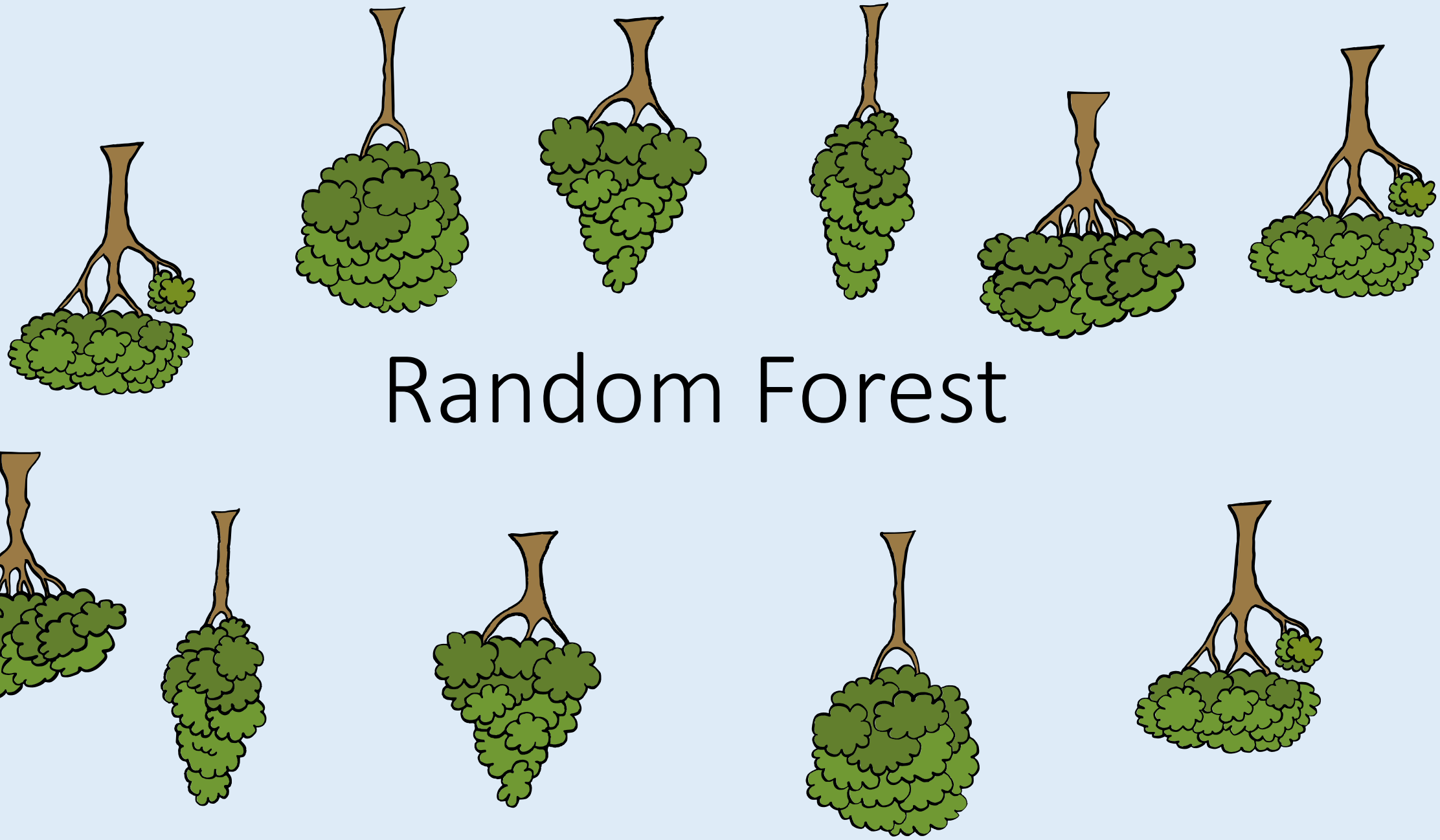

Random Forest

Step 1: Create Bootstrap Dataset

Randomly select samples from main dataset

We are allowed to pick same sample twice

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

Original Dataset

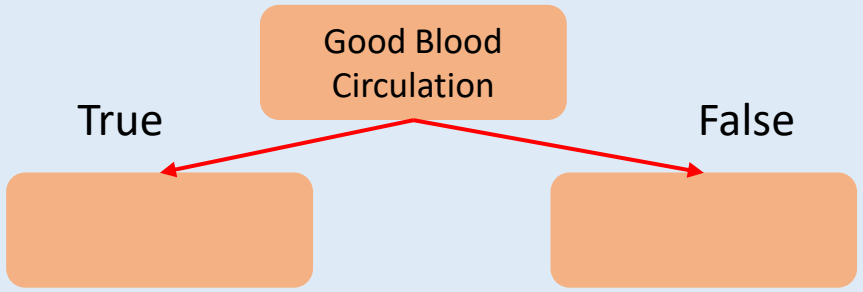
Chest Pain	Good Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	**Yes**	**Yes**	**180**	**Yes**
No	**No**	**No**	**125**	**No**
Yes	**No**	**Yes**	**167**	**Yes**
Yes	**No**	**Yes**	**167**	**Yes**

Bootstrapped Dataset

Step 2: Create Decision Tree using Bootstrapped Dataset using Random Subset of Variables at each step

Here we are going to select Good Blood circulation and Blocked Arteries Randomly

