Random Forest

how it works

The traditional decision tree is good at dealing with the data used to create it, but not good for any new data. The good thing is that “random forest” is as easy as the traditional decision tree, and has a vast improvement in terms of accuracy

So how can we create a “random forest”

So how can we create a “random forest”

Step 1 create a bootstrap dataset

First step is that we need to create an bootstrapped dataset from the original dataset

Original dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes

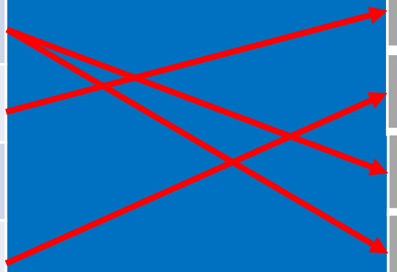
So how can we create a “random forest”

Step 1 create a bootstrap dataset

First step is that we need to create an bootstrapped dataset from the original dataset

Original dataset

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes



Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example in this case, the first row gets selected twice, while the 3rd and last row are not selected at all

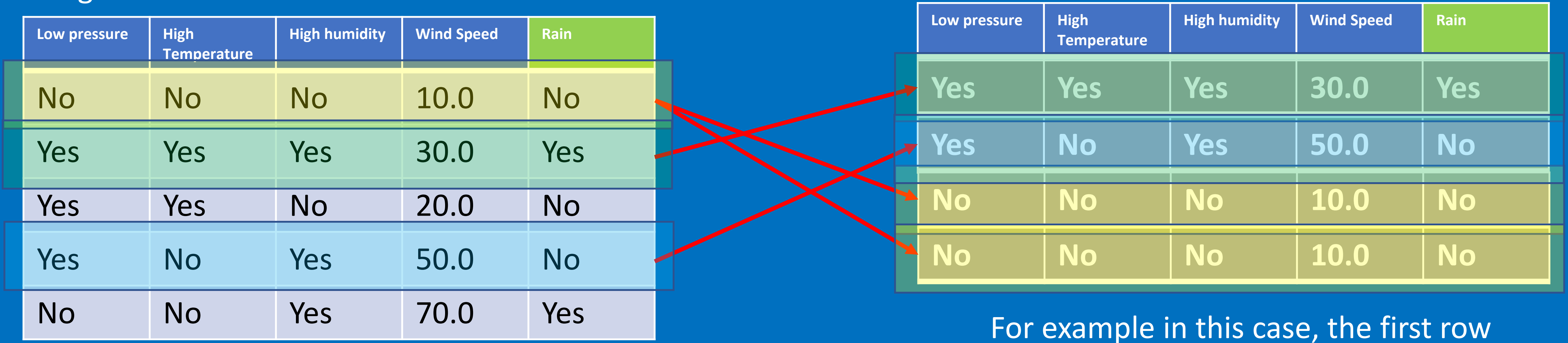
To create a bootstrap dataset with the same predictors as the original dataset, we just randomly select samples here

So how can we create a “random forest”

Step 1 create a bootstrap dataset

First step is that we need to create an bootstrapped dataset from the original dataset

Original dataset



Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes

Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

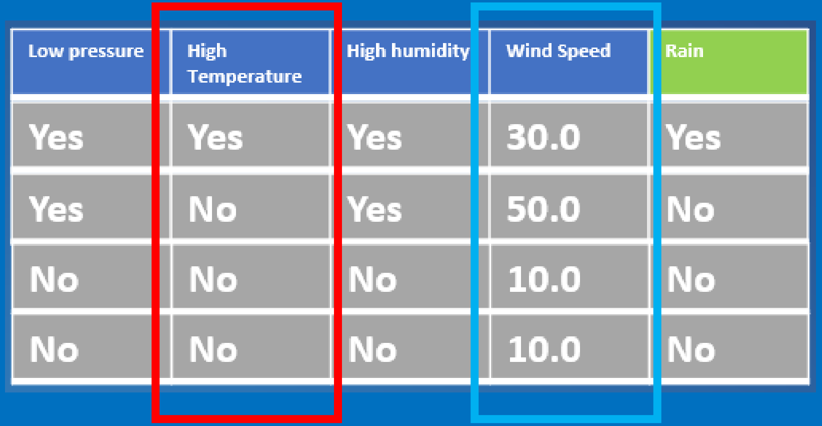
To create a bootstrap dataset with the same predictors as the original dataset, we just randomly select samples here

