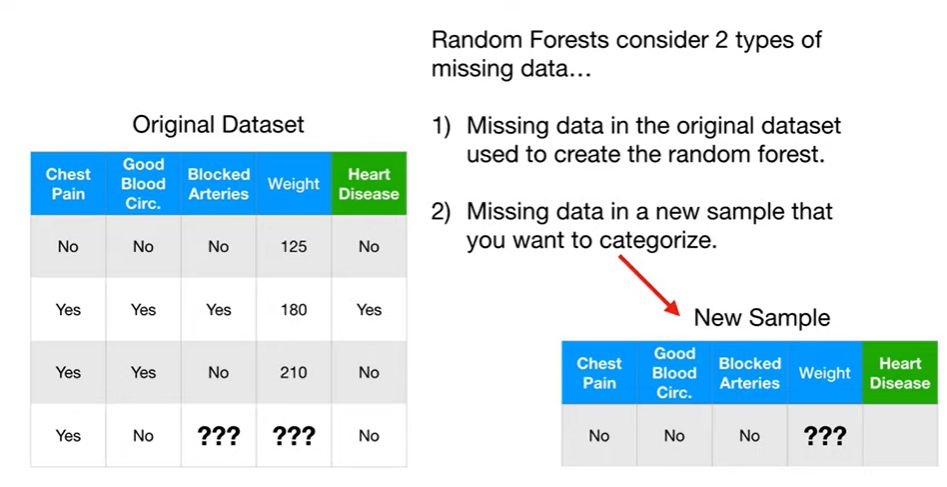
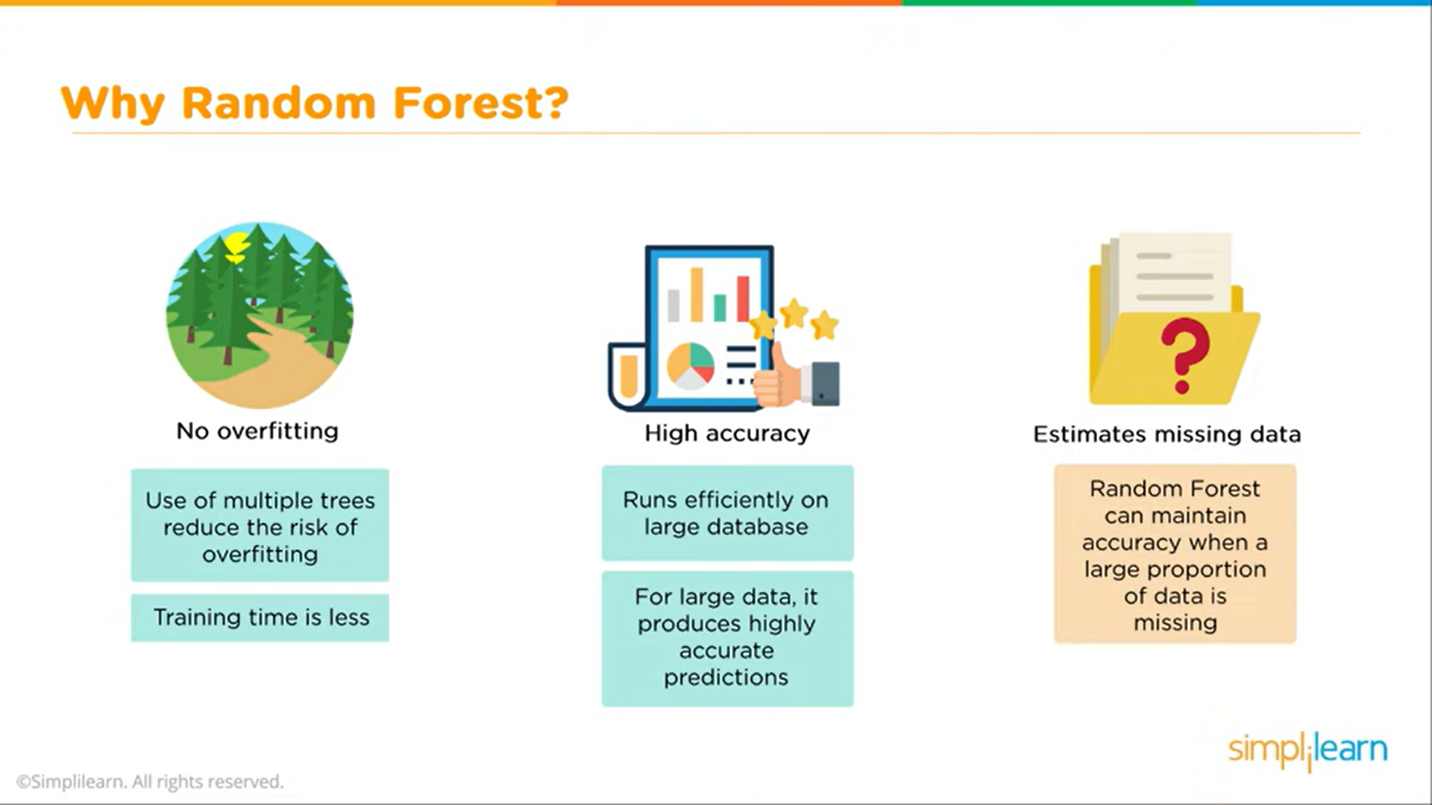
**RANDOM FOREST**

**Q: Why Random Forest when we have Decision Trees?**

* Decision Trees are highly prone to Overfitting, while this problem is solved in Random Forest
* Decision Trees are highly used for