Let's say out of two Good Blood Circulation has least Gini Impurity



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

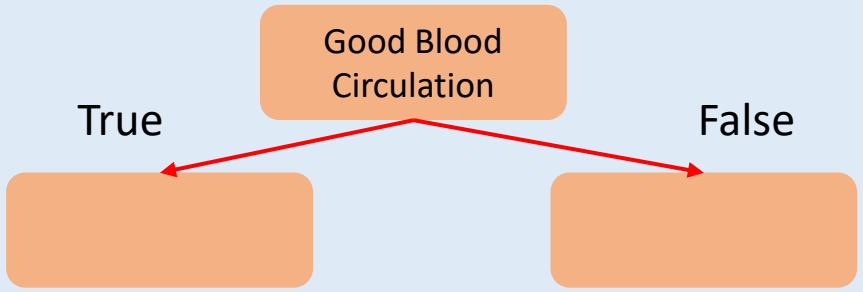
Bootstrapped Dataset

Here, the blood circulation will form a root node and data will be split in two branches, one for true and other for false

What should be the next spitting criterion for True and False branch?

Let's say we have to select the splitting criterion for True, again for that we will select two features out of remaining three randomly. Let's assume two selected variables are Chest Pain and Blocked arteries

Step 2: Create Decision Tree using Bootstrapped Dataset using Random Subset of Variables at each step