For example in this case, the first row gets selected twice, while the 3rd and last row are not selected at all (we are allowed to pick the same sample more than once …)

So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors

Bootstrapped dataset



Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

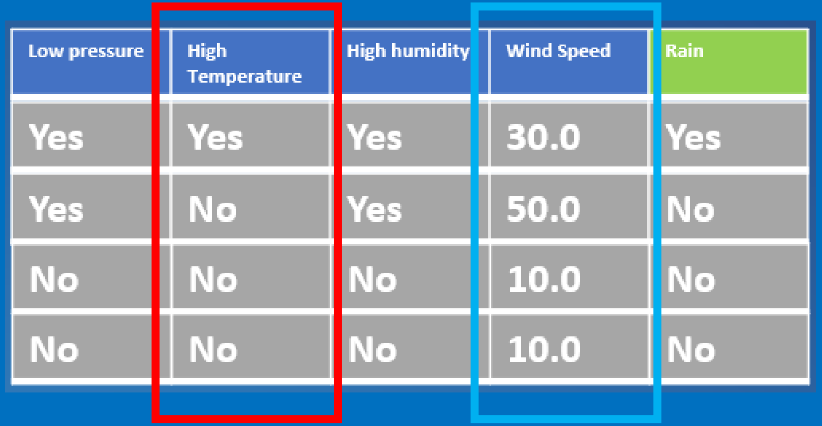
For example, instead of considering all four predictors to figure out which one should be the “root” node, we only consider two here ~ “high temp” and “wind speed” (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors

Bootstrapped dataset

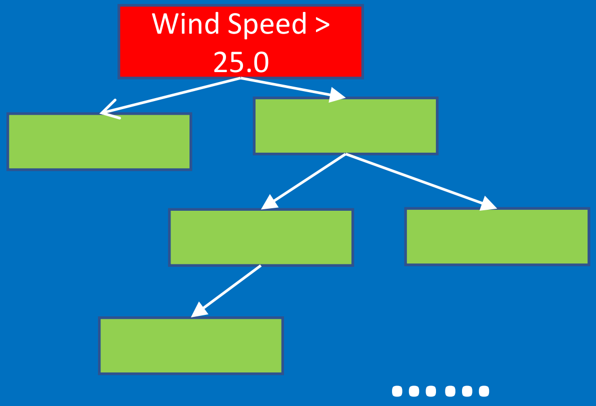


Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the “root” node, we only consider two here ~ “high temp” and “wind speed” (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

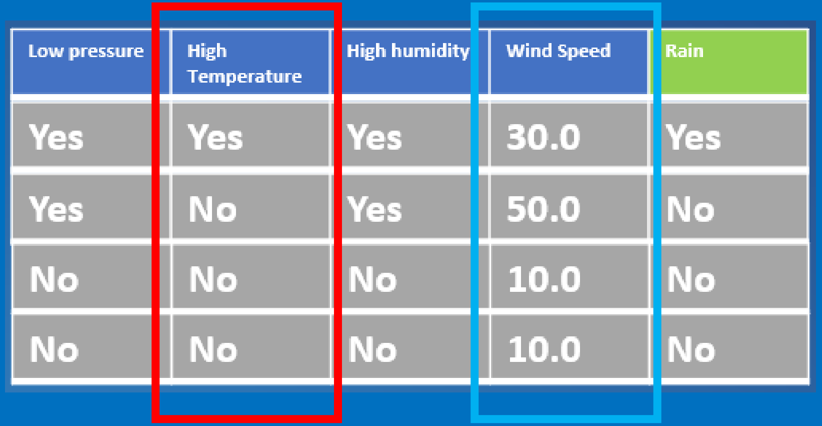
If “Wind Speed > 25” has the smallest Gini and is selected as the root node, we can construct the tree as



