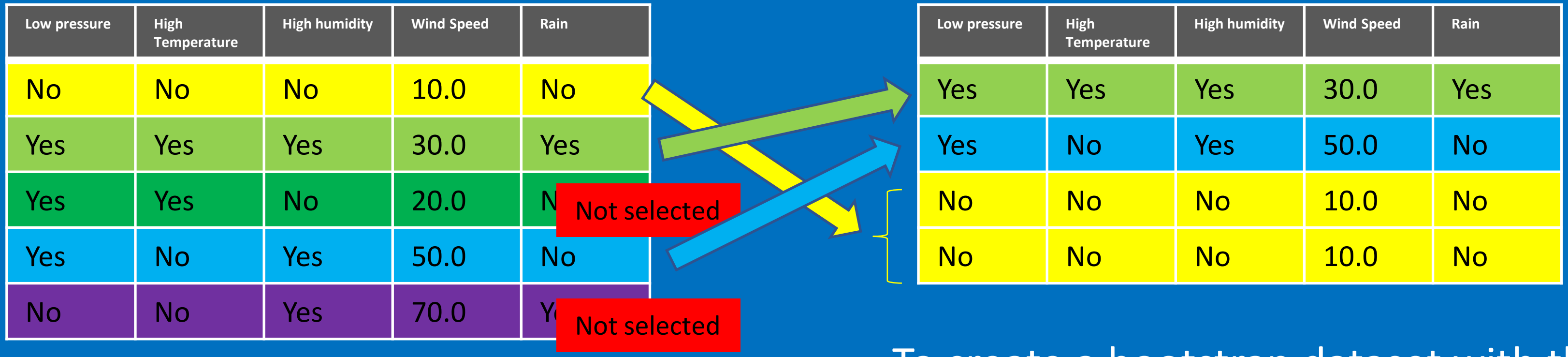
Bagging

Bagging

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

Assuming that we have the above data

Bagging



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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

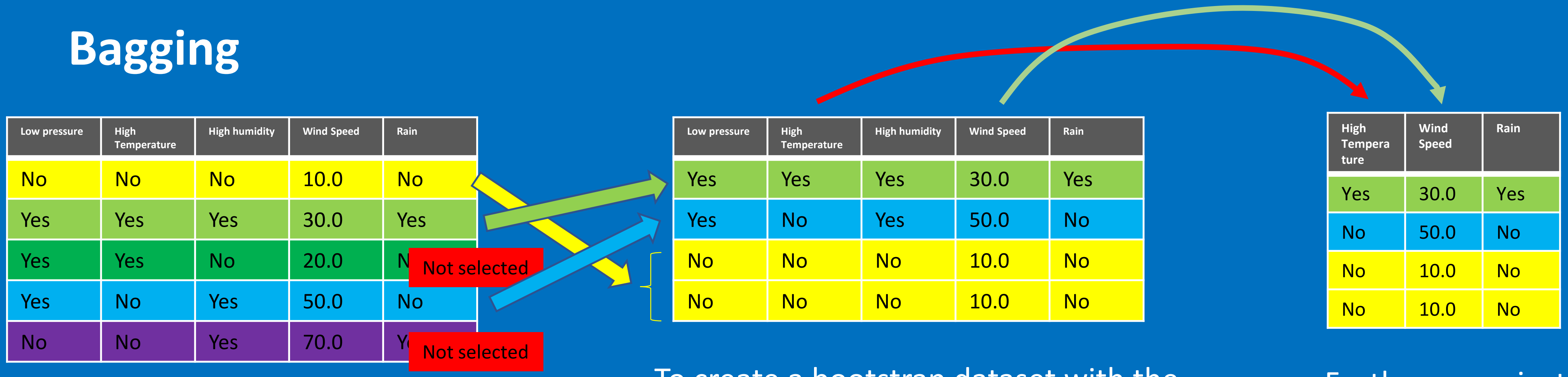
Low pressure	High Temperature	High humidity	Wind Speed	Rain
Yes	Yes	Yes	30.0	Yes
Yes	No	Yes	50.0	No
No	No	No	10.0	No
No	No	No	10.0	No