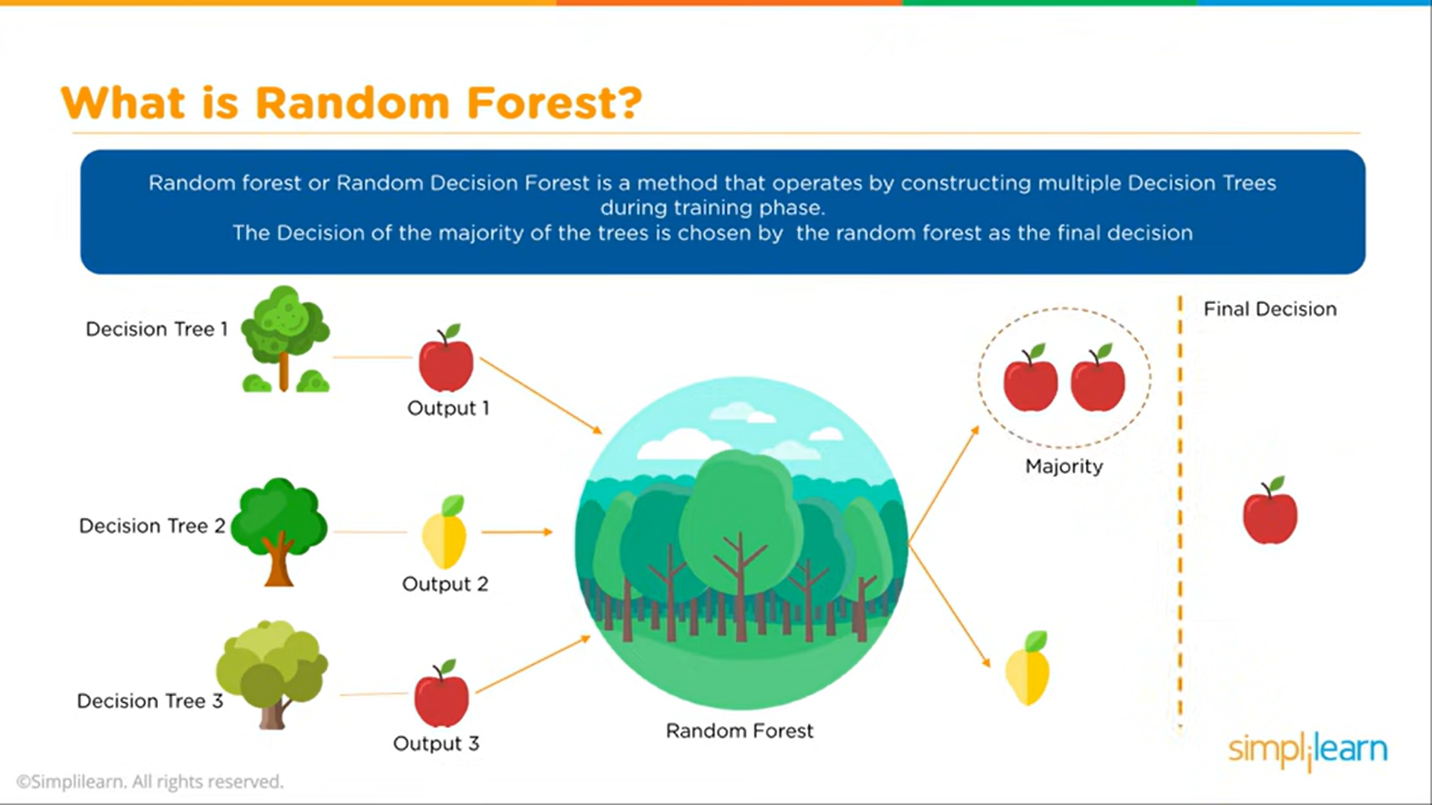
1. Estimating missing data : It is an Iterative method that uses “Proximity Matrics” and starts with initial ‘guesses’ for missing data, which after multiple iterations converges to give us ‘final values’ for the missing data.



* 

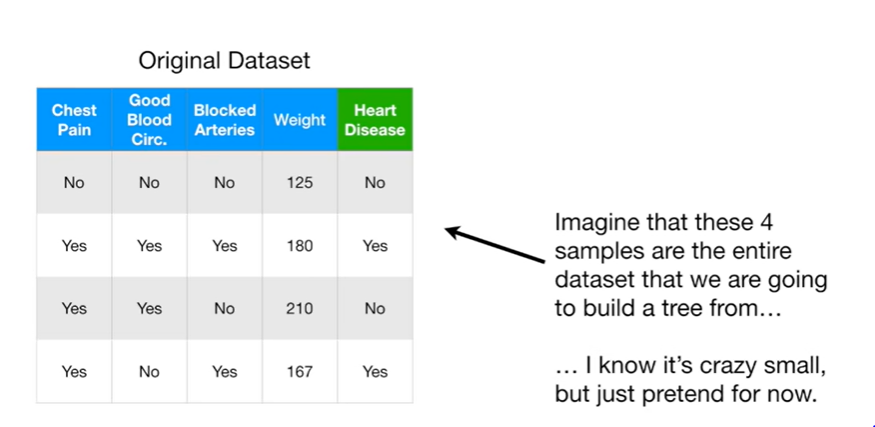
**Q: How does random forest work**

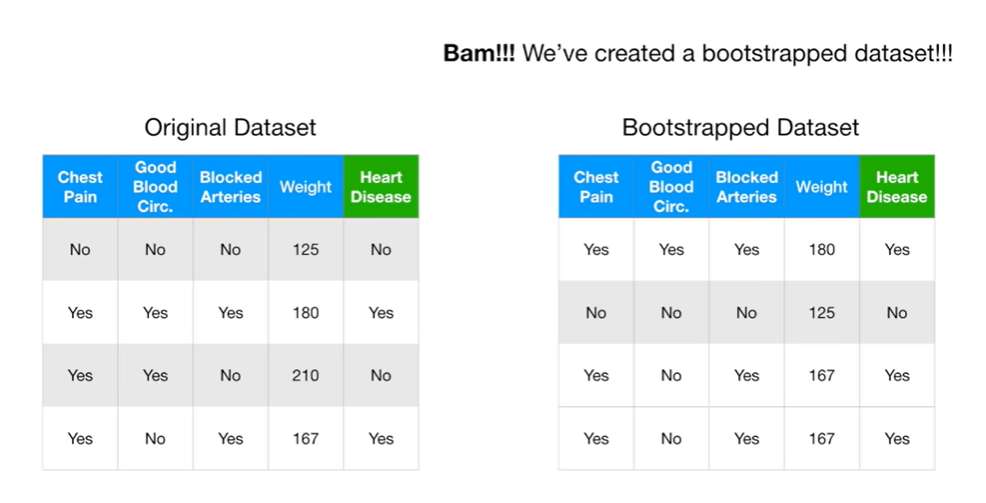
* It uses many Decision Trees to arrive at a decision