So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors

Bootstrapped dataset

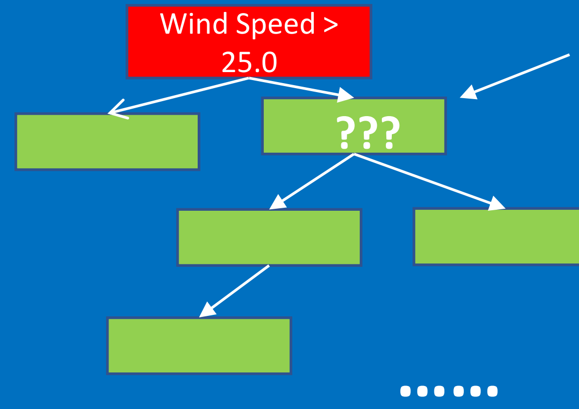


Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

For example, instead of considering all four predictors to figure out which one should be the “root” node, we only consider two here ~ “high temp” and “wind speed” (these are selected randomly)

High Temperature	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

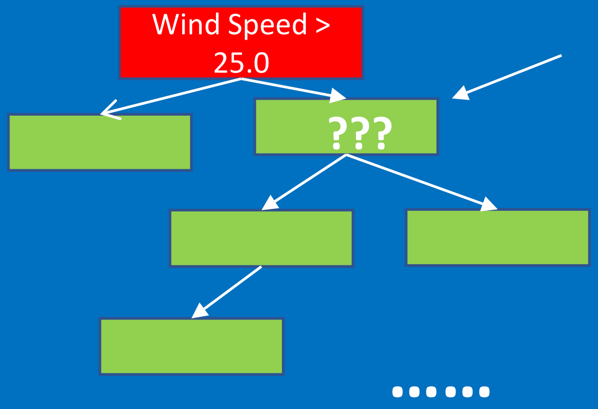
If “Wind Speed > 25” has the smallest Gini and is selected as the root node, we can construct the tree as



Next step we need to figure out which predictor is to be used for the next level of tree

So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors



Next step we need to figure out which predictor is to be used for the next level of tree

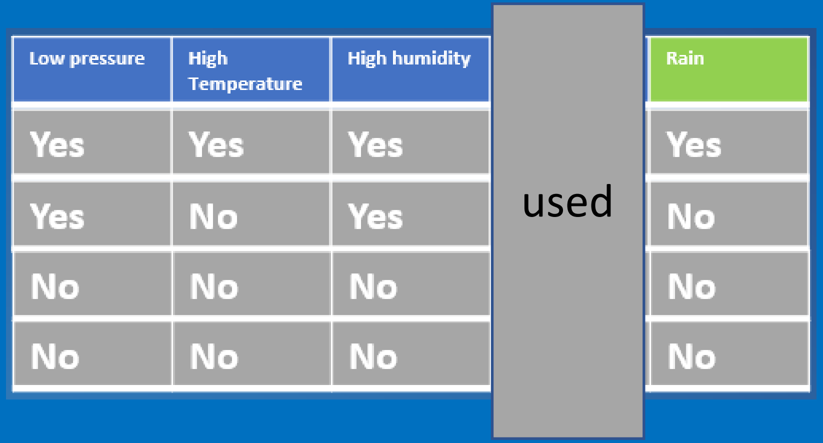
Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

Since the wind speed is chose, so we have 3 remaining predictors can be considered

Low pressure	High Temperature	High humidity	used	Rain
Yes	Yes	Yes		Yes
Yes	No	Yes		No
No	No	No		No
No	No	No		No