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

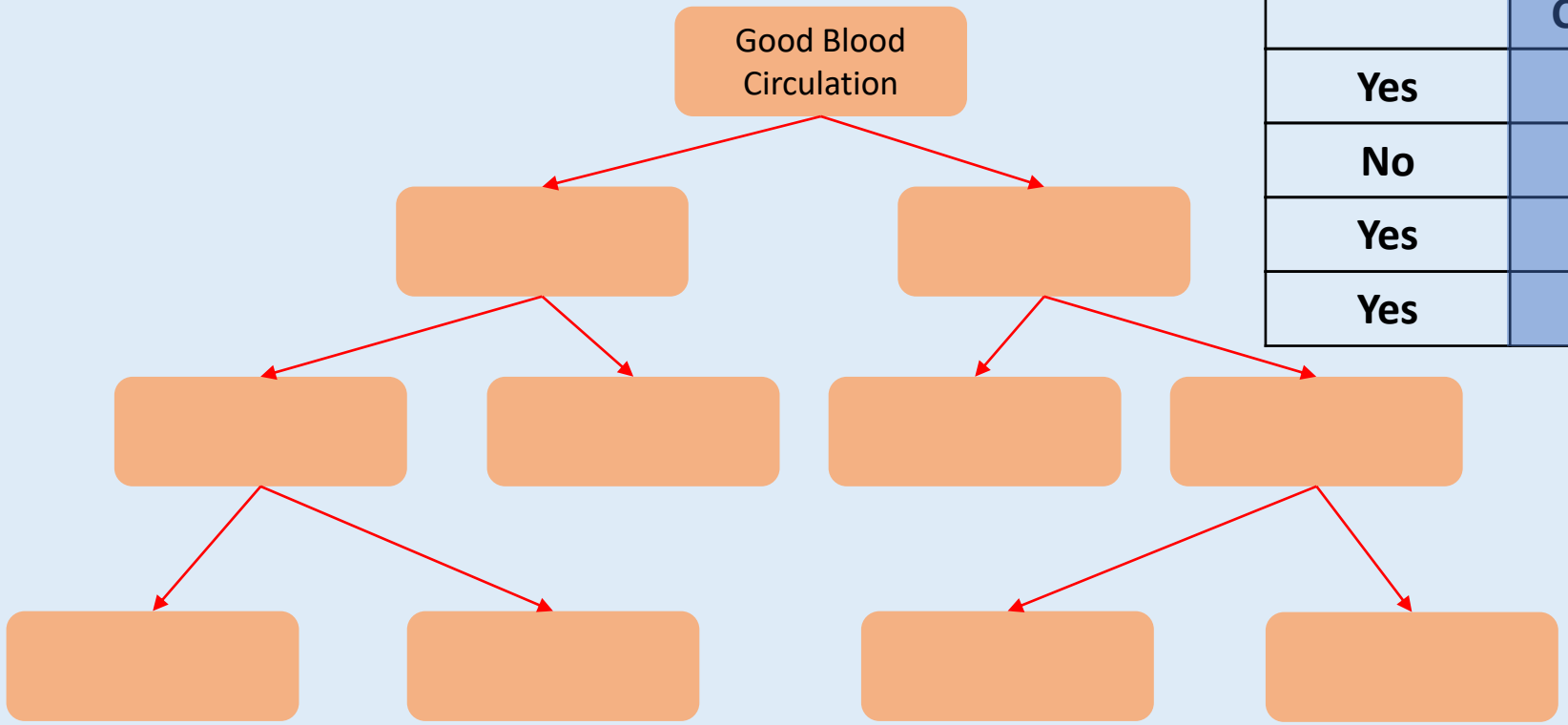
Bootstrapped Dataset

Out of these two selected variables i.e. Chest Pain and Blocked Arteries, we will figure which one has low Gini impurity

The variable with low Gini impurity will be the next splitting criterion for the True branch

Same process will be followed to find the next splitting criterion of False branch, i.e. randomly selecting two variables and finding the one with low Gini impurity. This is how the tree will be grown by selecting random subset of variables at each step

Step 2: Create Decision Tree using Bootstrapped Dataset using Random Subset of Variables at each step