* Steps involved in creating a Random Forest

1. Creating a Bootstrapped Dataset (pick up any rows with Replacement is allowed) – Here we randomly choose rows of dataset(with repeatability allowed for these rows) and create a new Dataset called Bootstrapped Dataset

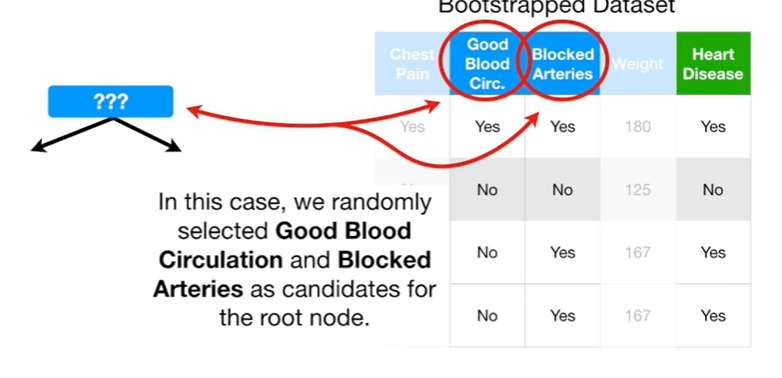
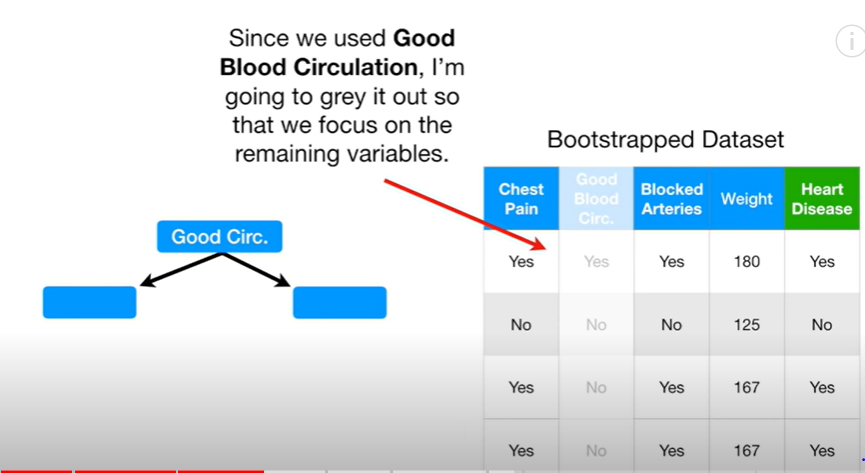




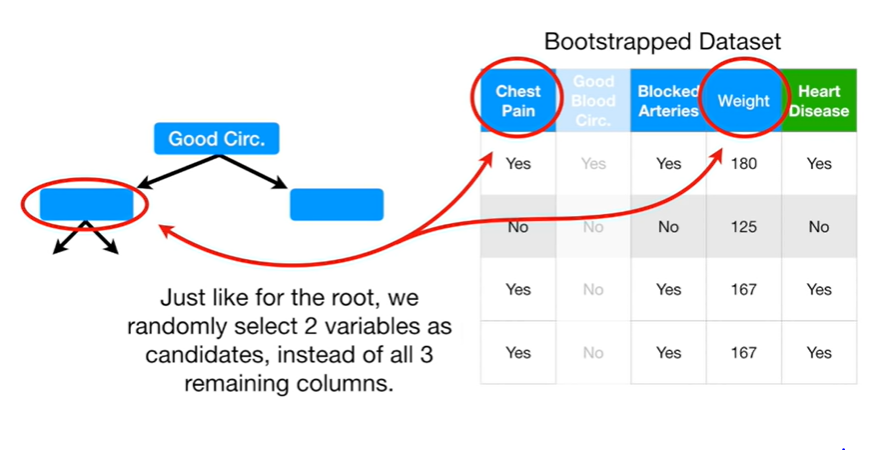
Observe : that Bootstrapped dataset has last 2 rows as same/repeated (i.e. taken from the last row of the Original dataset twice).

1. Create a Decision Tree using Bootstrapped Dataset(modified) BUT only use RANDOM SUBSET OF VARIABLES/COLUMNS AT EACH STEP .

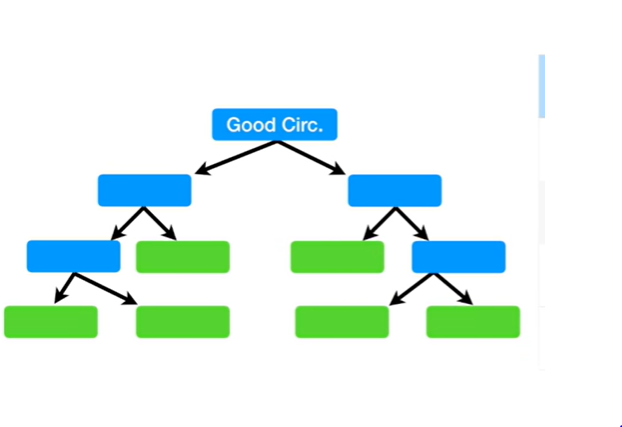
Example: below for Root node, we use only 2 variables / columns at each step/node of Decision tree i.e. say Good Blood Circulation(i.e column2/Variable 2) and Blocked Art(i.e. Column 3/ Variable 3) . Here say, Variable2 does a better job at classifying than Variable3 . Thus, at Root Node we have Var.2 i.e. Good Blood Circulation.