So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors

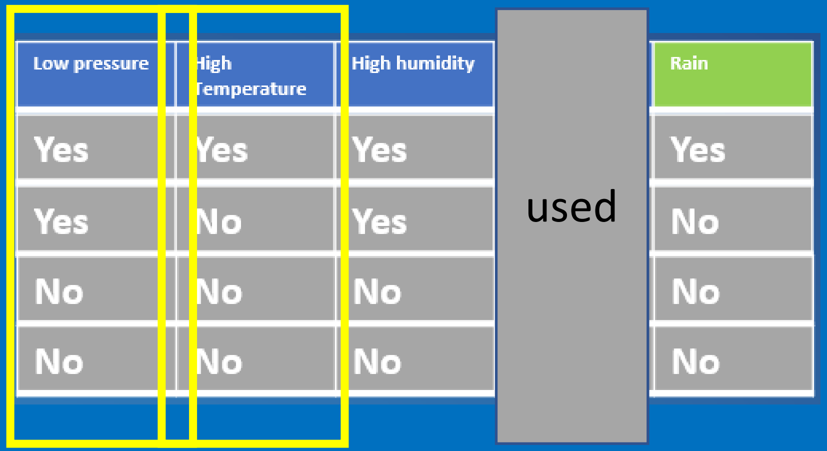


Low pressure	High Temperature	High humidity	used	Rain
Yes	Yes	Yes		Yes
Yes	No	Yes		No
No	No	No		No
No	No	No		No

Then we randomly select two predictors from the remaining predictors, and build the tree as usual

So how can we create a “random forest”

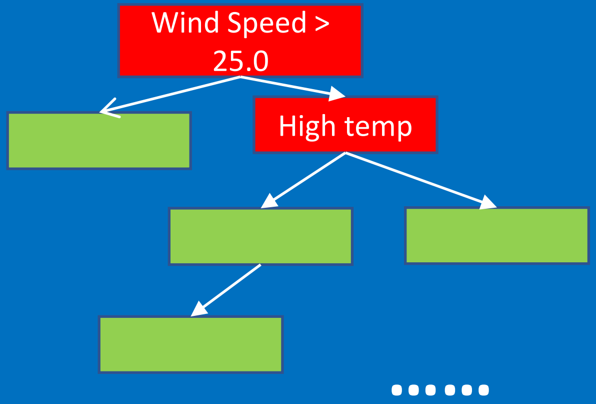
Step 2 create decision tree using the bootstrapped dataset with only subset of predictors



Low pressure	High Temperature	High humidity	used	Rain
Yes	Yes	Yes		Yes
Yes	No	Yes		No
No	No	No		No
No	No	No		No

Then we randomly select two predictors from the remaining predictors, and build the tree as usual

For example, we if we randomly select low pressure and high temp, and high temp has a smaller Gini, then the next level of leaf will be built based on “high temp”



So how can we create a “random forest”

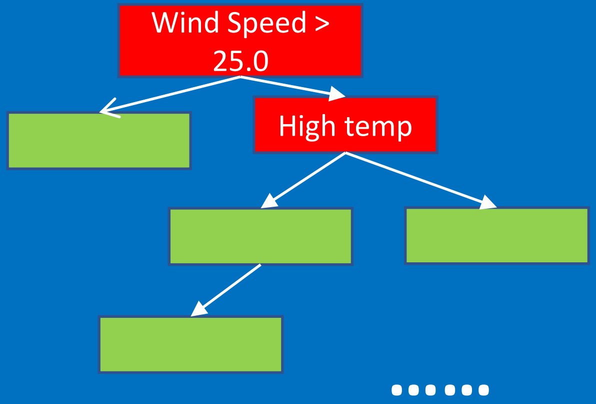
Step 2 create decision tree using the bootstrapped dataset with only subset of predictors

