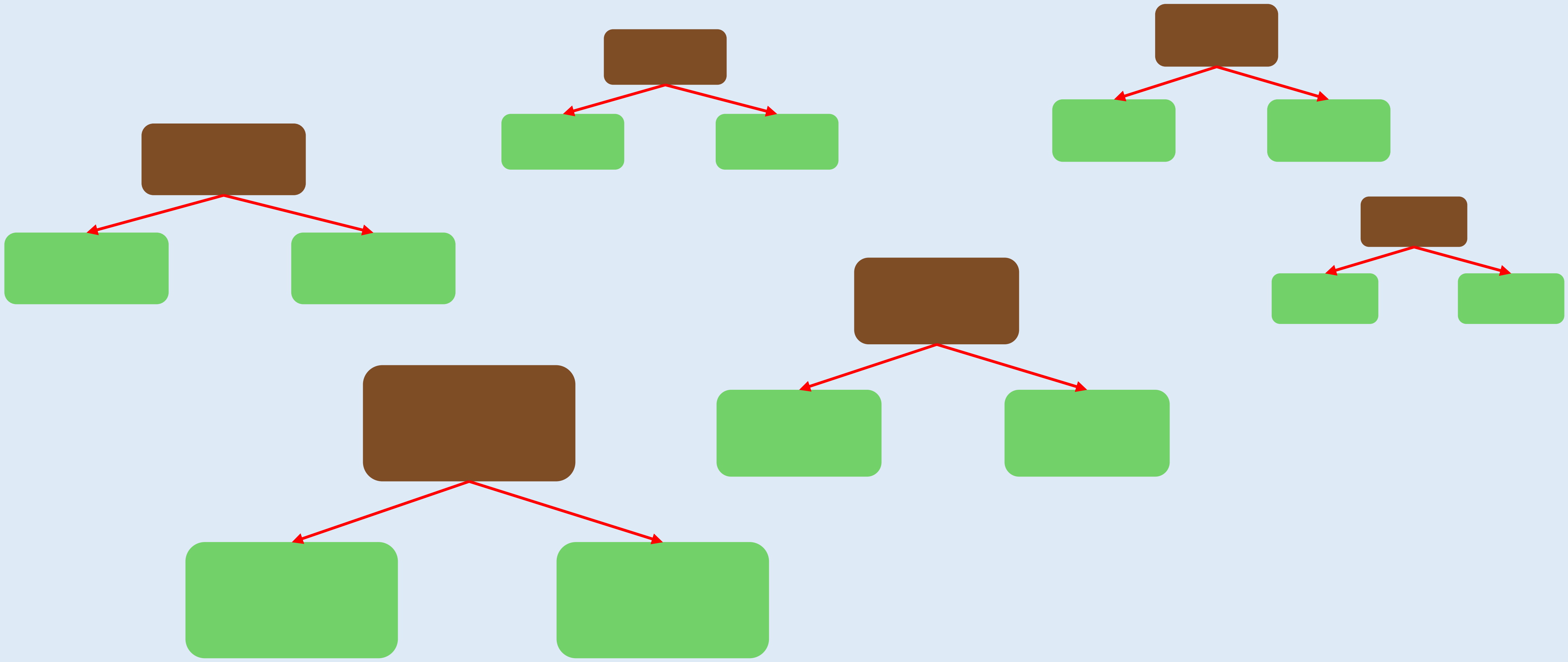


But in case of AdaBoost, the order of formation of trees matters a lot

Here in Adaboost, the error first stump makes while giving output will decide how the second stump is made

And the error made by second stump will decide how the third stump is made

Same way the formation of rest of the stumps is influenced by the error made buy it’s predecessor



To summaries AdaBoost….