1. Next, we again choose 2 random Variables/columns (say , Column1/Variable1 =Chest pain and column4/Variable4 = Weight) and say among these 2 the Variable 1 did a better job of classifying. So , At 2nd Decision Node we put Variable 1 = Chest Pain.

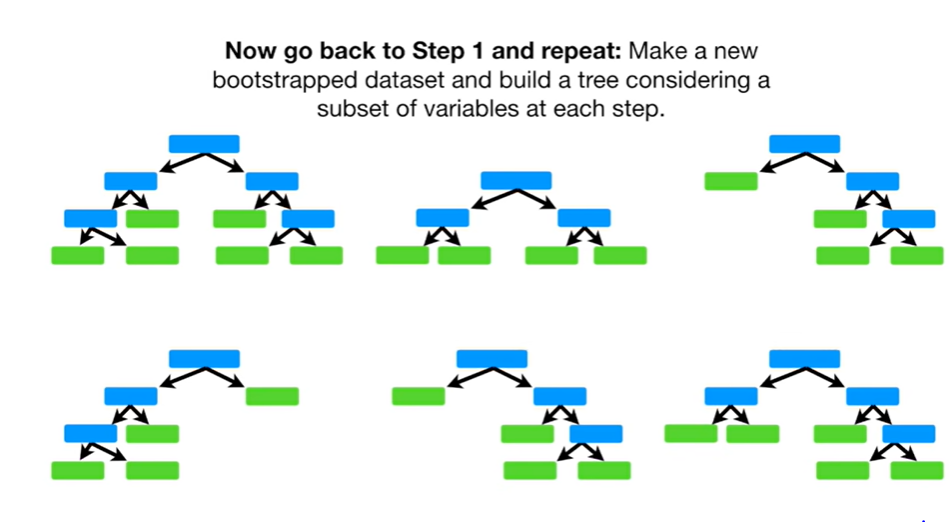


1. We keep on building this Decision tree only using the a subset of variables (subset of 2 variables in this example) . Say we get something like this at the end.



1. Now we have 1 Decsion Tree made. Now, we go back and create a new Bootstrapped Dataset and build a New Tree using this Bootstrapped dataset and only a subset of variables at each step. Repeat this till you get 100s or 1000s of new Decision trees.

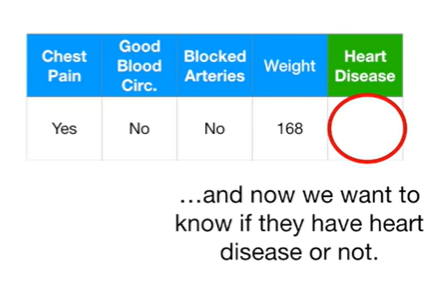
**NOTE:** below we have shown only 6 of the Decision Trees because of simplicity.



**NOTE: USING THESE MANY VARIETIES OF TREES IS WHAT MAKES RANDOM FOREST MORE EFFECTIVE THAN DECISION TREES.**

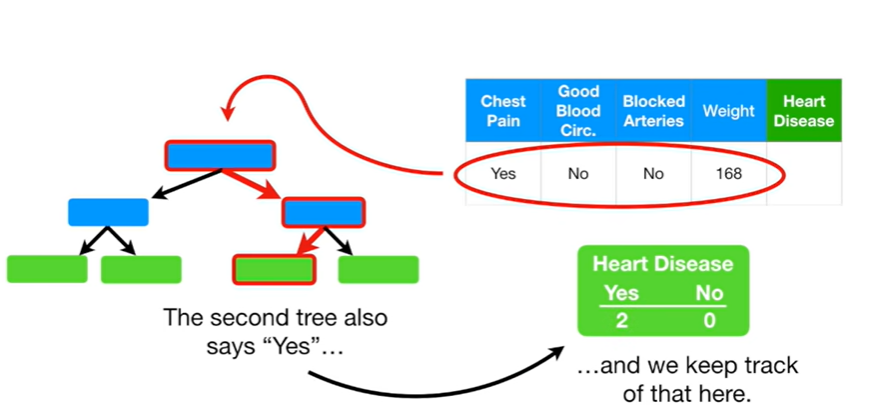
1. Now that we have created Random Forest (with 6 tress in this case, but, the algorithm would have created 100s of such trees).

We can now take new Dataset (of new patient ) for which we want to predict/classify the patient as having Heart disease or Not, and pass it down the 1st tree.

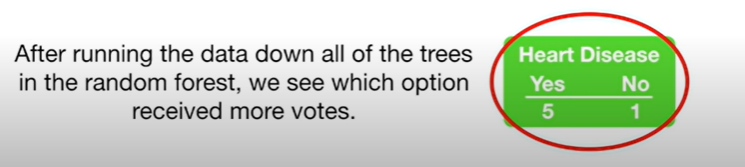


The 1st Tree says “YES” . We keep track of this record in a tabular form (as shown above).

1. Slly, we pass this data down the 2nd tree…….and it also says “YES”. Again we keep its track in the same table.



1. Slly, we run this same new data down the rest of the trees and the their outcome as record.



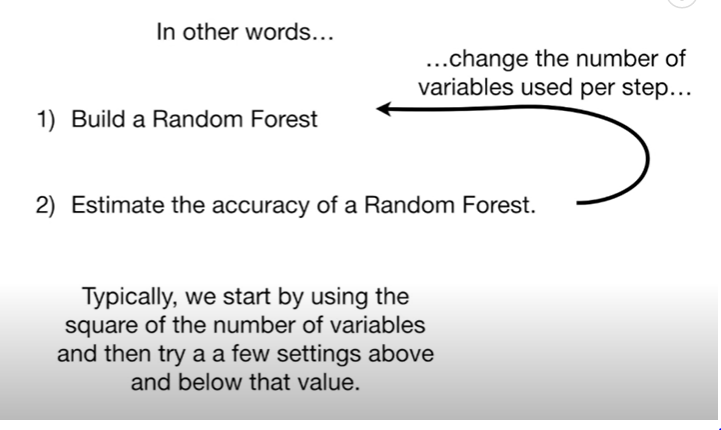
1. We now see which option has received the most votes . Clearly, here “YES” has received the most votes (= 5 votes ) and “NO” has received less votes (=1 vote).

