Adaboost

Weak Learners/Stumps

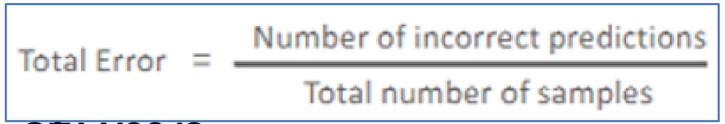
- In a Random Forest, each tree will be a full sized tree. There is no specified depth that each may adhere to
- Whereas in AdaBoost, the trees are usually one node and two leaves
- Such a tree is call a **“stump”**
- So, in AdaBoost, we will have a forest of stumps rather than trees
- Stumps are not good at making classifications as they can use only a single variable to make decision
- Whereas, a tree will use all the variables to make classifications
- So, we can say that stumps are **“weak learners”**

AdaBoost

- AdaBoost creates a Forest of Weak Learners to make classifications
- In AdaBoost, some stumps have a higher say in the final classification (Note: in a Random Forest, each tree has an equal weight)
- In AdaBoost, the stumps are made sequentially. The mistakes that the first stump makes influences how the next stump is made and this goes on for all the following stumps (Note: In a RF, each tree is made independently from each other)

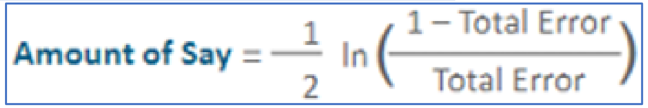
Steps

1. Assign weight to each sample row (to start with equal weight)
2. Get the Stump with the lowest Gini index
3. Calculate 'Total Error' for this stump



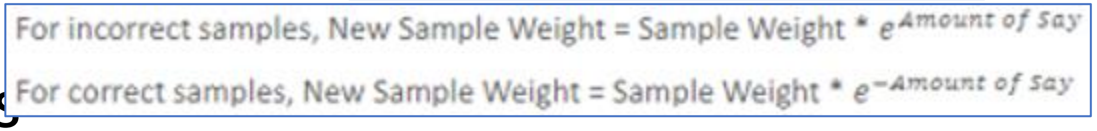
$$\text{Total Error} = \frac{\text{Number of incorrect predictions}}{\text{Total number of samples}}$$

4. Calculate the 'Amount of Say' for this stump



$$\textbf{Amount of Say} = \frac{1}{2}\ln\left(\frac{1-\text{Total Error}}{\text{Total Error}}\right)$$

5. Calculate the New sample weights



For incorrect samples, New Sample Weight = Sample Weight * $e^{Amount\ of\ Say}$

For correct samples, New Sample Weight = Sample Weight * $e^{-Amount\ of\ Say}$

6. Normalize the new sample weights
7. Resample based on the new weights
8. Create the next stump ->Go to 2 (using the new sample with weights)

Step-by-Step Implementation of AdaBoost

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes
171	Yes	45.4	Yes
122	Yes	28.8	Yes
197	No	30.5	Yes
189	No	25.6	Yes
116	No	30.1	No
139	No	27.1	No
92	Yes	32	No
85	No	26.6	No

Dataset for AdaBoost

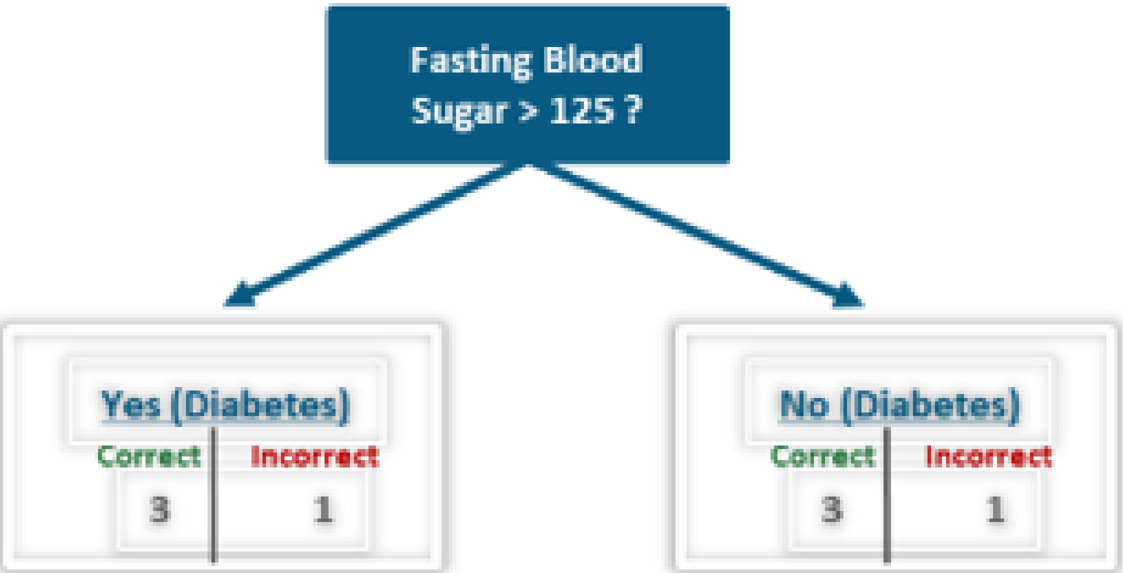
Step 1: Giving Sample Weight to Each Sample