Assuming that we have the above data

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

Bagging



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	N
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y

Not selected

Not selected

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

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

Assuming that we have the above data

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 3^{rd} and last row are not selected at all

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

Bagging

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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

By “randomly” repeat the above process, we can have many bootstrapped dataset

Bagging

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	N
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y

Not selected

Not selected

Assuming that we have the above data

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

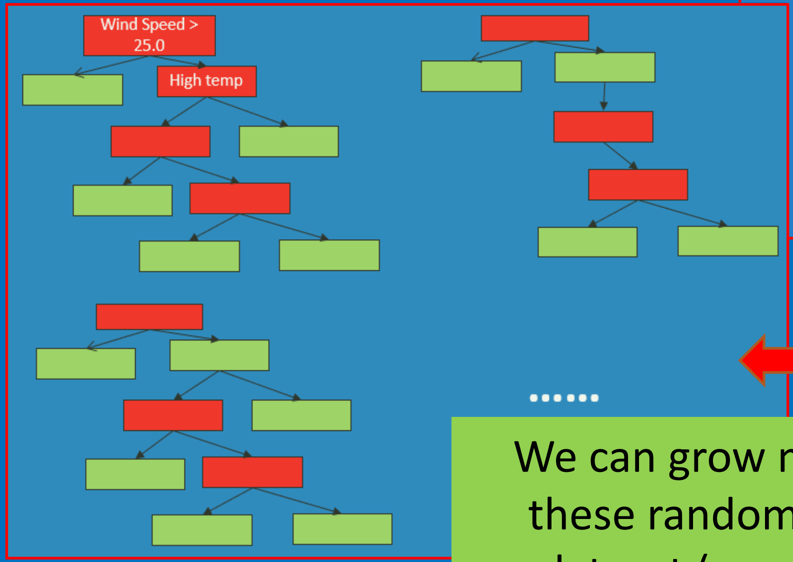
To create a bootstrap dataset with the same predictors as the original dataset, we just randomly select samples here

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

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

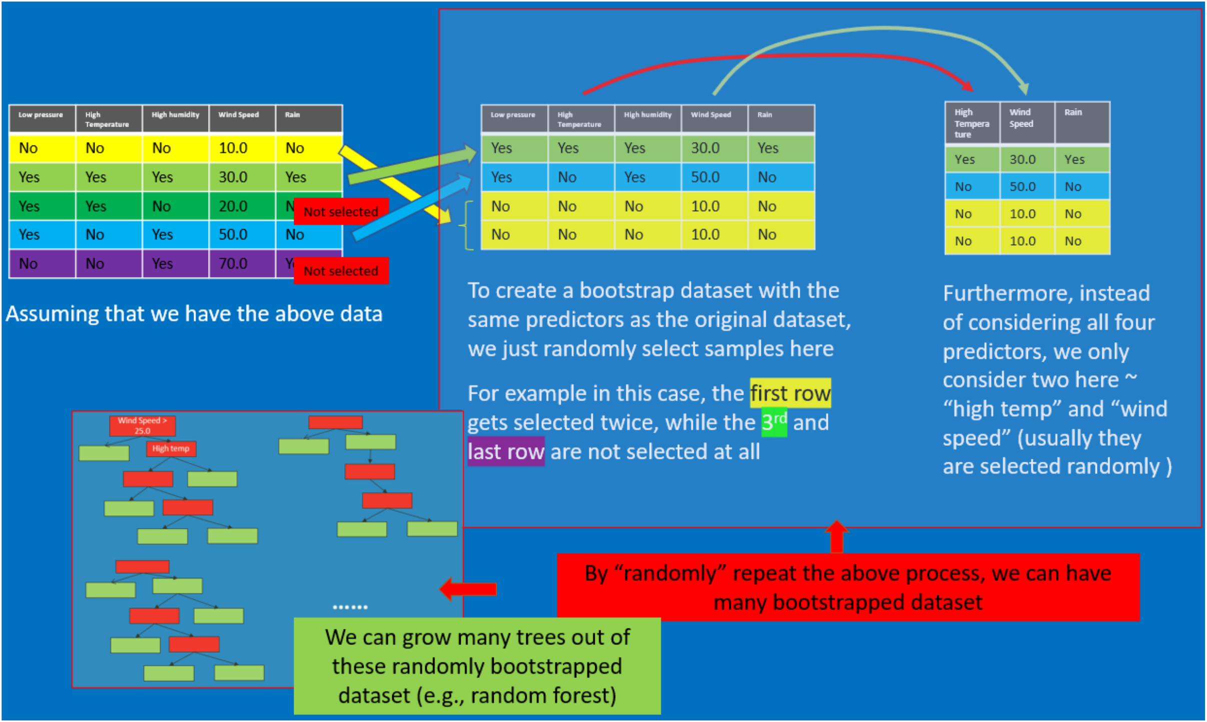
Furthermore, instead of considering all four predictors, we only consider two here ~ "high temp" and "wind speed" (usually they are selected randomly)