THUS, WE CLASSIFY THIS PATIENT AS “HAVING HEART DISEASE”.

**NOTE:**

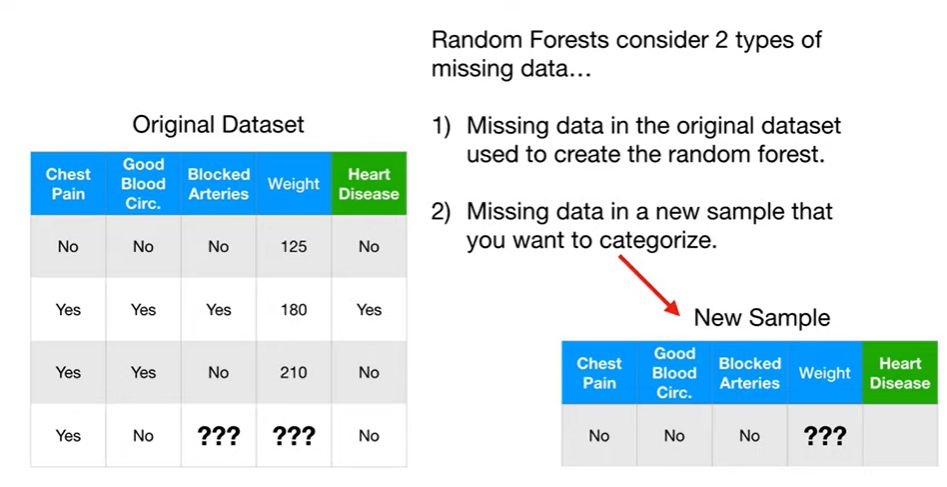
1. ***BOOTSTRAPPING THE DATA PLUS USING THE AGGREGATE (as shown in above steps) TO MAKE A DECISION PIS CALLED IS CALLED ‘BAGGING’.***
2. ***How do we decide the ‘No. of Variables to be used in Each Step’?*** *(recall in above Example we had used only 2 variables/2 columns from Original Datset, in each step, while creating a Bootstrapped Data).*

***THUMB RULE: We start by using the square root of the Number of variables in the original dataset (=4 in above example. So, root = 2 variables per step for creating Bootstrapped dataset). We can then try a few numbers above and below (like 3 or 1 variables per step) and check which gives us the highest accuracy.***

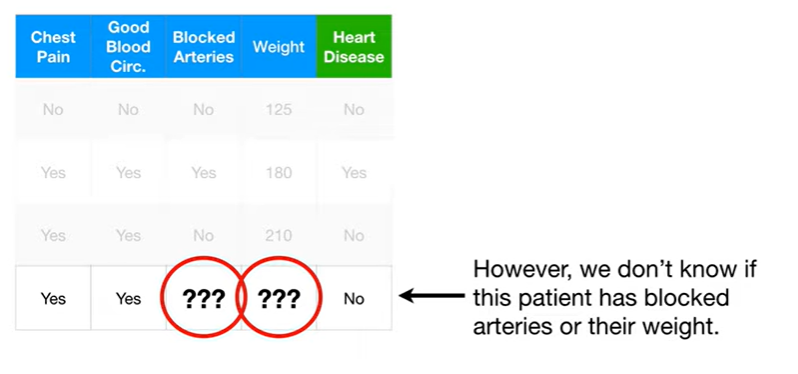


**USING RANDOM FOREST TO DETERMINE MISSING DATA**

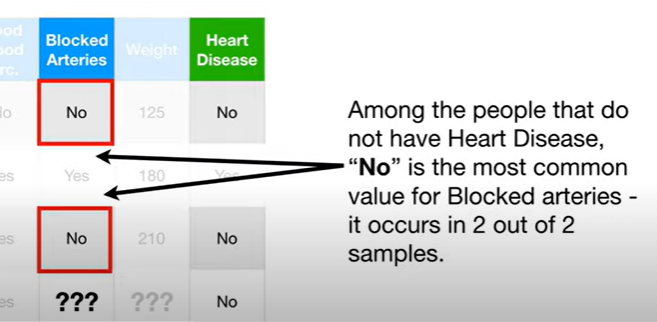
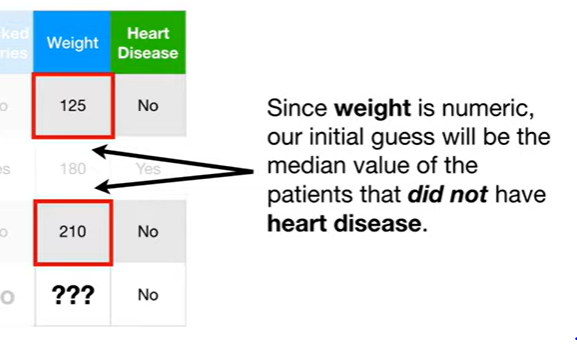
2 TYPES OF PROBLEM STATEMENT:



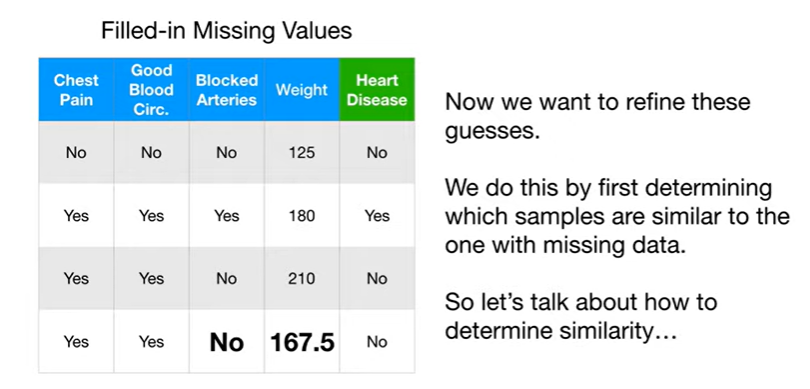
1. MISSING DATA IN THE ORIGINAL DATASET TO BE DETERMINED BEFORE WE CAN USE R.F. TO MAKE CLASSIFICATION:



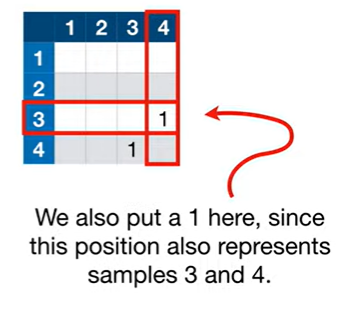
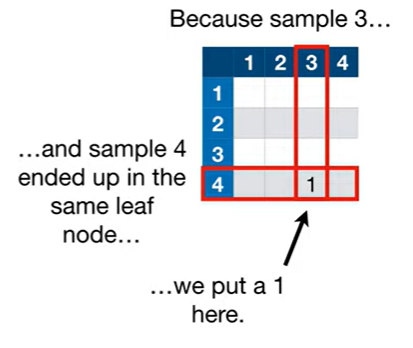
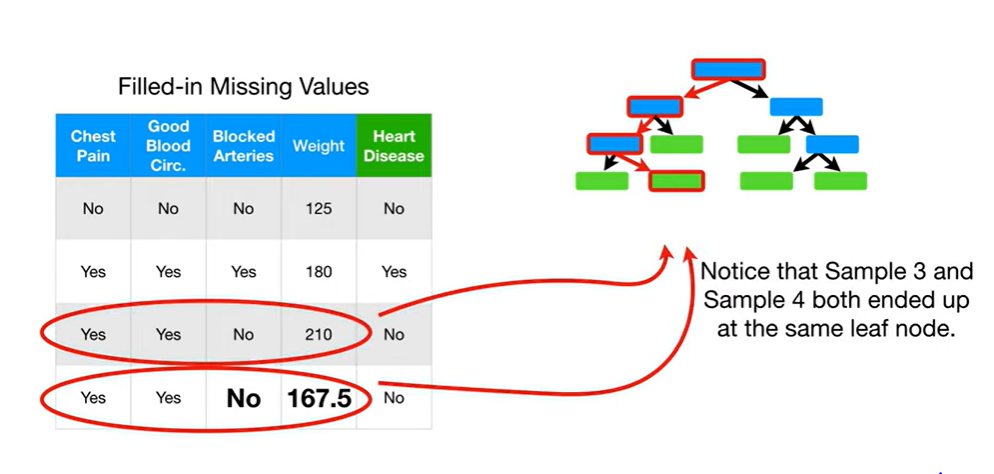
1. Start with a initial (usually a Bad) guess. It is just the **most common value** for ‘Blocked Art’ from the other entries in the **dataset which do NOT have Heart Disease**, which is ‘NO’ here for both the cases (see fig 1 below), and, for ‘Weight’ missing numeric value, it is the Median of the values of the patients **that did NOT have Heart disease (=median of 125, 210 which is 167.5)** (see fig 2 below).

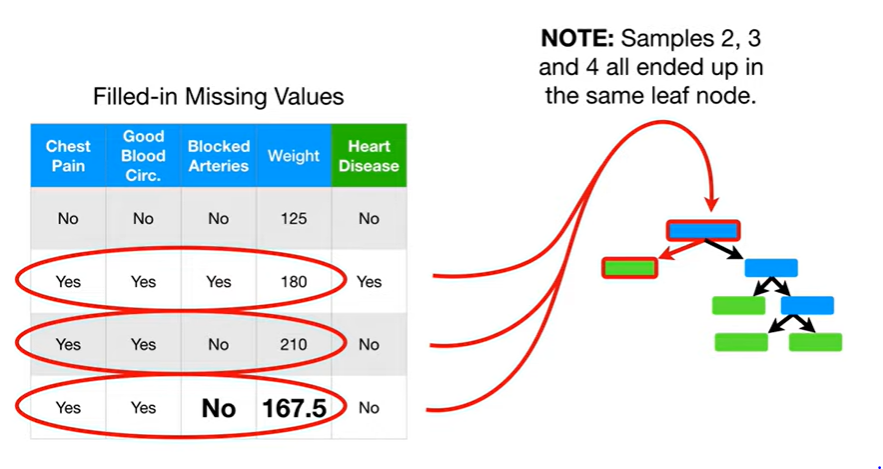
So, our Initial guess is



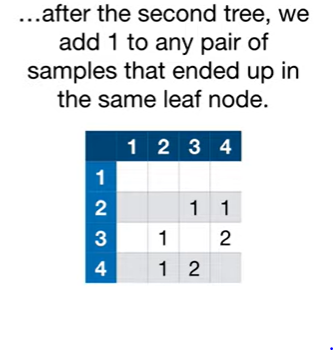
1. Now, refine these Initial Guesses.
2. Build a Random Forest, which may contain 100s of trees.
3. Run all of these Data from above dataset through all these trees.
4. Lets say we are running the data ‘row by row’ down the 1st tree. And we find that the 4th row(with Missing data and initial guesses) ends up **the same Leaf** as the 3rd row data. That means they are ‘SIMILAR’.



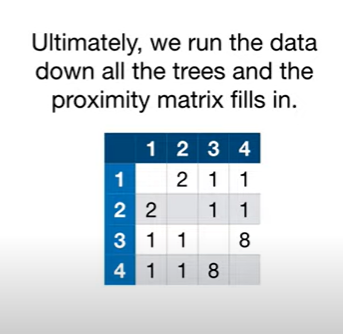
1. We keep the track of this Similarity for all other Trees as well, using a Proximity Matrix. A Proximity Matrix has a Row and Column for each sample/entry/row of data. Since here we have 4 rows/entries , thus Proximity matrix is 4X4.
2. Now, we run the Data down the 2nd tree. Now, the sample/row 2nd and 3rd end up at the same leaf as the 4th row (the initiall,y guessed row), we say they are similar .