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight
171	Yes	45.4	Yes	1/8
122	Yes	28.8	Yes	1/8
197	No	30.5	Yes	1/8
189	No	25.6	Yes	1/8
116	No	30.1	No	1/8
139	No	27.1	No	1/8
92	Yes	32	No	1/8
85	No	26.6	No	1/8

Sample Weight = $\frac{1}{Total\ number\ of\ samples} = \frac{1}{8}$ → This makes all samples equally important

Step 2: Creating Stumps based on Fasting Blood Sugar

Prediction based on Fasting Blood Sugar:



Fasting Blood Sugar > 125 ?

Yes (Diabetes)

Correct 3

Incorrect 1

No (Diabetes)

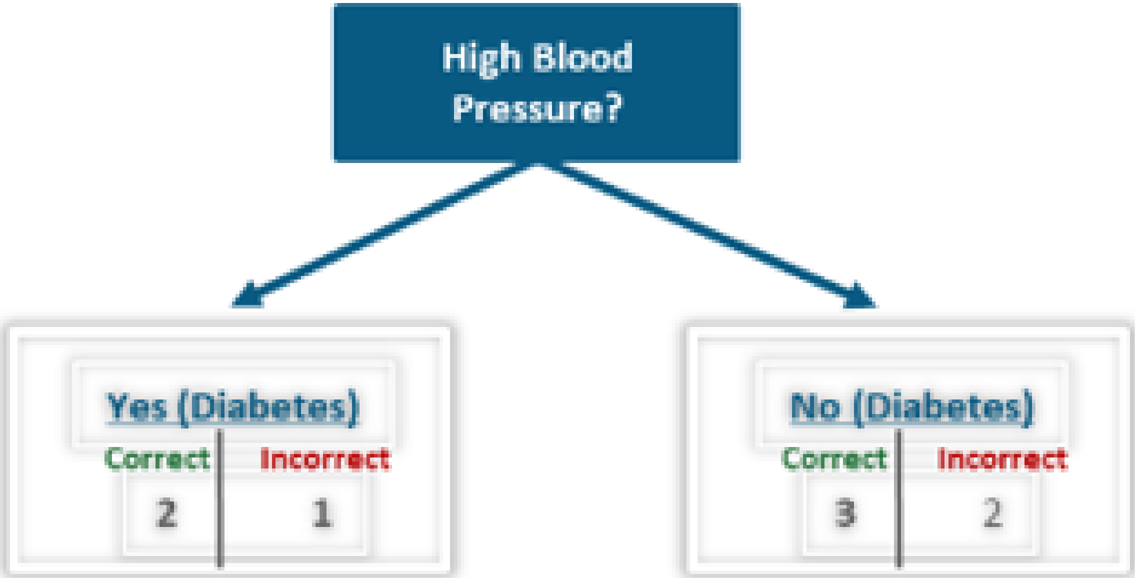
Correct 3

Incorrect 1

Fasting Blood Sugar	Diabetes
171	Yes
122	Yes
197	Yes
189	Yes
116	No
139	No
92	No
85	No

Step 3: Creating Stumps based on High Blood Pressure

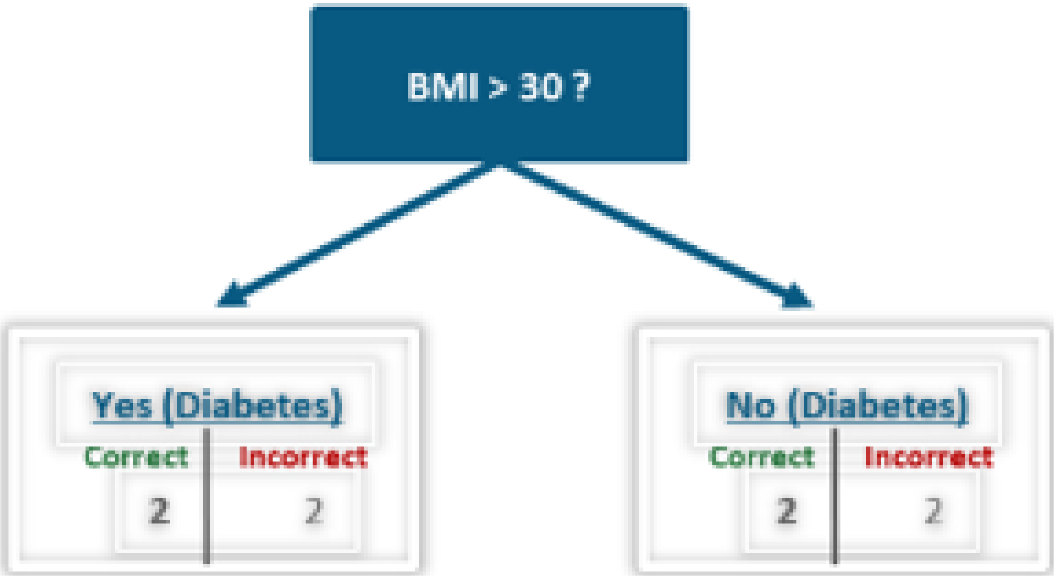
Prediction based on High Blood Pressure:



High Blood Pressure	Diabetes
Yes	Yes
Yes	Yes
No	Yes
No	Yes
No	No
No	No
Yes	No
No	No

Step 4: Creating Stumps based on BMI

Prediction based on BMI:

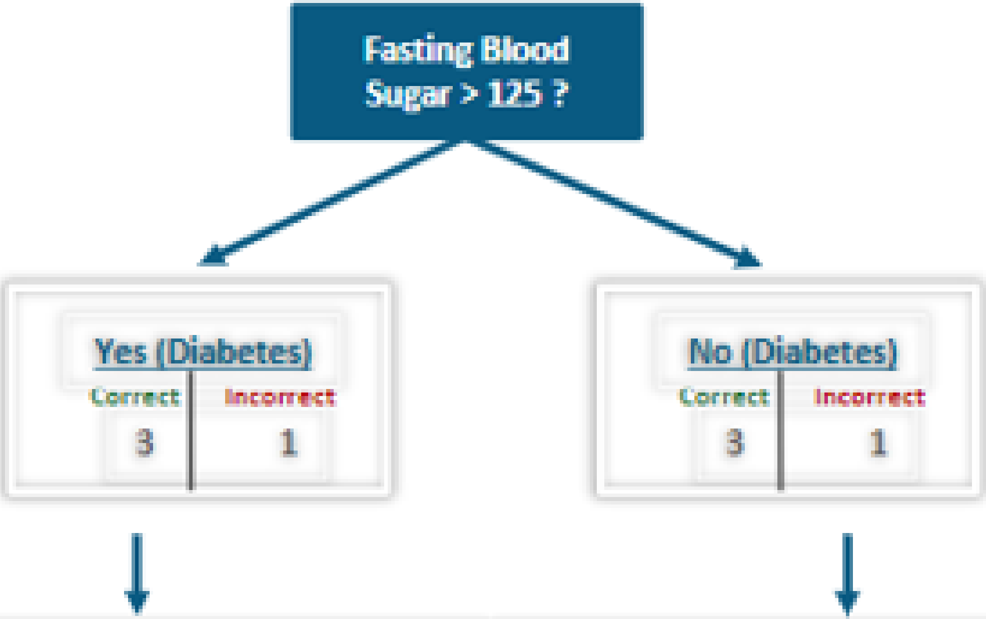


BMI	Diabetes
45.4	Yes
28.8	Yes
30.5	Yes
25.6	Yes
30.1	No
27.1	No
32	No
26.6	No

Step 5: Calculate Gini Index for each Stump

Gini Index = 1 – [(Probability of Correct Prediction)^2 + (Probability of Incorrect Prediction)^2]

For Example:



$$\text{Gini Index} = 1 - \left[\left(\frac{3}{3+1} \right)^2 + \left(\frac{1}{3+1} \right)^2 \right]$$

= 0.375

$$\text{Gini Index} = 1 - \left[\left(\frac{3}{3+1} \right)^2 + \left(\frac{1}{3+1} \right)^2 \right]$$