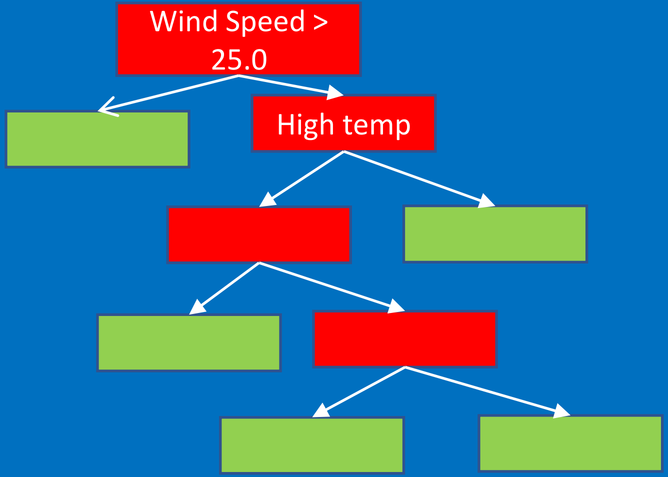
Low pressure	High Temperature	High humidity	used	Rain
Yes	Yes	Yes		Yes
Yes	No	Yes		No
No	No	No		No
No	No	No		No

Then we randomly select two predictors from the remaining predictors, and build the tree as usual

So we just build the tree iteratively like this, until we are not able to split the tree anymore !

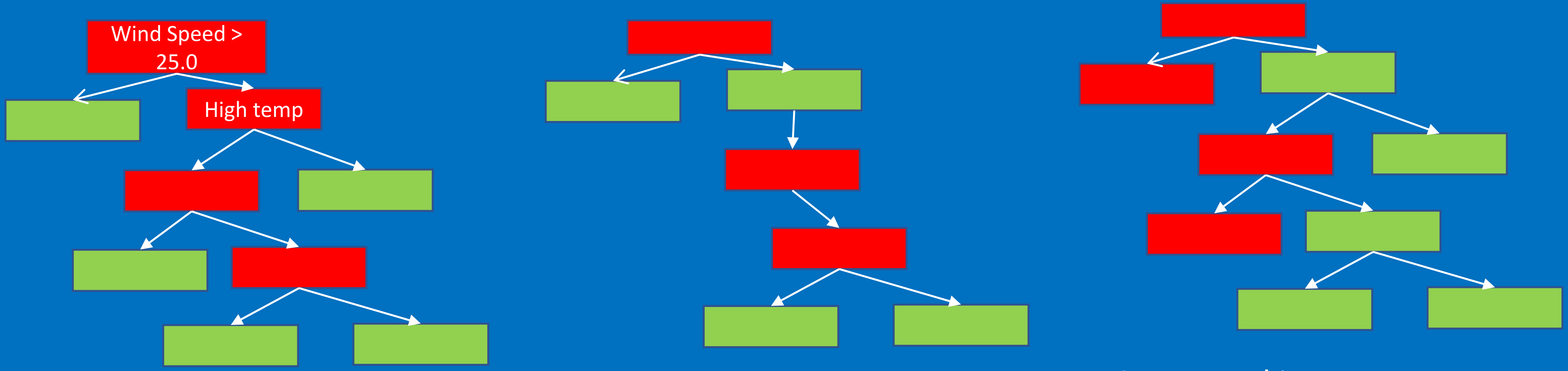
For example, we if we randomly select low pressure and high temp, and high temp has a smaller Gini, then the next level of leaf will be built based on “high temp”

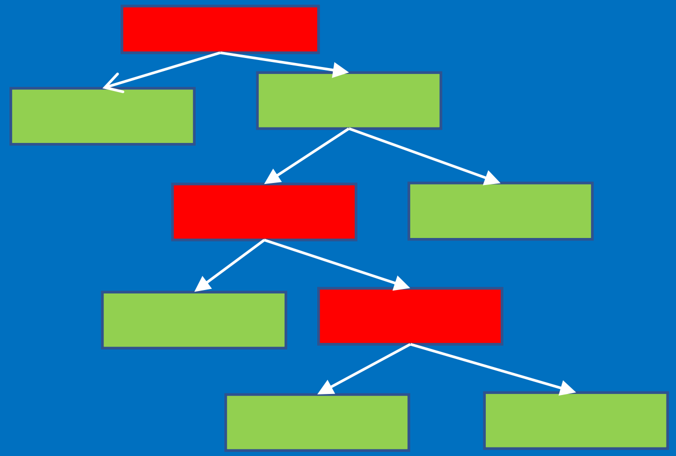




So how can we create a “random forest”