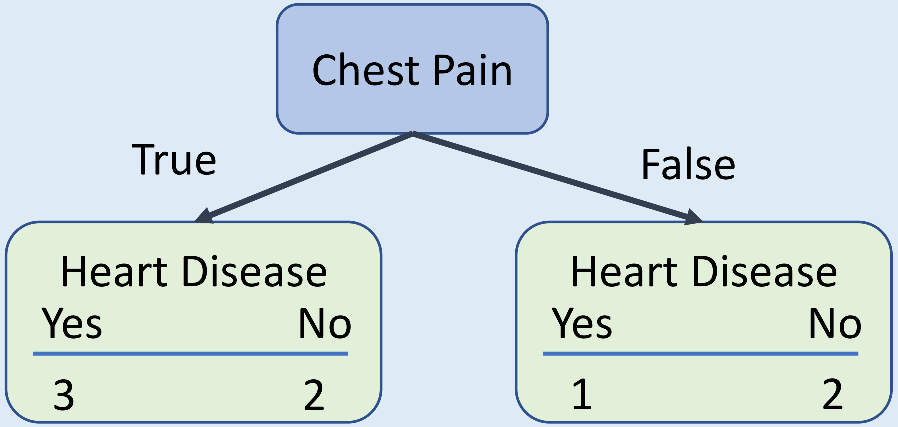
1. Adaboost combines many weak learners, also called as stumps or decision stump
2. Some stumps have more weightage than other in giving the final output
3. Each stump is made by taking the error made by it’s predecessor into account

Chest Pain	Blocked Arteries	Weight	Heart Disease
Yes	Yes	205	Yes
No	Yes	180	Yes
Yes	No	210	Yes
Yes	Yes	167	Yes
No	Yes	156	No
No	Yes	125	No
Yes	No	168	No
Yes	Yes	172	No

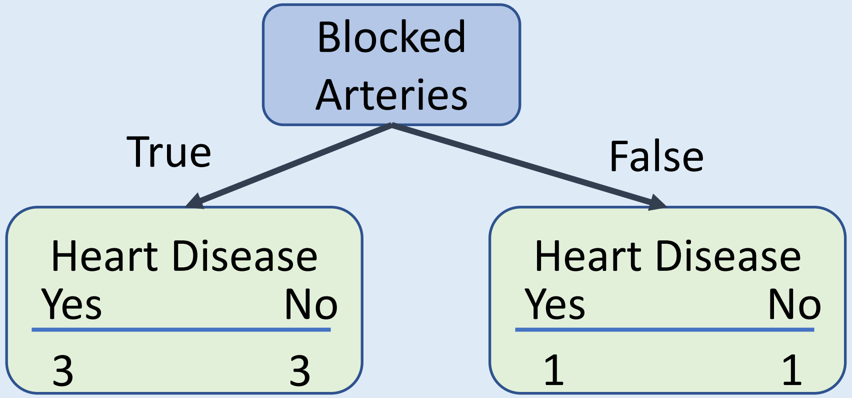
Sample Weight
1/8
1/8
1/8
1/8
1/8
1/8
1/8
1/8

$$\frac{1}{Total\ Number\ of\ Samples} = \frac{1}{8}$$

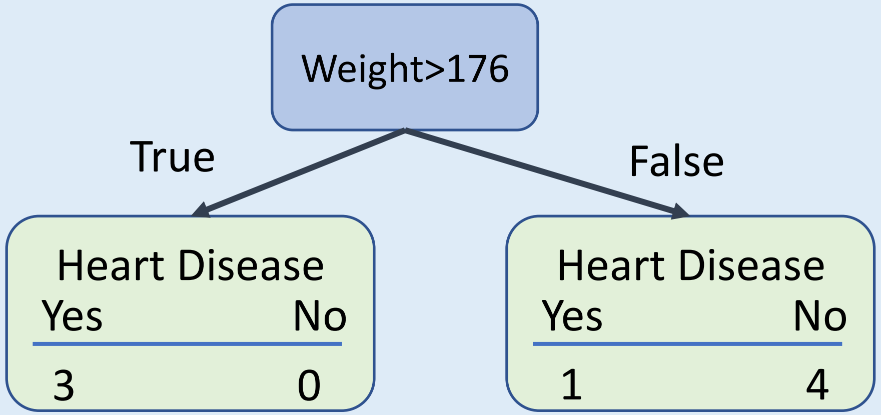
Chest Pain	**Blocked Arteries**	**Weight**	**Heart Disease**	**Sample Weight**
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8