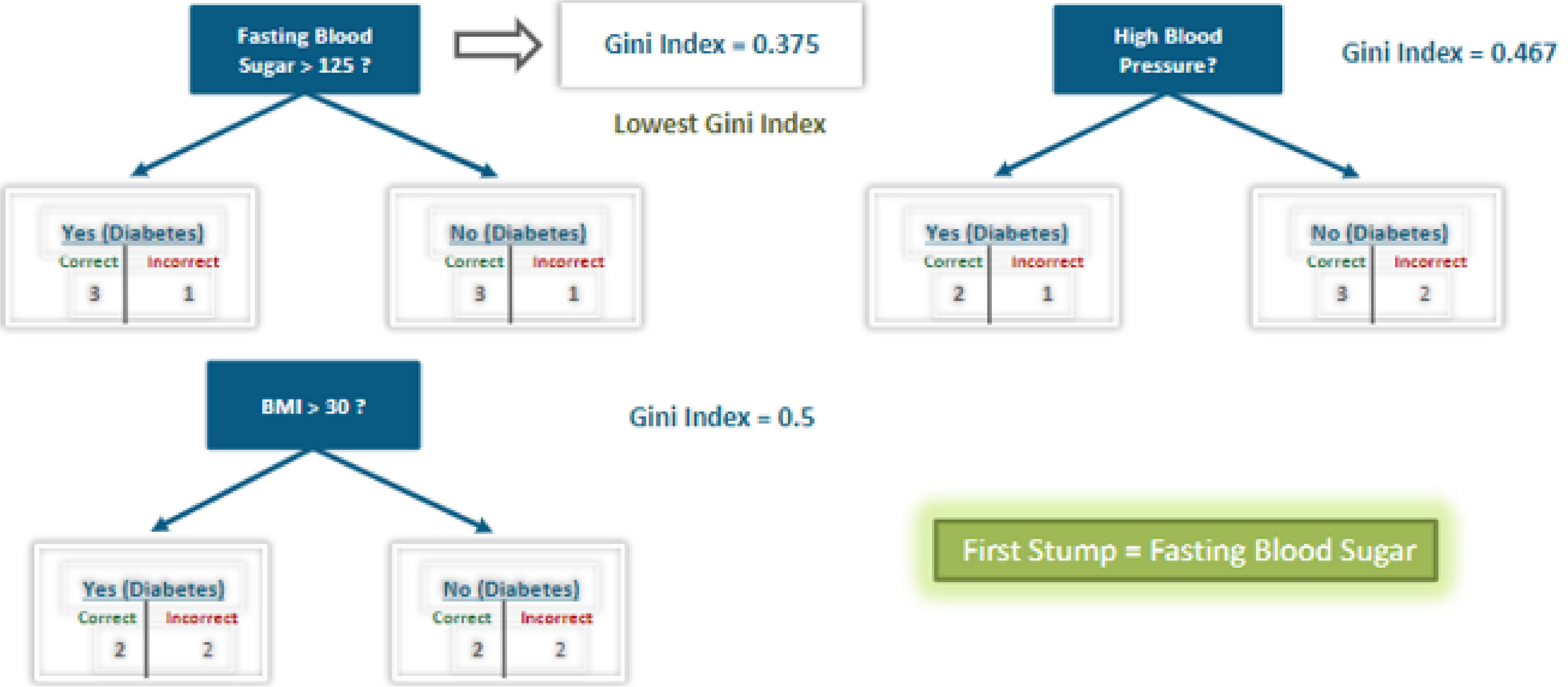
= 0.375

Average Gini Index

$$= \left(\frac{4}{8} \right) * 0.375 + \left(\frac{4}{8} \right) * 0.375$$

= 0.375

Step 6: Select the First Stump based on Gini Index



Fasting Blood Sugar > 125 ?

Gini Index = 0.375

Lowest Gini Index

Yes (Diabetes)

Correct	Incorrect
3	1

No (Diabetes)

Correct	Incorrect
3	1

High Blood Pressure?

Gini Index = 0.467

Yes (Diabetes)

Correct	Incorrect
2	1

No (Diabetes)

Correct	Incorrect
3	2

BMI > 30 ?

Gini Index = 0.5

Yes (Diabetes)

Correct	Incorrect
2	2

No (Diabetes)

Correct	Incorrect
2	2

First Stump = Fasting Blood Sugar

Step 7: Calculate Total Error for Selected Stump

Let us calculate the “Amount of Say” for the selected stump in the final classification

For this, let us first calculate the “Total Error” made by this stump

$$\text{Total Error} = \frac{\text{Number of incorrect predictions}}{\text{Total number of samples}}$$

$$\text{Total Error for Fasting Blood Sugar} = \frac{2}{8} = 0.25$$

Fasting Blood Sugar	Diabetes
171	Yes
122	Yes
197	Yes
189	Yes
116	No
139	No
92	No
85	No