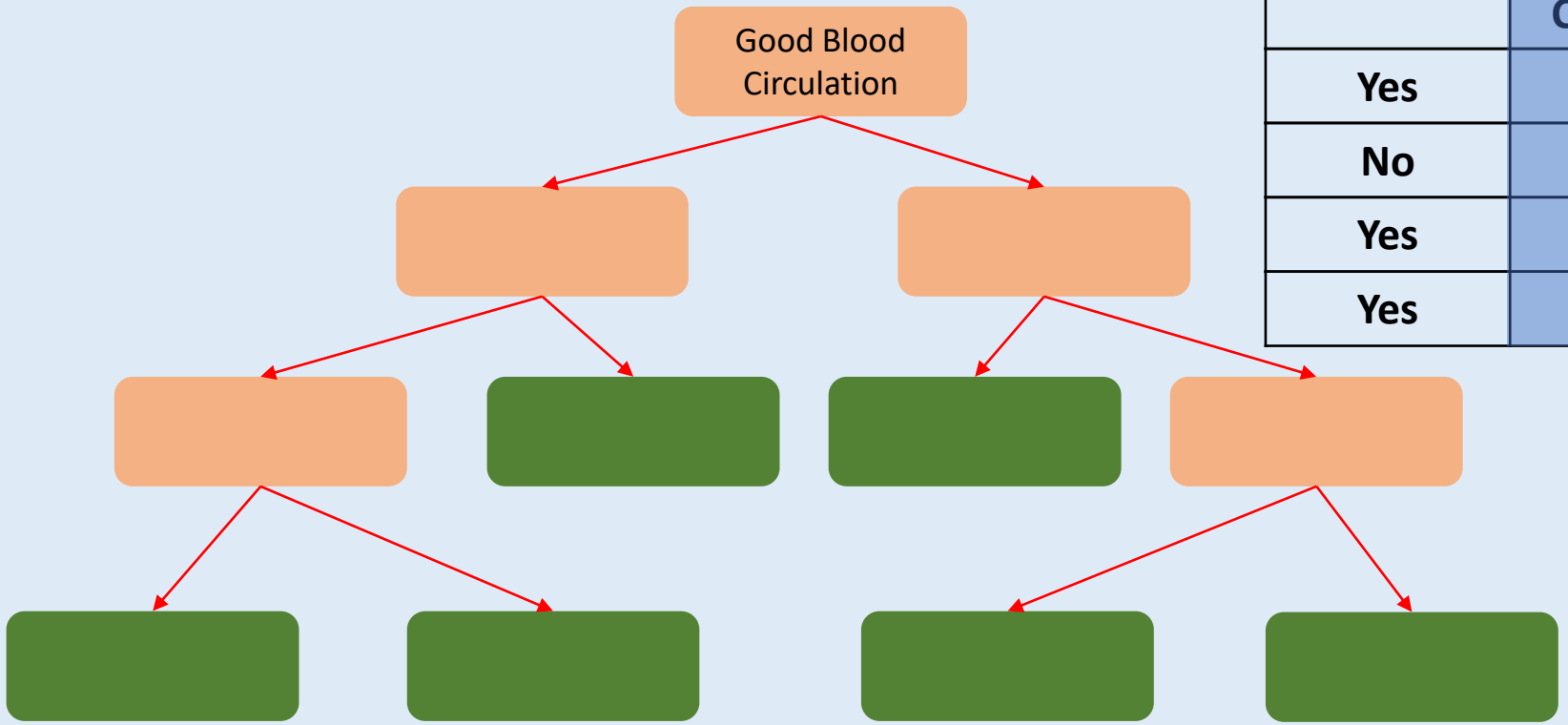
Chest Pain	Good Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

Bootstrapped Dataset

Step 2: Create Decision Tree using Bootstrapped Dataset using Random Subset of Variables at each step

Here we are going to select Good Blood circulation and Blocked Arteries Randomly

Let's say out of two Good Blood Circulation has least Gini Impurity

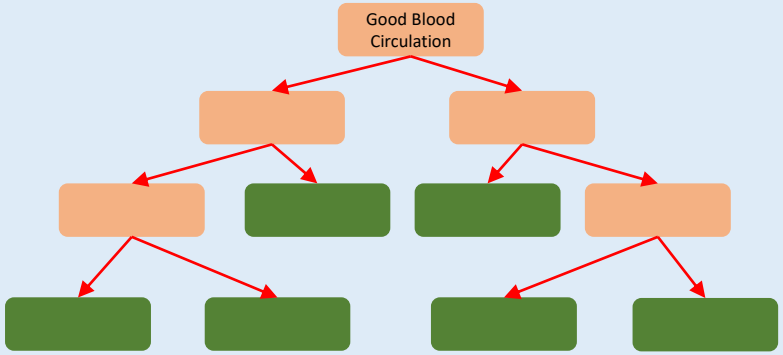


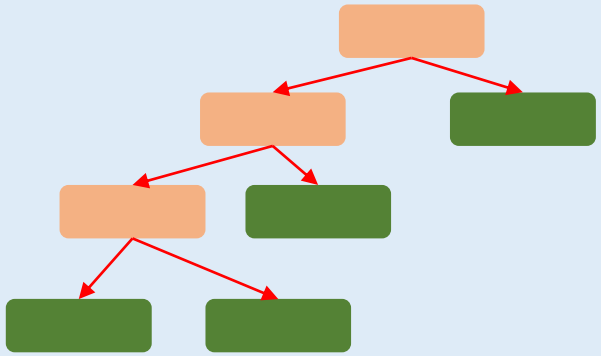
Chest Pain	Good Blood Circulation	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