Gini Index = 0.47

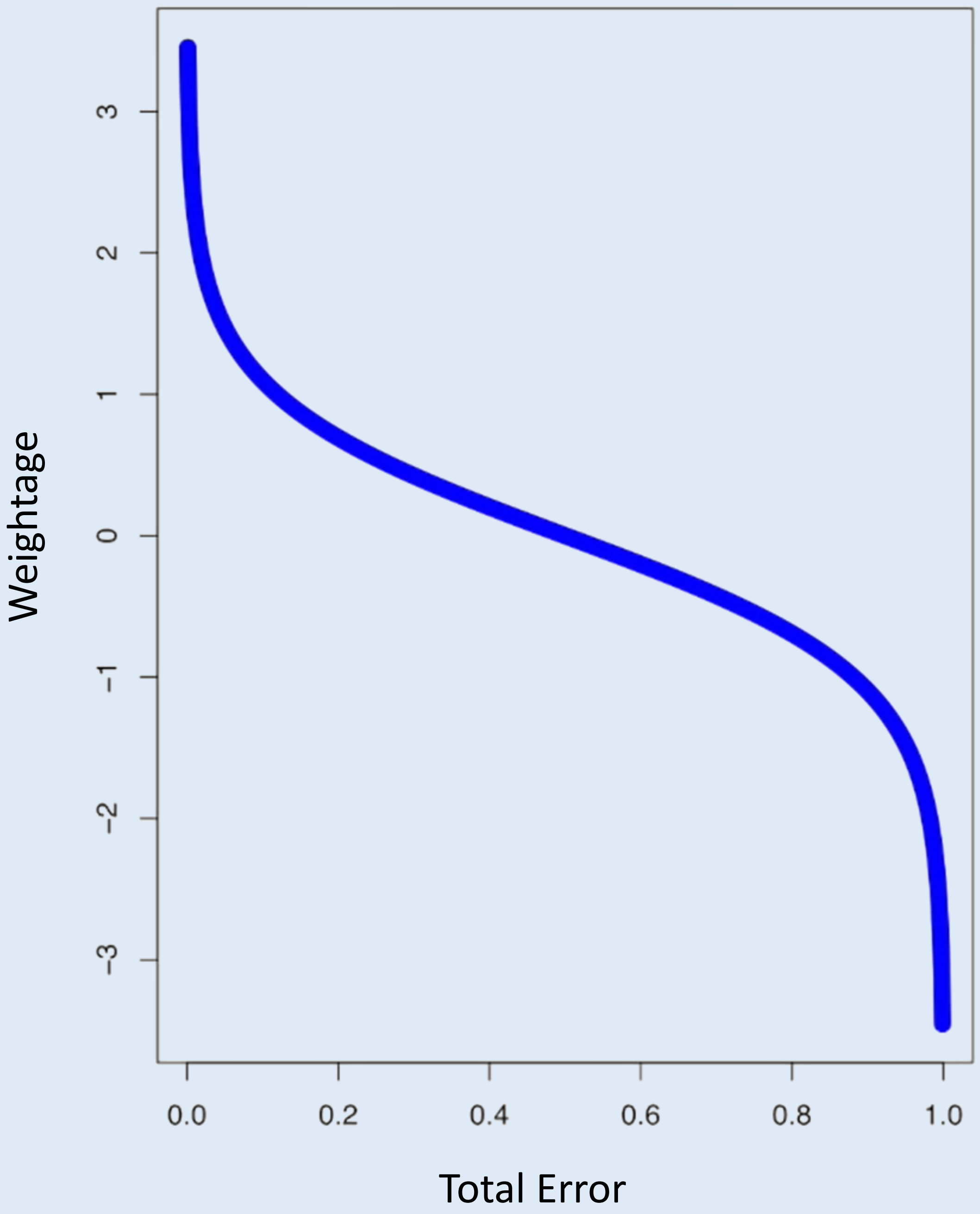


Gini Index = 0.5



Gini Index = 0.2

Stump with lowest Gini index will be the first stump in classification process

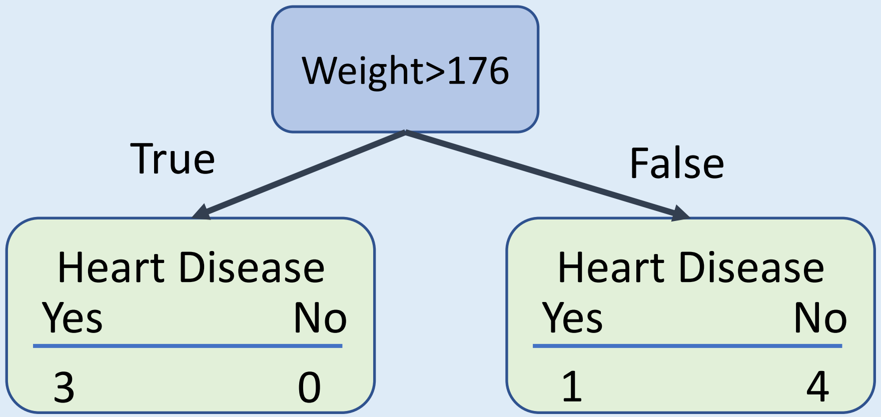


How much weightage will stump have?

It depends on how well it has classified the data

$$\mathit{Stump\ Weightage} = \frac{1}{2}\log\left(\frac{1 - \mathit{Total\ Error}}{\mathit{Total\ Error}}\right)$$

Now if the total error is zero or one the Weightage would be too large, therefore the small error is added to avoid weightage from getting too high



Weight>176

True

False

Heart Disease

Yes	No
3	0

Heart Disease

Yes	No
1	4

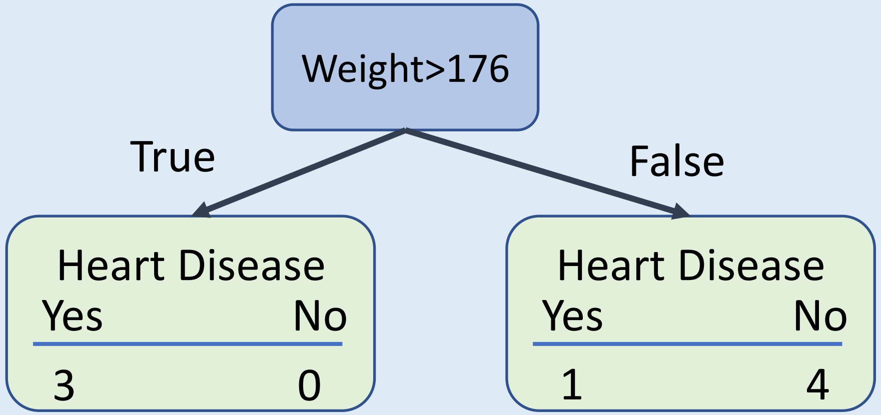
Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1 - Total\ Error}{Total\ Error}\right)$$

Total error for the stump will be equal to sum of weights of incorrectly classified samples

$$Therefore\ the\ total\ error\ here\ for\ the\ stump\ is = \frac{1}{8}$$

Since the error is equally distributed equally for each sample, error for a stump will be always between 1 and 0



Chest Pain	Blocked Arteries	Weight	Heart Disease
Yes	Yes	205	Yes
No	Yes	180	Yes
Yes	No	210	Yes
Yes	Yes	167	Yes
No	Yes	156	No
No	Yes	125	No
Yes	No	168	No
Yes	Yes	172	No

Sample Weight
1/8
1/8
1/8
1/8
1/8
1/8
1/8
1/8

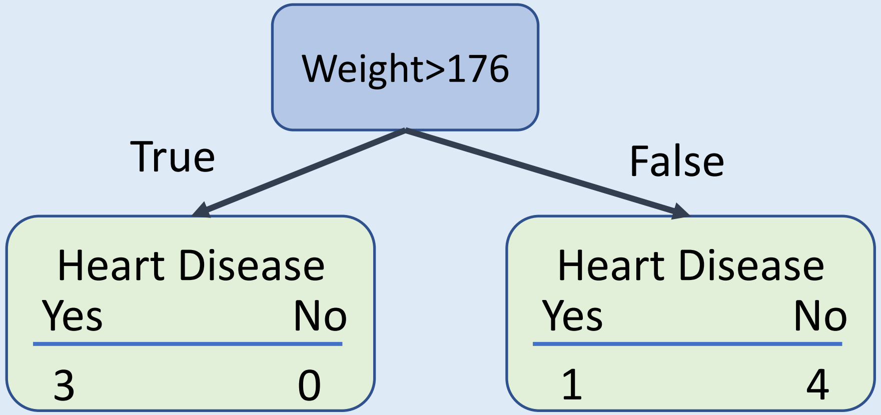
$$Therefore\ the\ total\ error\ here\ for\ the\ stump\ is = \frac{1}{8}$$