By "randomly" repeat the above process, we can have many bootstrapped dataset



We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bagging



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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

By “randomly” repeat the above process, we can have many bootstrapped dataset

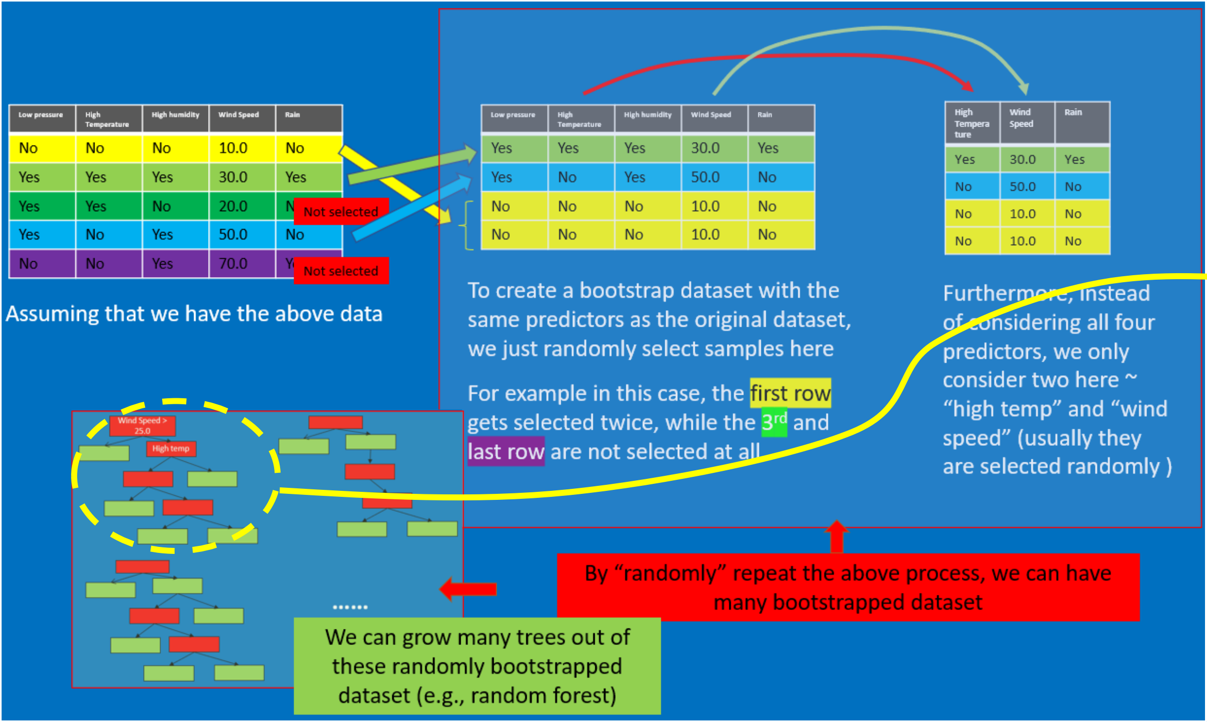
We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