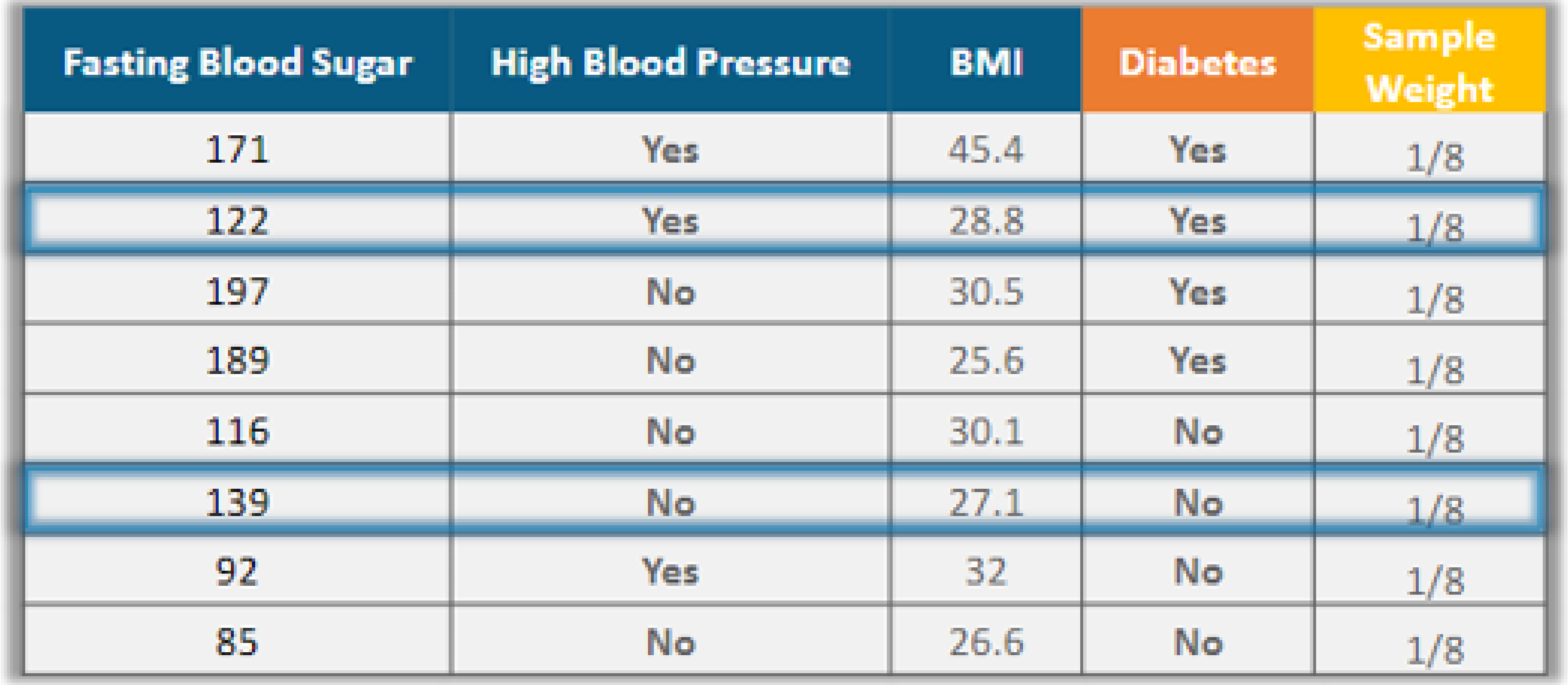
Step 8: Calculate Amount of Say for Selected Stump

$$\textbf{Amount of Say} = \frac{1}{2} \ln\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

- Total error value is always between 0 to 1
- Total error is small for good stumps, Amount of Say will have a high positive value
- Total error is high for bad stumps, Amount of Say will have a large negative value