Let’s calculate the weightage for the stump

$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1 - Total\ Error}{Total\ Error}\right)$$

$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1 - 1/8}{1/8}\right)$$

$$Stump\ Weightage = 0.97$$



Chest Pain	Blocked Arteries	Weight	Heart Disease
Yes	Yes	205	Yes
No	Yes	180	Yes
Yes	No	210	Yes
Yes	Yes	167	Yes
No	Yes	156	No
No	Yes	125	No
Yes	No	168	No
Yes	Yes	172	No

Sample Weight
1/8
1/8
1/8
1/8
1/8
1/8
1/8
1/8

Let’s say, Chest Pain was the best stump

How much weightage would it have?

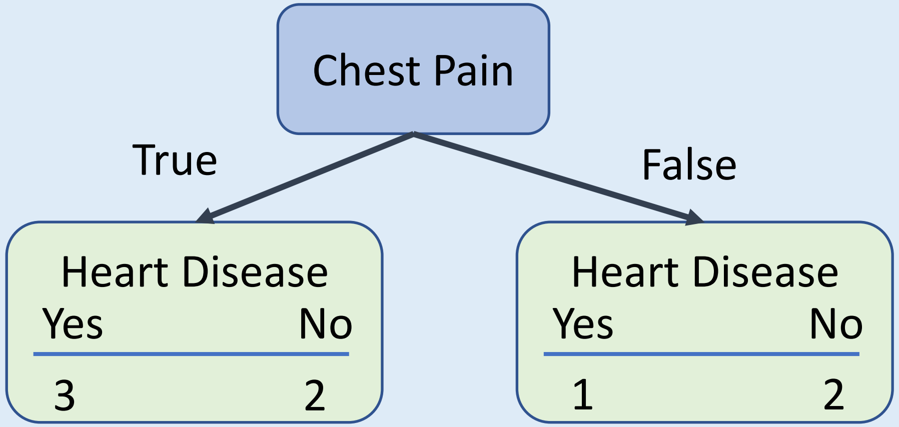
$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1-Total\ Error}{Total\ Error}\right)$$

We have to check how much error has it made.

It depends upon how many samples are miss classified

Assuming no chest pain means no heart disease and chest pain means heart disease we have

2 samples miss classified when Chest Pain is true
1 sample miss classified when Chest pain is false
Therefore there are total three samples miss classified



Chest Pain	Blocked Arteries	Weight	Heart Disease
Yes	Yes	205	Yes
No	Yes	180	Yes
Yes	No	210	Yes
Yes	Yes	167	Yes
No	Yes	156	No
No	Yes	125	No
Yes	No	168	No
Yes	Yes	172	No

Sample Weight
1/8
1/8
1/8
1/8
1/8
1/8
1/8
1/8

Let's say, Chest Pain was the best stump

How much weightage would it have?

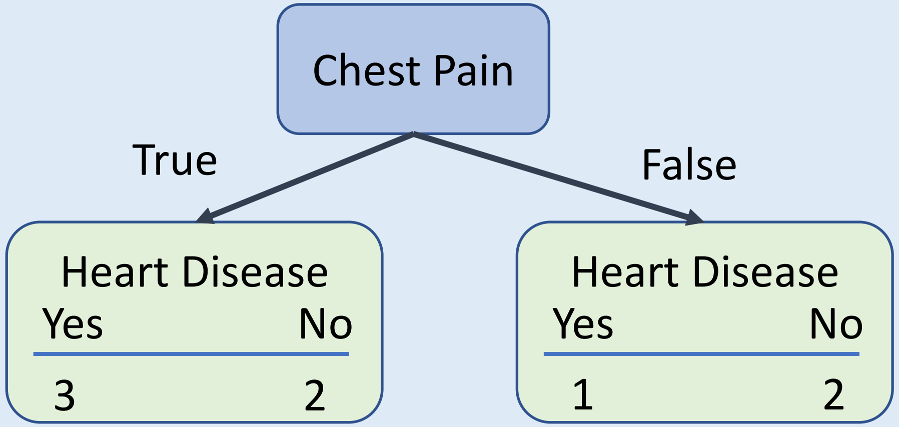
$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1 - Total\ Error}{Total\ Error}\right)$$