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

Bagging



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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ "high temp" and "wind speed" (usually they are selected randomly)

By "randomly" repeat the above process, we can have many bootstrapped dataset

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

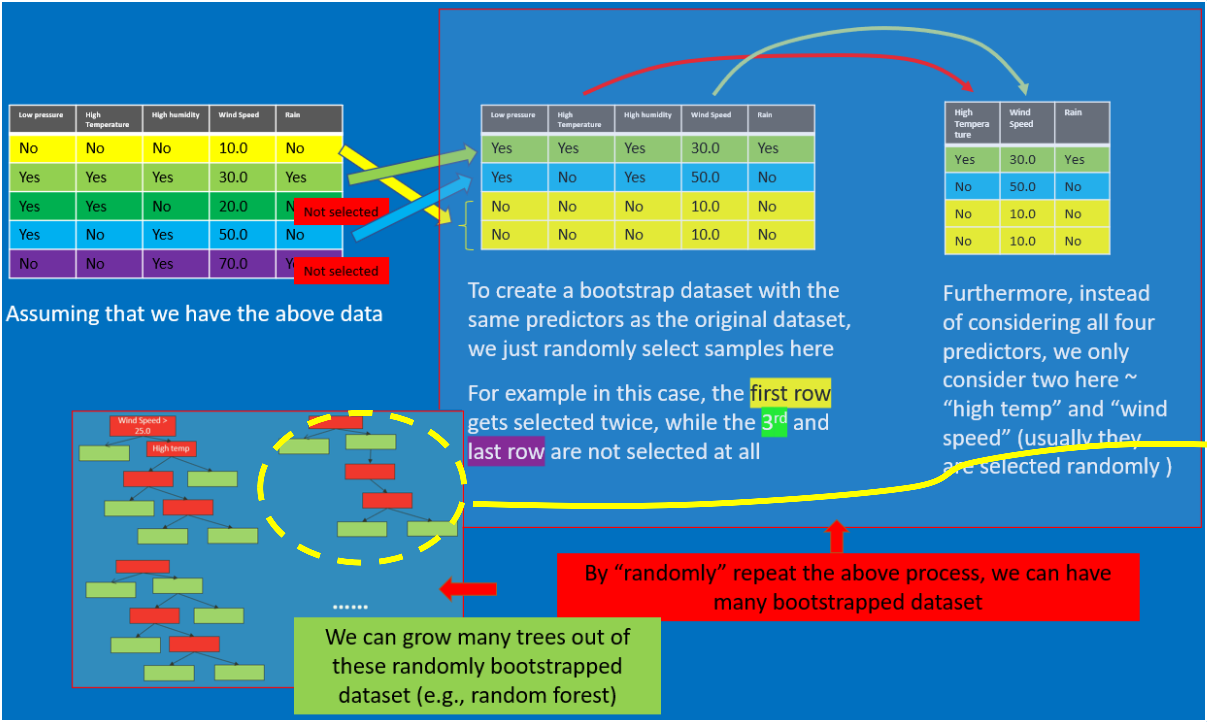
Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

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

➡ The first tree says "YES"

(It will rain)

Bagging



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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ "high temp" and "wind speed" (usually they are selected randomly)