Amount of Say for Fasting Blood Sugar = 0.55

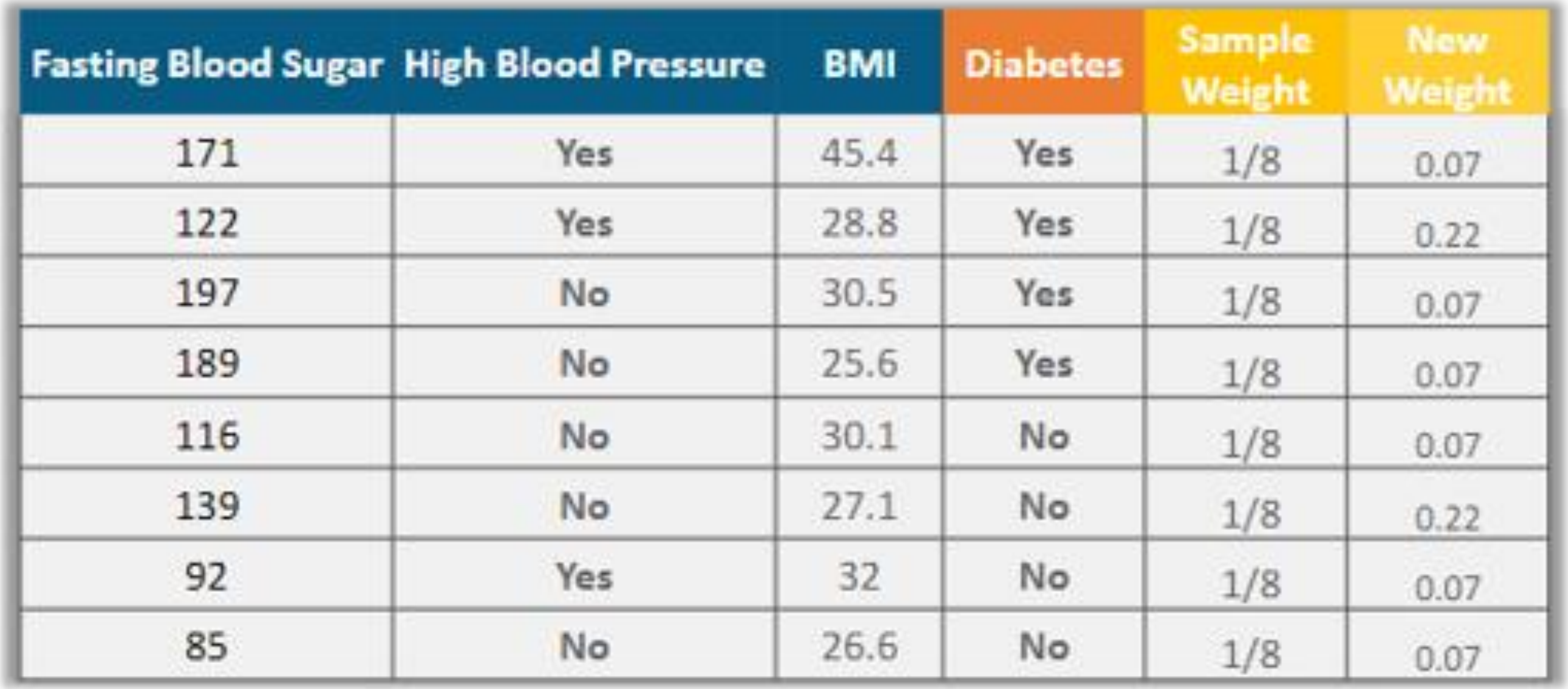
Step 9: Calculate new Sample Weights



Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight
171	Yes	45.4	Yes	1/8
122	Yes	28.8	Yes	1/8
197	No	30.5	Yes	1/8
189	No	25.6	Yes	1/8
116	No	30.1	No	1/8
139	No	27.1	No	1/8
92	Yes	32	No	1/8
85	No	26.6	No	1/8

- Increase the weight of incorrectly predicted samples and decrease the weight of correctly predicted samples
  - For incorrect samples, New Sample Weight = Sample Weight * $e^{Amount\ of\ Say}$ = 0.22
  - For correct samples, New Sample Weight = Sample Weight * $e^{-Amount\ of\ Say}$ = 0.07

Step 10: Add new Sample Weights



Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight	New Weight
171	Yes	45.4	Yes	1/8	0.07
122	Yes	28.8	Yes	1/8	0.22
197	No	30.5	Yes	1/8	0.07
189	No	25.6	Yes	1/8	0.07
116	No	30.1	No	1/8	0.07
139	No	27.1	No	1/8	0.22
92	Yes	32	No	1/8	0.07
85	No	26.6	No	1/8	0.07

Step 11: Add Normalized Weights

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight	New Weight	Norm. Weight
171	Yes	45.4	Yes	1/8	0.07	0.08
122	Yes	28.8	Yes	1/8	0.22	0.26
197	No	30.5	Yes	1/8	0.07	0.08
189	No	25.6	Yes	1/8	0.07	0.08
116	No	30.1	No	1/8	0.07	0.08
139	No	27.1	No	1/8	0.22	0.26
92	Yes	32	No	1/8	0.07	0.08
85	No	26.6	No	1/8	0.07	0.08

- Normalized Weight = New Weight of each sample/ (Sum of all New Weights)
- **Normalized Weight will be used to identify the next stumps**

Step 11: Revised weights

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight
171	Yes	45.4	Yes	0.125
122	Yes	28.8	Yes	0.125
197	No	30.5	Yes	0.125
189	No	25.6	Yes	0.125
116	No	30.1	No	0.125
139	No	27.1	No	0.125
92	Yes	32	No	0.125
85	No	26.6	No	0.125

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight Normalized
171	Yes	45.4	Yes	0.08
122	Yes	28.8	Yes	0.26
197	No	30.5	Yes	0.08
189	No	25.6	Yes	0.08
116	No	30.1	No	0.08
139	No	27.1	No	0.26
92	Yes	32	No	0.08
85	No	26.6	No	0.08

Step 12: Resampling based on weights

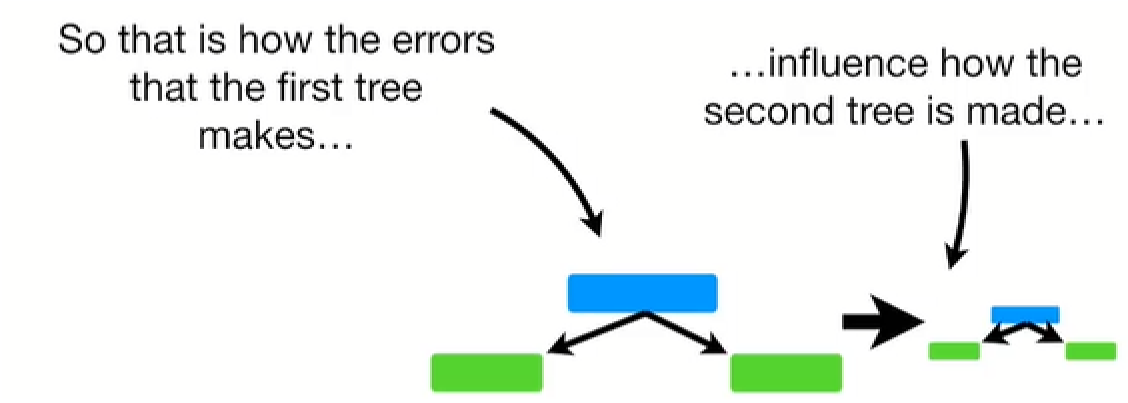
- We start by making a new, but empty, dataset that is the same size as original
- We pick a random number between 0 and 1
- And we see where the number falls when we use the **Sample Weights** like a distribution
- If the number is between 0.42 and 0.50, then we put the 4th record in the new dataset
- If the next number is between 0.08 and 0.34, then we put the 2nd record from left to the new dataset
- …and so on, we create the new dataset with exactly the same size as original
- **As we can see, the observations in error will reflect more times in the new dataset**
- Now with the new dataset, we can make the weights equal and start creating a new stump on this dataset

Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight	Cumulative
171	Yes	45.4	Yes	0.08	0.08
122	Yes	28.8	Yes	0.26	0.34
197	No	30.5	Yes	0.08	0.42
189	No	25.6	Yes	0.08	0.50
116	No	30.1	No	0.08	0.58
139	No	27.1	No	0.26	0.84
92	Yes	32	No	0.08	0.92
85	No	26.6	No	0.08	1.00

Resampling

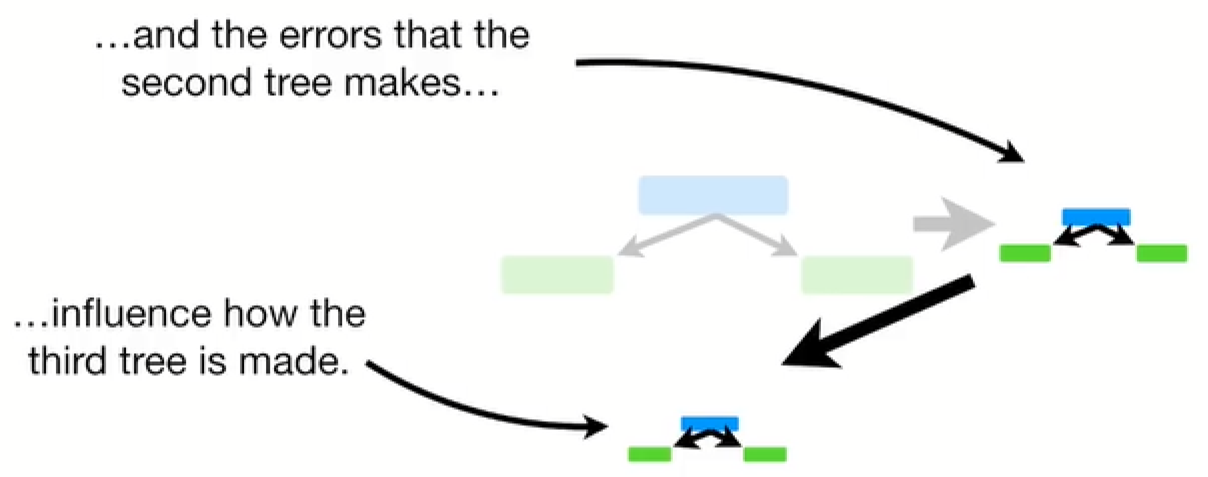
Fasting Blood Sugar	High Blood Pressure	BMI	Diabetes	Sample Weight
189	No	25.6	Yes	0.08
122	Yes	28.8	Yes	0.26
116	No	30.1	No	0.08
139	No	27.1	No	0.26
122	Yes	28.8	Yes	0.26
139	No	27.1	No	0.26
92	Yes	32	No	0.08
139	No	27.1	No	0.26

Sequential tree creation



So that is how the errors that the first tree makes…

…influence how the second tree is made…

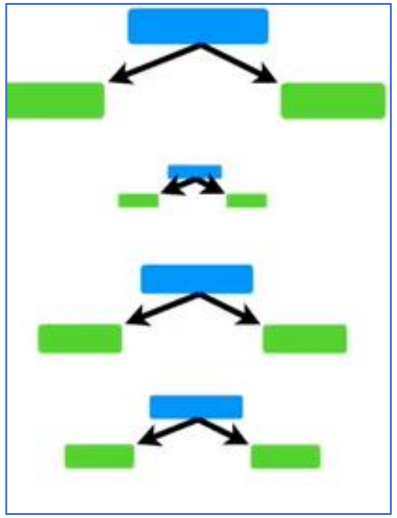


…and the errors that the second tree makes…

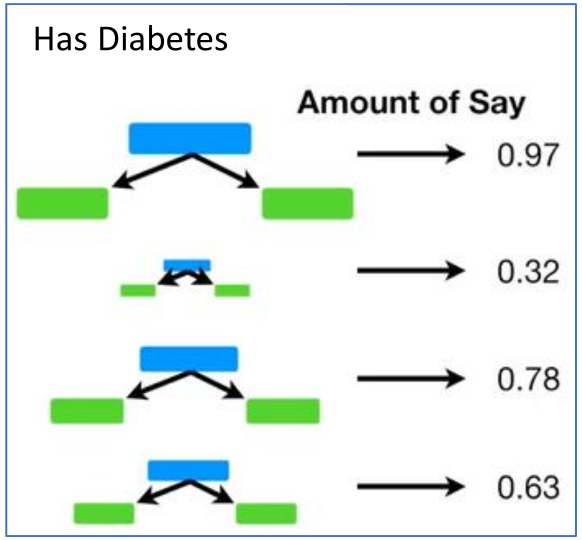
…influence how the third tree is made.