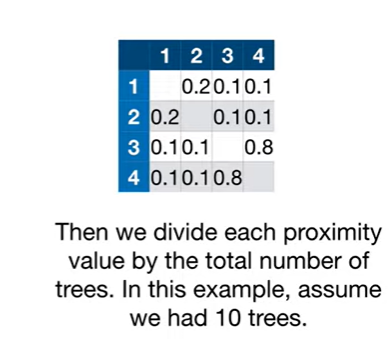
1. We now, update the Proximity matrix by simply adding 1 to the pairs which are similar i.e adding 1 to (2,3 ), (3,2), (2,4), (4,2), (3,4), (4,3).



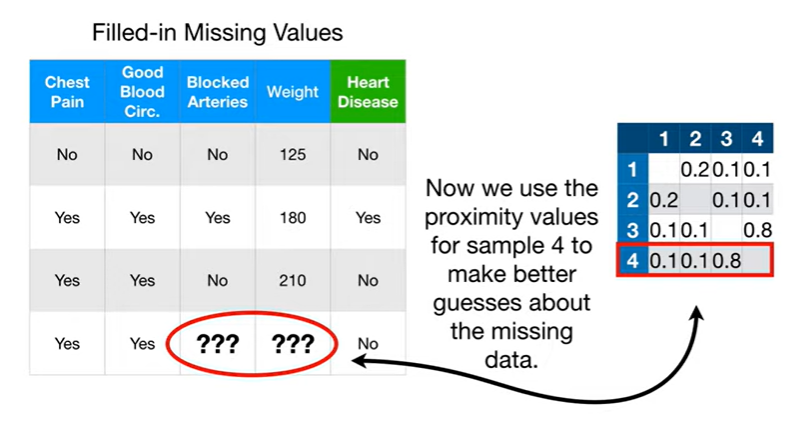
1. After running down the data through all the trees( say there were 10 trees in total), we have



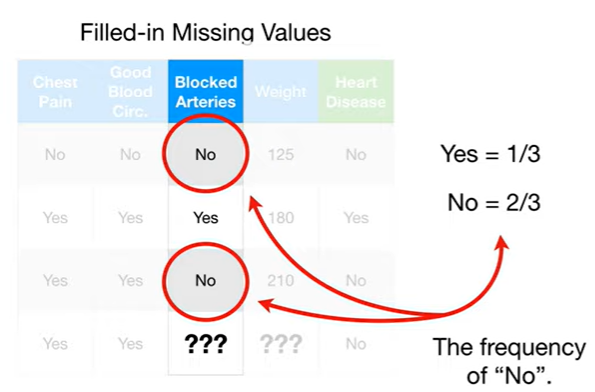




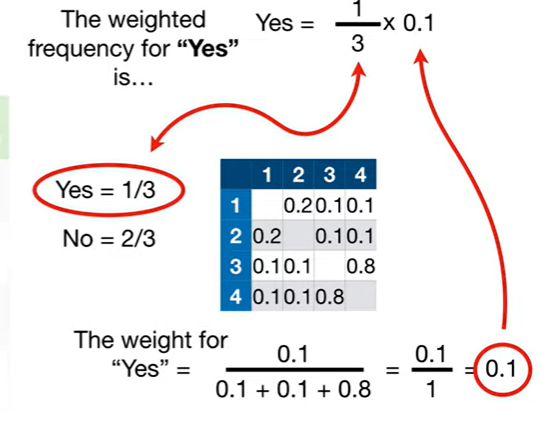
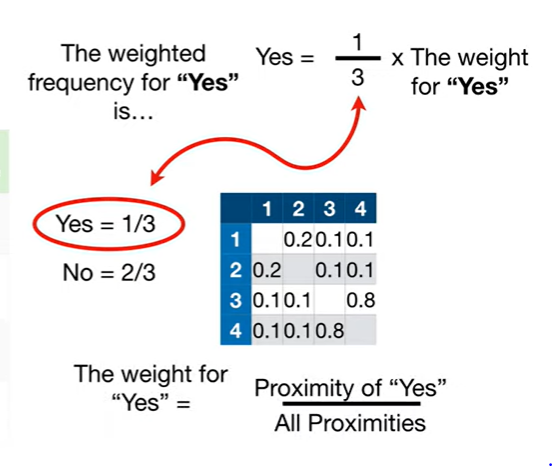
1. Now use these Proximities values for sample4 or 4th row to better estimate the originally missing values.



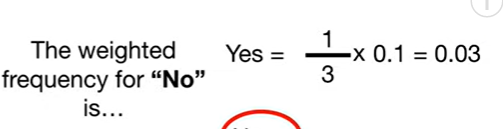
1. To determine “Yes” or “No” for Blocked Art. missing value, we have to calculate “The Weighted frequency” for both “Yes” and “No”. Whichever has more weighted frequency, will be chosen as the final entry for missing value for Blocked art.



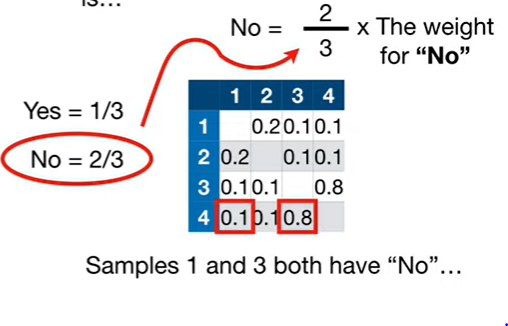
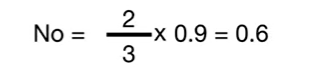
1. Calculation for weighted freq for o”YES”:



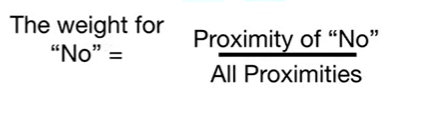
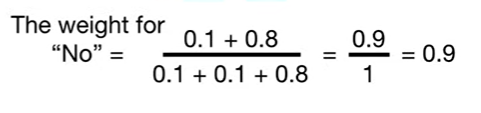
Thus………….