Step 2 create decision tree using the bootstrapped dataset with only subset of predictors





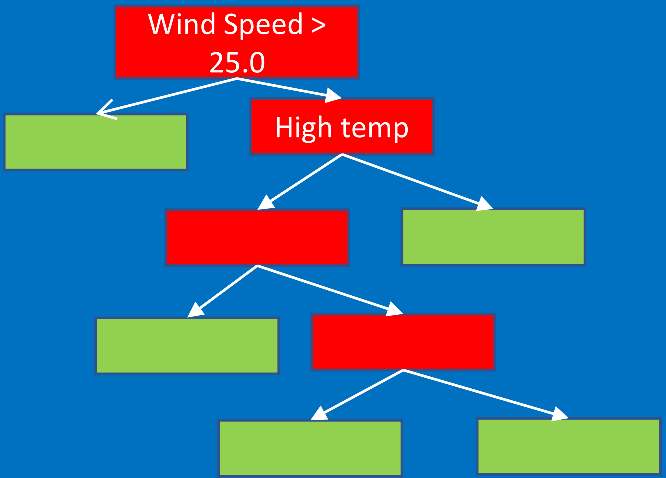
Then we just use the same technology to build many independent/random trees

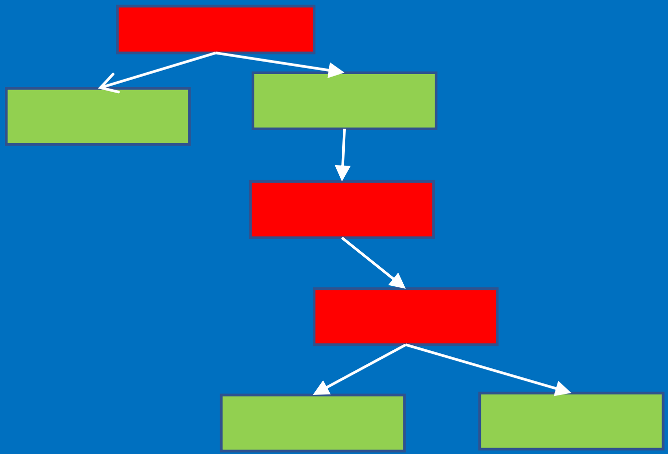
Ideally, you do this 100’s times

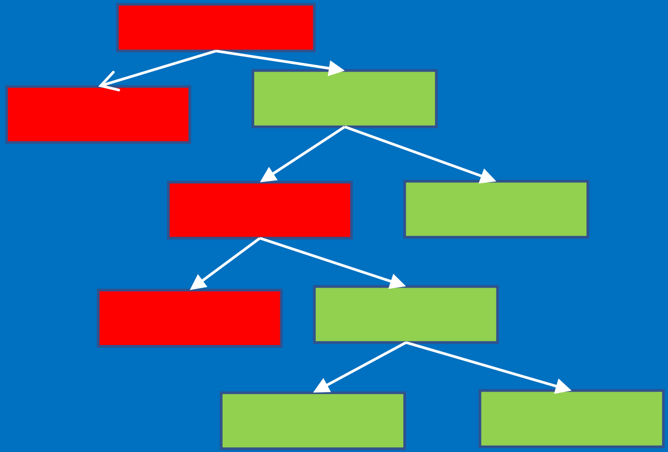
<u>Step 1</u>:Making a new bootstrapped dataset
<u>Step 2</u>: select random predictors from the bootstrapped data and make trees