By "randomly" repeat the above process, we can have many bootstrapped dataset

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

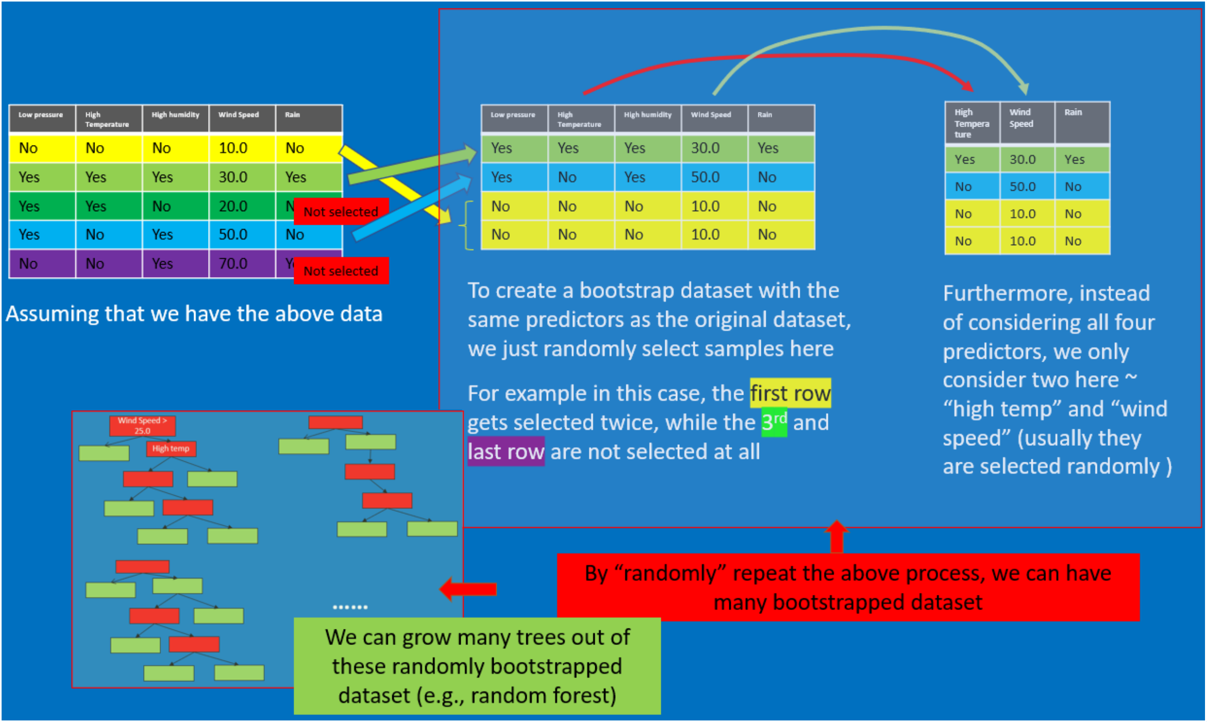
Bootstrapping

Then we have a testing data

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

- So we take the test data, run it through the first tree we made ⇨ The first tree says "YES" *(It will rain)*
- We take the test data, run it through the 2nd tree we made ⇨ The 2nd tree says "YES" *(It will rain)*

Bagging



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	N
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y