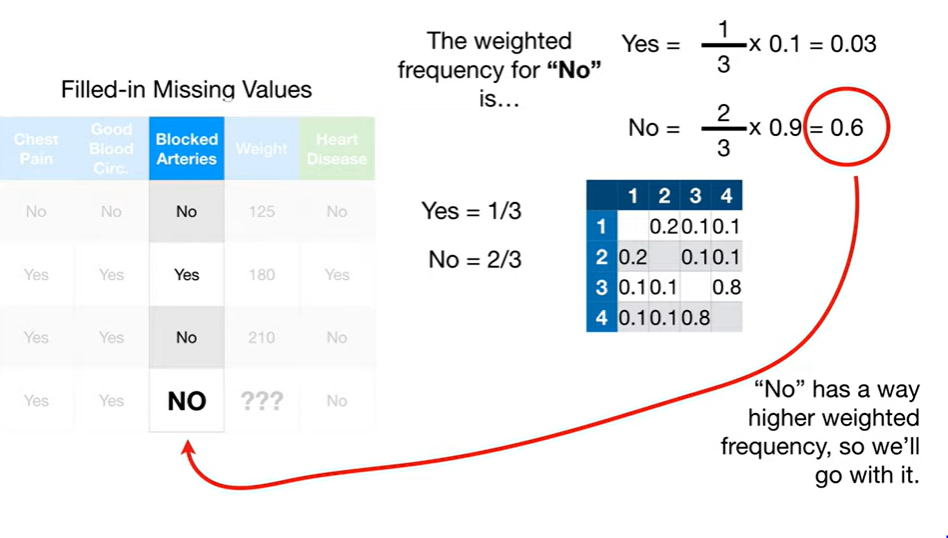
1. Calculation for Weighted freq for “ NO” :

 THUS, 

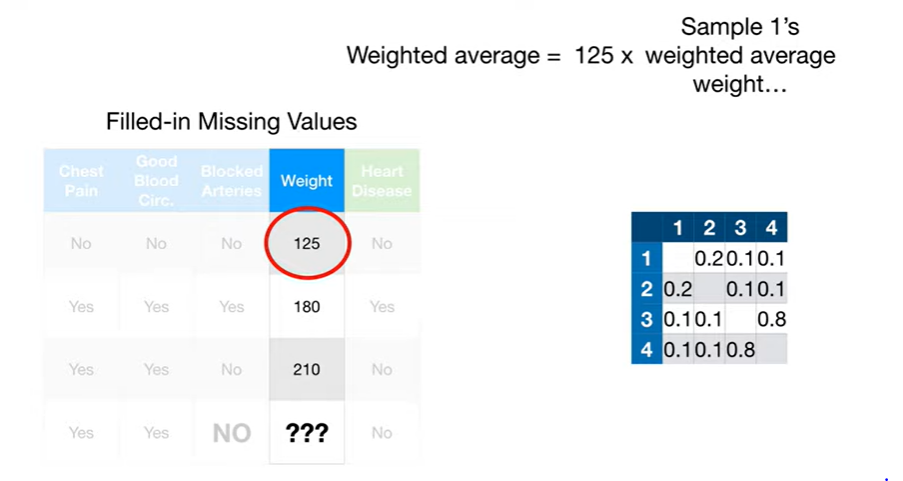
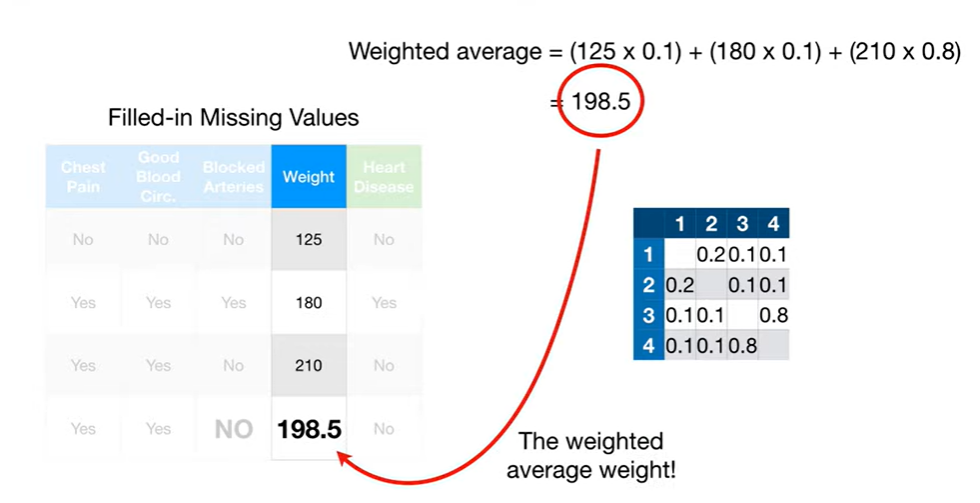
NOTE:

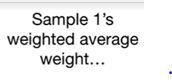
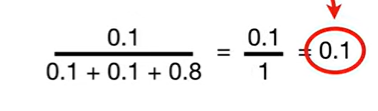
1. Since, “NO” has a much higher Weighted frq value of 0.6 (>0.03), our final entry for missing value for Blocked Art is NO.



1. Slly, we calculate “Weighted Average” numeric value for all the Samples/rows to calculate the “Weight” missing value.

NOTE:

Thus, our final guess for Missing value for “Weight” in 4th row is 198.5 .

1. However, “NO” and 198.5 are not our Absolute final enteries as we need to rebuild 6-7 new R.F. and do these whole exercise again with these new enteries as input to the next R.F. This done untill the “Missing values converge i.e. no longer changes with new set of trials” .