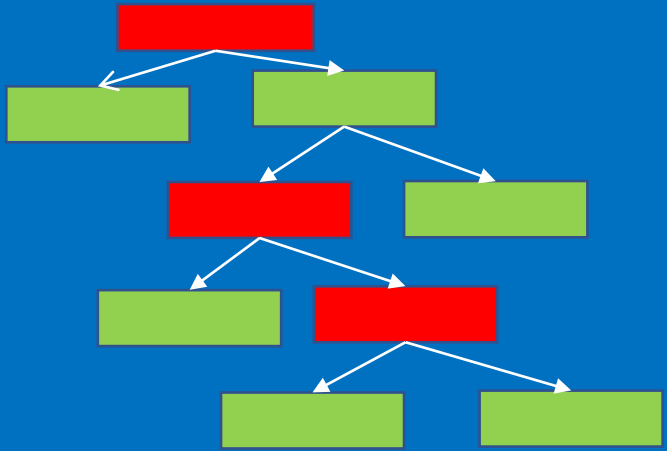
So how can we create a “random forest”

Step 3 Create many “random” trees









Then we just use the same technology to build many independent/random trees

Step 1:Making a new bootstrapped dataset
Step 2: select random predictors from the bootstrapped data and make trees

- Ideally, you do this 100’s times
- By doing this we will have a variety of trees, and the variety is what makes RF so robust

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The second tree says “YES”

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The second tree says “YES”
- So we take the test data, run it through the 3rd tree we made ➡ The 3rd tree says “YES”

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The second tree says “YES”
- So we take the test data, run it through the 3rd tree we made ➡ The 3rd tree says “NO”

……

- So we take the test data, run it through the n’rd tree we made ➡ The n’rd tree says “NO”

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The second tree says “YES”
- So we take the test data, run it through the 3rd tree we made ➡ The 3rd tree says “NO”

••••••

- So we take the test data, run it through the n’rd tree we made ➡ The n’rd tree says “NO”

After running the dataset through all the “random” trees, we see which option gets more votes, e.g.,

Rain: YES	Rain: NO
15	3

So how can we create a “random forest”

Step 4 How we use RF

Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The second tree says “YES”
- So we take the test data, run it through the 3rd tree we made ➡ The 3rd tree says “NO”

……

- So we take the test data, run it through the n’rd tree we made ➡ The n’rd tree says “NO”

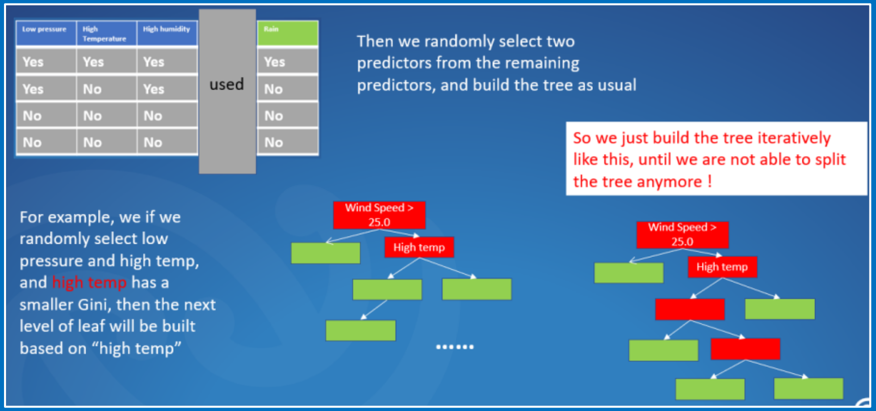
After running the dataset through all the “random” trees, we see which option gets more votes, e.g.,

Rain: YES	Rain: NO
15	3

➡

In this case, Rain: YES gets more votes, so the prediction will be “YES”

So how can we create a “random forest”