Not selected

Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

By “randomly” repeat the above process, we can have many bootstrapped dataset

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

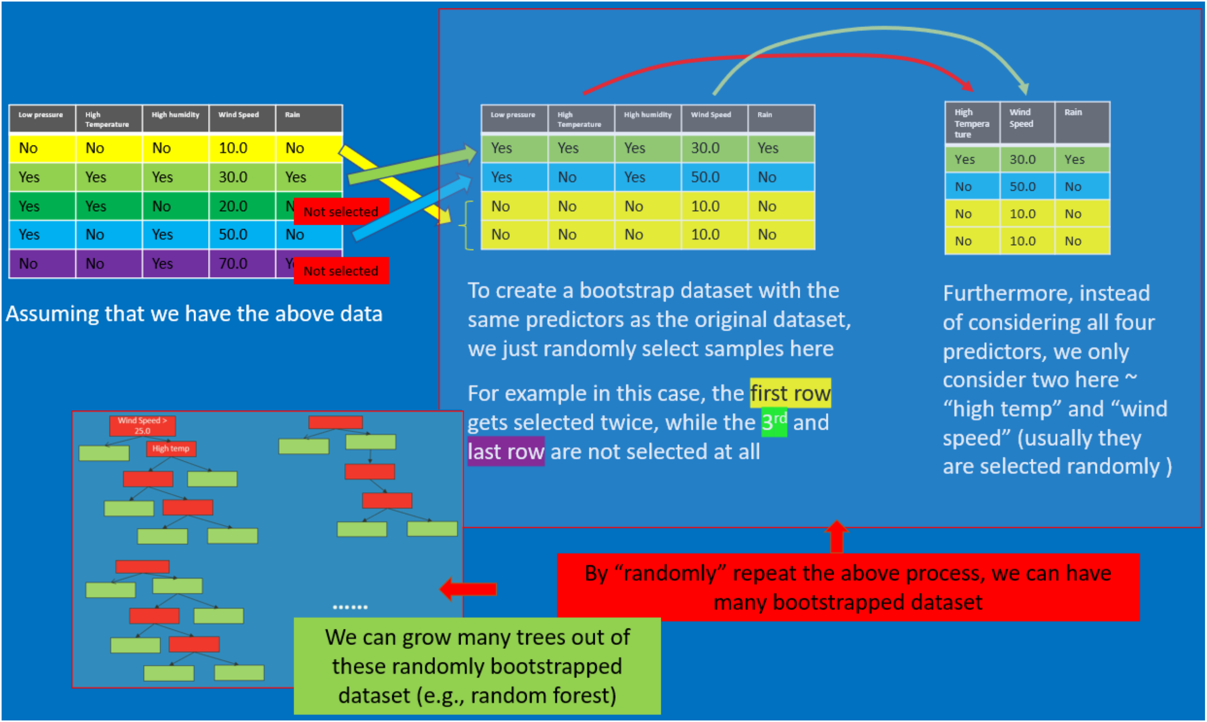
Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	Yes	No	40.0	?

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- We take the test data, run it through the 2nd tree we made ➡ The 2nd tree says “YES” *(It will rain)*

......

- We take the test data, run it through the nth tree we made ➡ The nth tree says “NO” *(It won’t rain)*

Bagging



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	N
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y

Not selected

Not selected

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ "high temp" and "wind speed" (usually they are selected randomly)

By "randomly" repeat the above process, we can have many bootstrapped dataset

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

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

- So we take the test data, run it through the first tree we made ➡ The first tree says "YES" *(It will rain)*
- We take the test data, run it through the 2nd tree we made ➡ The 2nd tree says "YES" *(It will rain)*