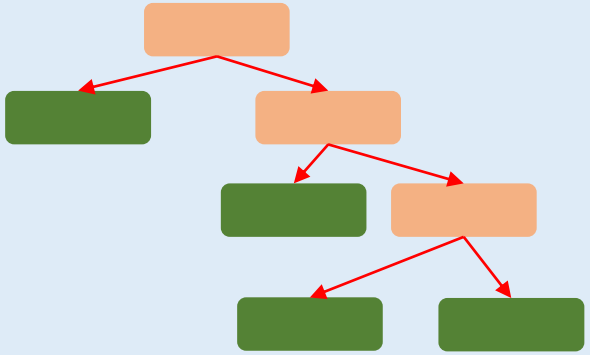
Bootstrapped Dataset

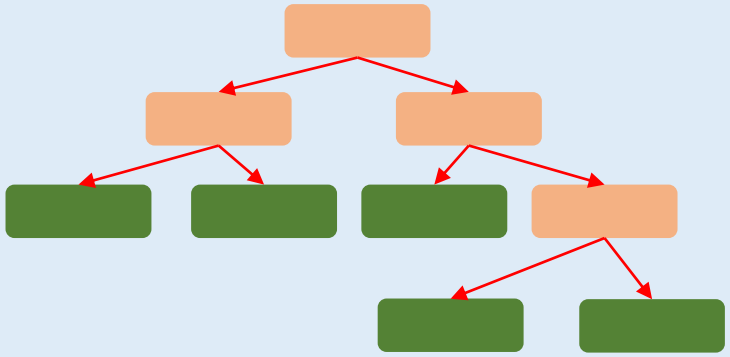
Step 3: Keep Repeating step 1 and 2

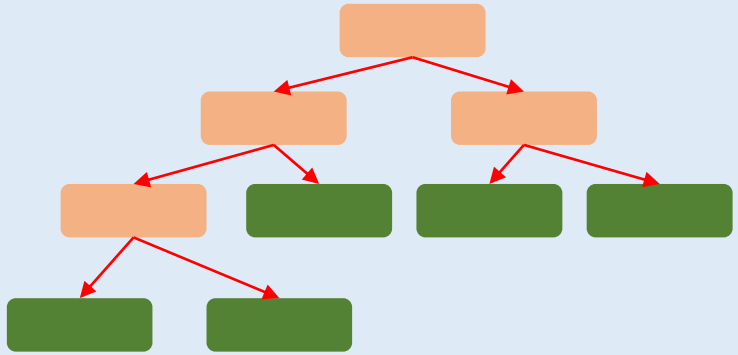
- Keep repeating step 1 and step 2
- That is, prepare bootstrap dataset by selecting the random rows
- And then build tree by selecting random variable at each step
- This will be done 100s of time
- This leads to formation of 100s of wide variety of trees
- This variety makes random forest better than single decision tree

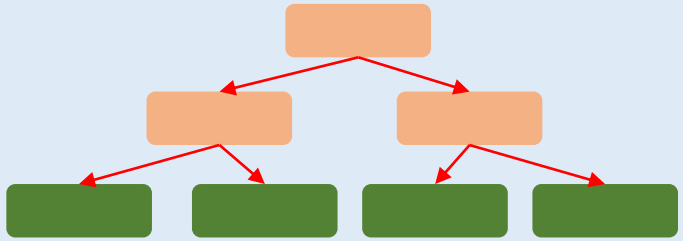

Good Blood
Circulation











Step 3: Keep Repeating step 1 and 2

- This variety makes random forest better than single decision tree
- Once these tree are ready, an unknown datapoint will be sent to run down all the trees and we will have the classification problem solved by each tree
- Here each tree of random forest might give us different class outcome for the given input datapoint
- Now question is which class should be considered as a final outcome?
- For this voting is carried out, for example, there is a binary classification problem, for this problem assume there are 100 trees in a random forest
- Let's say, 52 trees say the class for the given input datapoint is 0 and 48 trees say the class is 1, in this case the final outcome would be class 0 as it has more number of trees speaking in its favour
- This step of taking voting is called aggregating
- This is how the random forest got it's name of being the Bagging technique, i.e. Bootstrapping + Aggregating = Bagging

Assessment of Random Forest

- When the bootstrapped datasets are formed, there are chances that few of the sample rows never get selected in bootstrapped dataset
- Such samples are called as **OUT OF BAG SAMPLE**
- Now such samples are used for assessment of the Random Forest

For Regression

- In case of regression, 100s of regression trees are created by selecting random variable at each step of tree formation
- Now each of this tree will give different outcome for the regression problem
- Here, the final outcome is decided by taking average of outcomes given by all 100 trees