There are total three samples miss classified

$$Total\ Error = 3 \times \frac{1}{8}$$

$$Stump\ Weightage = \frac{1}{2}\log\left(\frac{1 - 3/8}{3/8}\right)$$

$$Stump\ Weightage = 0.42$$



Chest Pain

True

False

Heart Disease

Yes	No
3	2

Heart Disease

Yes	No
1	2

Let’s get beck to original first stump

We have to create new stump since a single stump won’t be suffice to get right prediction

Before doing so, we need to adjust the sample weights so that it would emphasize the mistake made by the first stump

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

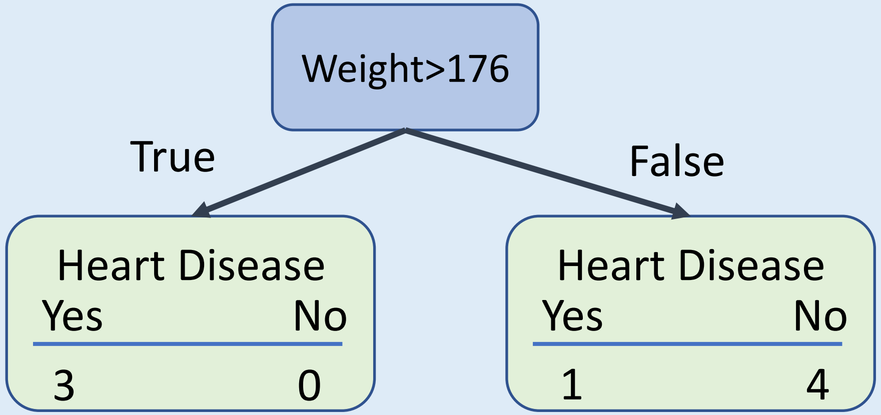
Increase the sample weight of incorrectly classified sample

And decrease the sample weight of correctly classified sample

Adjustments are done such a way that the total sample weight stays 1

$$New\ Sample\ Weight = Sample\ Weight \times e^{Weight\ of\ Stump}$$

$$New\ Sample\ Weight = \frac{1}{8} \times e^{0.97} \quad = \frac{1}{8} \times 2.64 \ \ = 0.33$$



Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

New Sample Weight for incorrectly classified sample

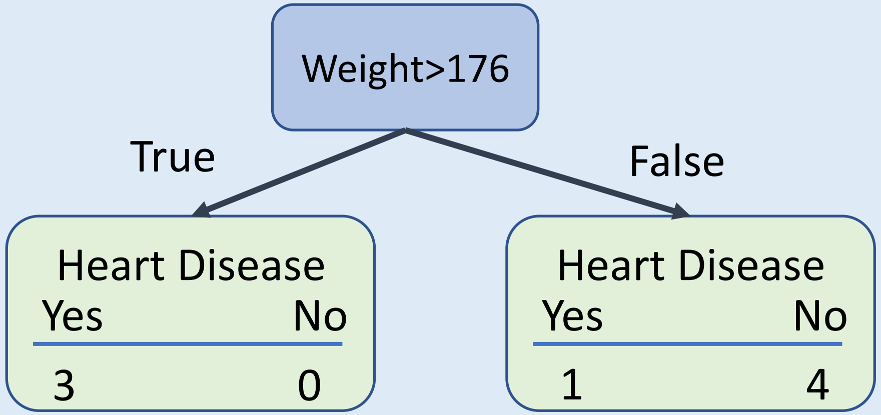
$$New\ Sample\ Weight = Sample\ Weight \times e^{Weight\ of\ Stump}$$

$$New\ Sample\ Weight = \frac{1}{8} \times e^{0.97} \quad = \frac{1}{8} \times 2.64 \quad = 0.33$$