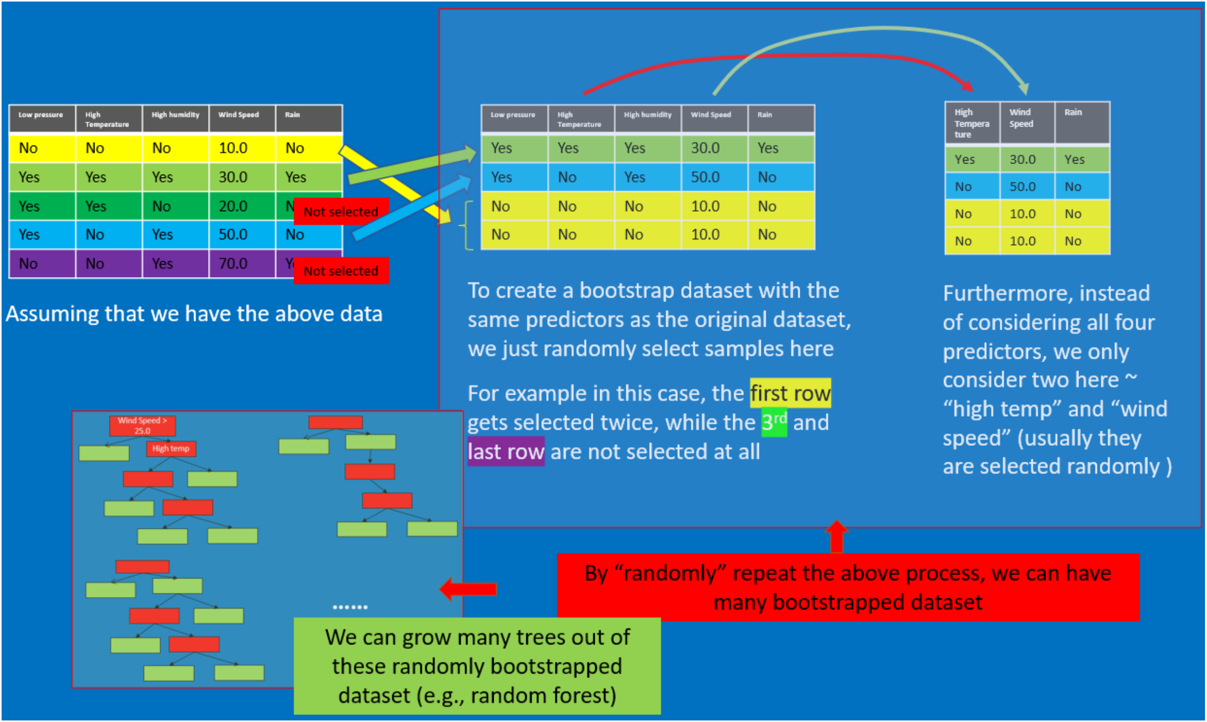
……

- We take the test data, run it through the nth tree we made ➡ The nth tree says "NO" *(It won't rain)*

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

Rain: YES	Rain: NO
15	3

Bagging



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	N
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y

Assuming that we have the above data

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

By “randomly” repeat the above process, we can have many bootstrapped dataset

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

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

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- We take the test data, run it through the 2nd tree we made ➡ The 2nd tree says “YES” *(It will rain)*

……

- We take the test data, run it through the nth tree we made ➡ The nth tree says “NO” *(It won’t rain)*

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

Rain: YES	Rain: NO
15	3

➡ Final output: YES

Bagging

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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

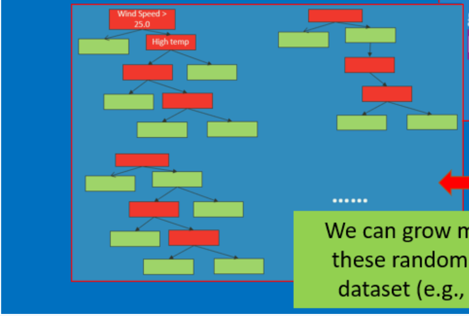
To create a bootstrap dataset with the same predictors as the original dataset, we just randomly select samples here

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

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

Furthermore, instead of considering all four predictors, we only consider two here ~ “high temp” and “wind speed” (usually they are selected randomly)

By “randomly” repeat the above process, we can have many bootstrapped dataset



We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

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

- So we take the test data, run it through the first tree we made ➡ The first tree says “YES” *(It will rain)*
- We take the test data, run it through the 2nd tree we made ➡ The 2nd tree says “YES” *(It will rain)*

......

- We take the test data, run it through the nth tree we made ➡ The nth tree says “NO” *(It won’t rain)*

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

Rain: YES	Rain: NO
15	3

➡ Final output: YES

Aggregation

Bagging

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	N Not selected
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Y Not selected

Assuming that we have the above data

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

High Tempera ture	Wind Speed	Rain
Yes	30.0	Yes
No	50.0	No
No	10.0	No
No	10.0	No

Furthermore, instead of considering all four predictors, we only consider two here ~ "high temp" and "wind speed" (usually they are selected randomly)

By "randomly" repeat the above process, we can have many bootstrapped dataset

Wind Speed > 25.0
High temp
......

We can grow many trees out of these randomly bootstrapped dataset (e.g., random forest)

Bootstrapping

Then we have a testing data

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

- So we take the test data, run it through the first tree we made ➡ The first tree says "YES" *(It will rain)*
- We take the test data, run it through the 2nd tree we made ➡ The 2nd tree says "YES" *(It will rain)*

......

- We take the test data, run it through the nth tree we made ➡ The nth tree says "NO" *(It won't rain)*

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

Rain: YES	Rain: NO
15	3

➡ Final output: YES

Aggregation

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