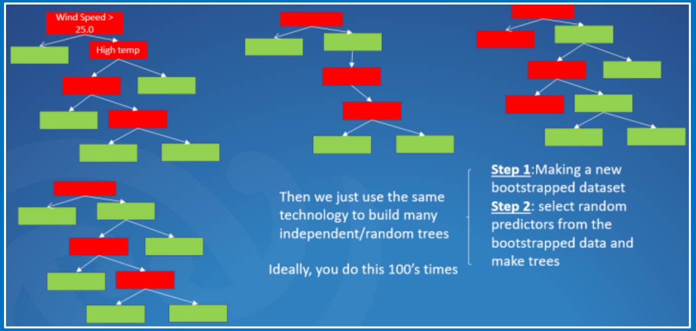


Low pressure	High Temperature	High humidity	used	Rain
Yes	Yes	Yes		Yes
Yes	No	Yes		No
No	No	No		No
No	No	No		No

Then we randomly select two predictors from the remaining predictors, and build the tree as usual

So we just build the tree iteratively like this, until we are not able to split the tree anymore !

For example, we if we randomly select low pressure and high temp, and high temp has a smaller Gini, then the next level of leaf will be built based on “high temp”

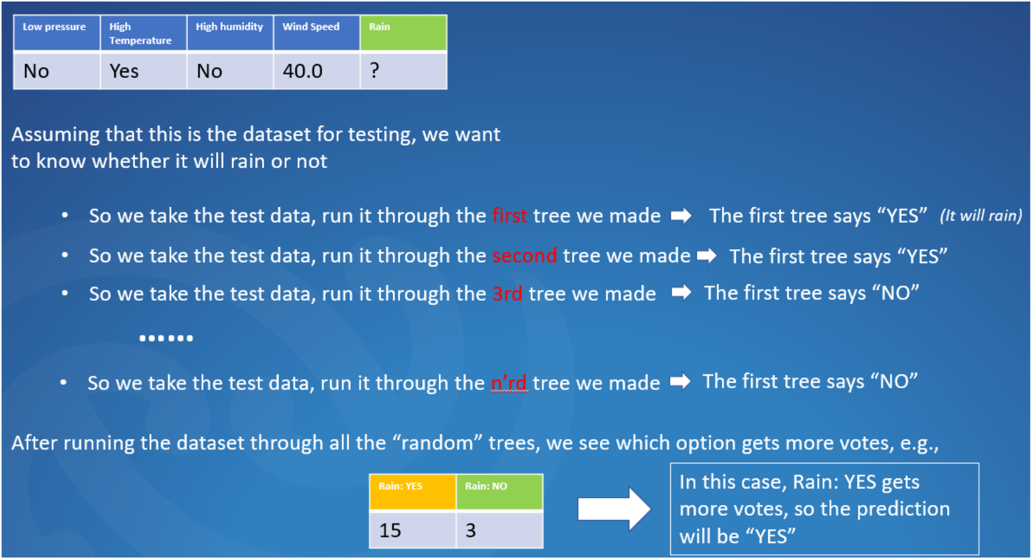


Then we just use the same technology to build many independent/random trees

Ideally, you do this 100’s times

Step 1:Making a new bootstrapped dataset
Step 2: select random predictors from the bootstrapped data and make trees

Bootstrapping the data



Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

Assuming that this is the dataset for testing, we want to know whether it will rain or not

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- So we take the test data, run it through the second tree we made ➡ The first tree says “YES”
- So we take the test data, run it through the 3rd tree we made ➡ The first tree says “NO”

……

- So we take the test data, run it through the n’rd tree we made ➡ The first tree says “NO”

After running the dataset through all the “random” trees, we see which option gets more votes, e.g.,

Rain: YES	Rain: NO
15	3

➡ In this case, Rain: YES gets more votes, so the prediction will be “YES”

Using aggregation to get the decision

Bootstrapping the data + Using AGGregation to get the decision = BAGGING

Decision Tree

Random forest: how to evaluate RF