How does Adaboost make Classifications using the forest of stumps

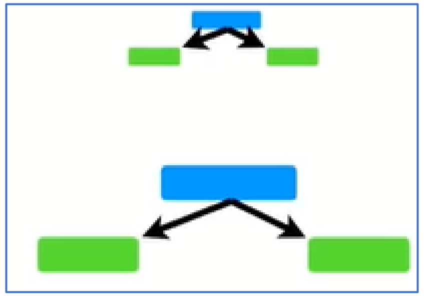
Imagine that these stumps classified a patient as **Has Diabetes**



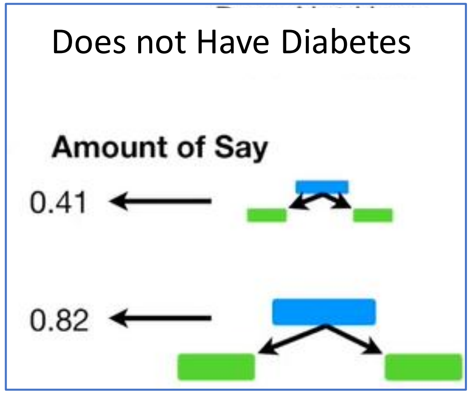
These are the **Amounts of Say** for these stumps…



…and these stumps classified the patient as **Does Not Have Diabetes**

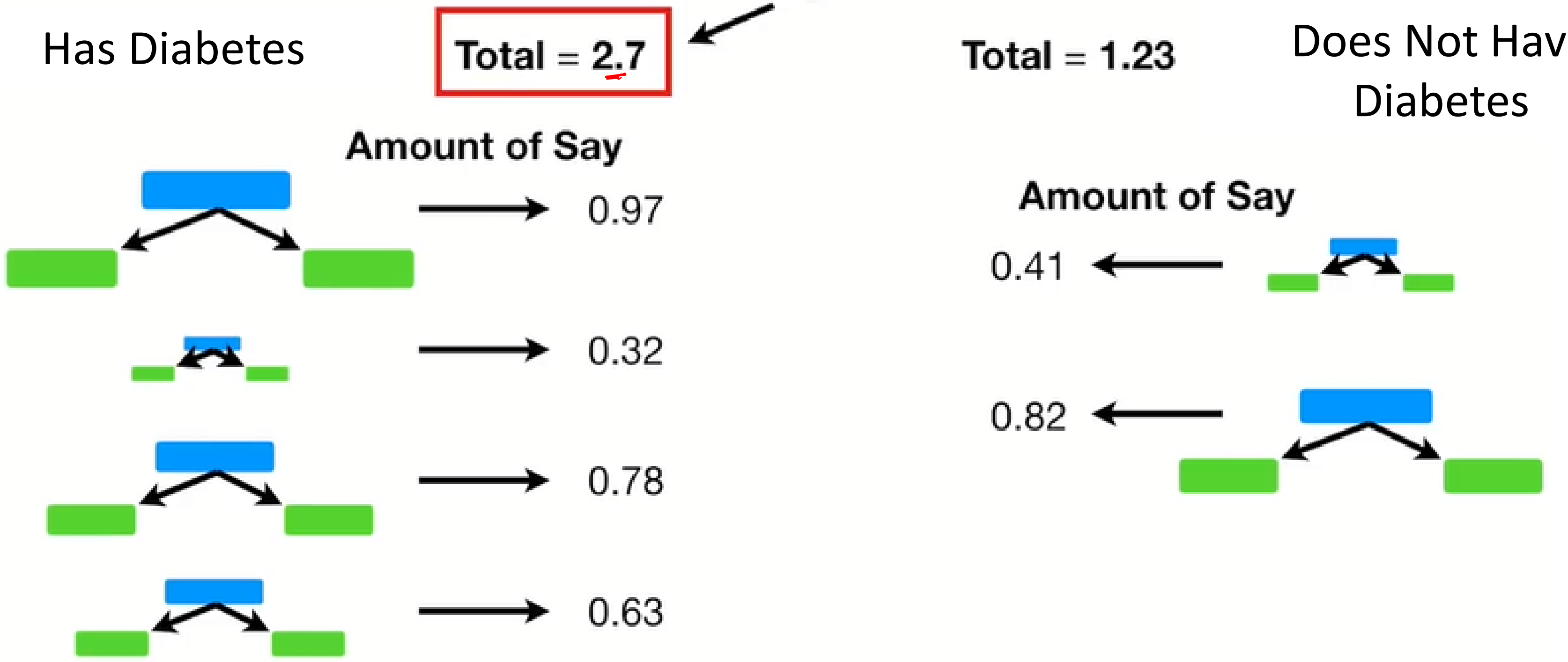


.. and these are the **Amounts of Say** for these stumps…



We add up the Amounts of Say for each group of stumps

Ultimately, the patient is classified as **Has Diabetes** because this is the larger sum



Has Diabetes

Does Not Have Diabetes