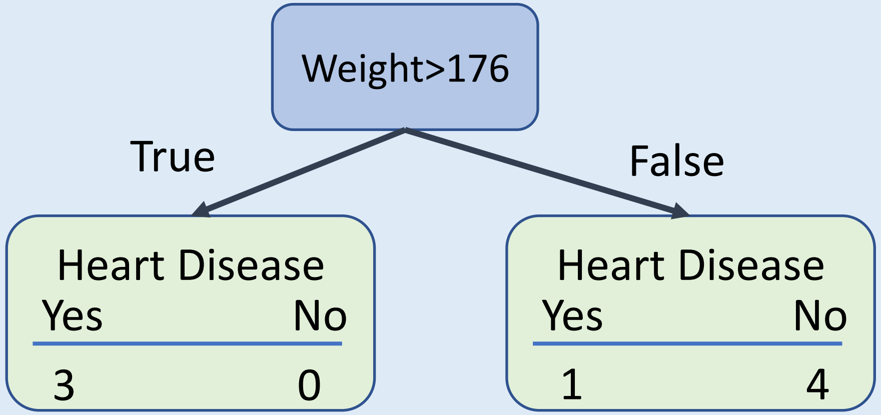
New Sample Weight for correctly classified sample

$$New\ Sample\ Weight = Sample\ Weight \times e^{-Weight\ of\ Stump}$$

$$New\ Sample\ Weight = \frac{1}{8} \times e^{-0.97} = \frac{1}{8} \times 0.38 \quad = 0.05$$



Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight	New Sample Weight	Normalised New Sample Weight
Yes	Yes	205	Yes	1/8	0.05	0.07
No	Yes	180	Yes	1/8	0.05	0.07
Yes	No	210	Yes	1/8	0.05	0.07
Yes	Yes	167	Yes	1/8	0.33	0.49
No	Yes	156	No	1/8	0.05	0.07
No	Yes	125	No	1/8	0.05	0.07
Yes	No	168	No	1/8	0.05	0.07
Yes	Yes	172	No	1/8	0.05	0.07



Weight>176

True

False

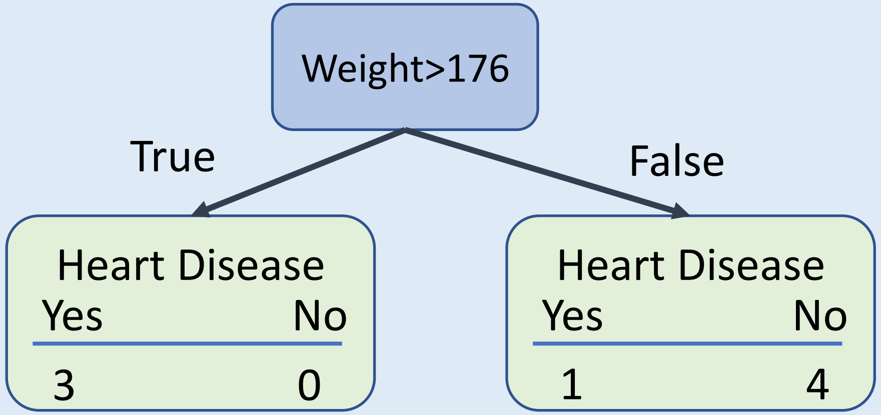
Heart Disease

Yes	No
3	0

Heart Disease

Yes	No
1	4

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07



Weight>176

True

False

Heart Disease

Yes No

3 0

Heart Disease

Yes No

1 4

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight

Later, random numbers are generated

If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	

Later, random numbers are generated

If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected

Random Number Generator

0.72

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	

Later, random numbers are generated

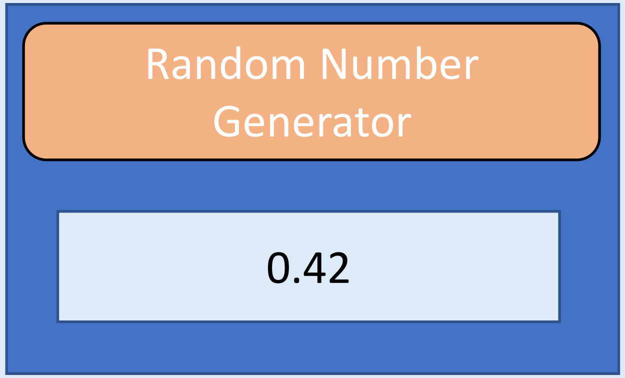
If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected



Random Number Generator

0.42

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	
No	Yes	125	No	

Later, random numbers are generated

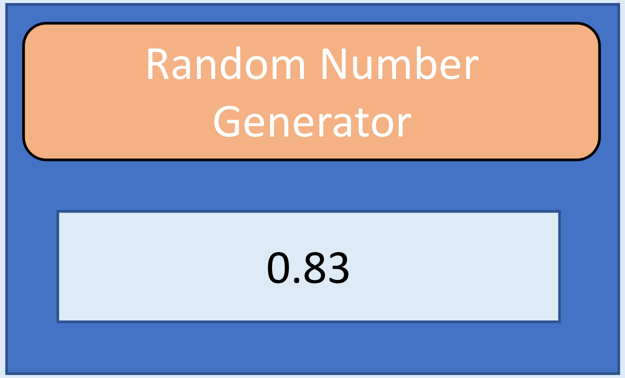
If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected



Random Number Generator

0.83

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	
No	Yes	125	No	
Yes	Yes	167	Yes	

Later, random numbers are generated

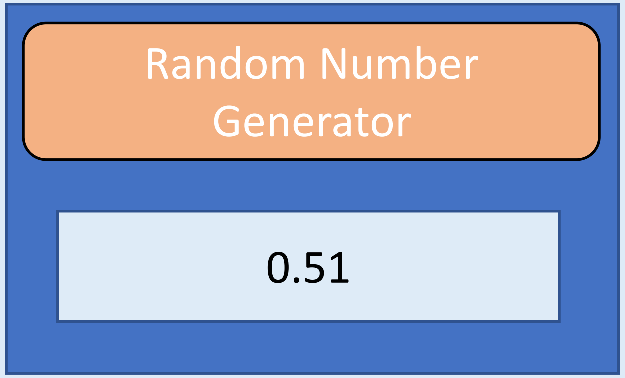
If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected



For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	
No	Yes	125	No	
Yes	Yes	167	Yes	
Yes	Yes	167	Yes	
Yes	Yes	172	No	
Yes	Yes	205	Yes	
Yes	Yes	167	Yes	

Later, random numbers are generated

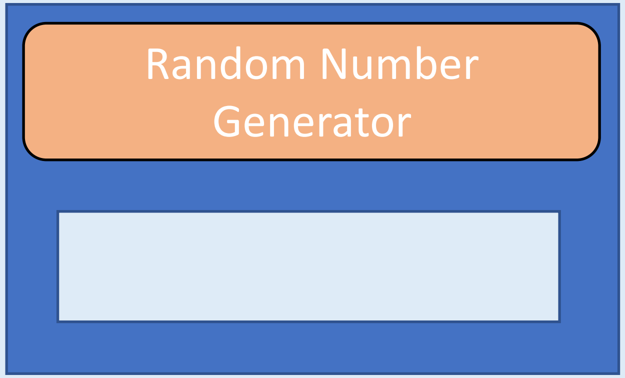
If the number generated is between 0 and 0.07, the first row is selected

If the number generated is between 0.07 and 0.14, the second row is selected

If the number generated is between 0.14 and 0.21, the third row is selected

If the number generated is between 0.21 and 0.21+0.49 = 0.7, the forth row is selected

If the number generated is between 0.7 and 0.7+0.07 = 0.77, the fifth row is selected



Random Number Generator

For creating new stump, we need to create a new dataset first, having same size as original

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	
No	Yes	125	No	
Yes	Yes	167	Yes	
Yes	Yes	167	Yes	
Yes	Yes	172	No	
Yes	Yes	205	Yes	
Yes	Yes	167	Yes	

Now new collection of samples will be used for creating new stump

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	
Yes	Yes	167	Yes	
No	Yes	125	No	
Yes	Yes	167	Yes	
Yes	Yes	167	Yes	
Yes	Yes	172	No	
Yes	Yes	205	Yes	
Yes	Yes	167	Yes	

Now new collection of samples will be used for creating new stump

Chest Pain	Blocked Arteries	Weight	Heart Disease	Sample Weight
No	Yes	156	No	1/8
Yes	Yes	167	Yes	1/8
No	Yes	125	No	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	172	No	1/8
Yes	Yes	205	Yes	1/8
Yes	Yes	167	Yes	1/8