Decision Tree

Classification

Let’s look at an example

predictors

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

We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

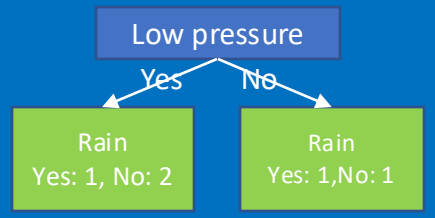
Let’s look at an example

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

We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

First let's look at “low pressure”



- When we have Low pressure, there are 1 case having rain, and 2 cases do not have rain
- When we don’t have low pressure, there are 1 case having rain, and 1 case don’t have rain

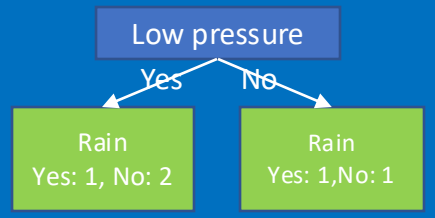
Let’s look at an example

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

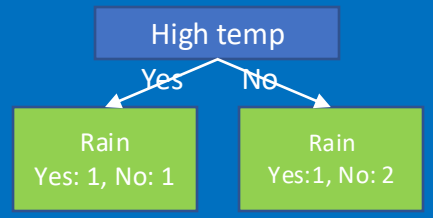
We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

First let's look at “low pressure”



Second let's look at “High temperature”



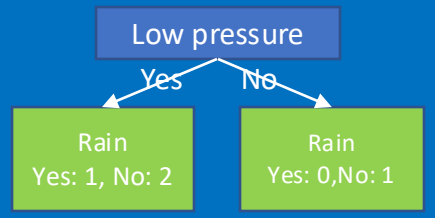
Let’s look at an example

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

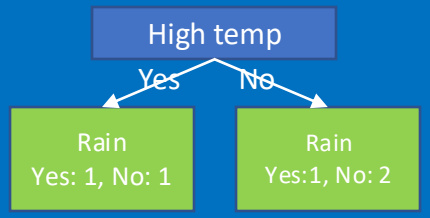
We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

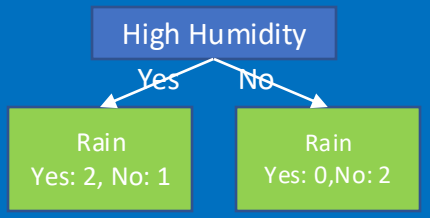
First let's look at “low pressure”



Second let's look at “High temp”



Third let's look at “High humidity”



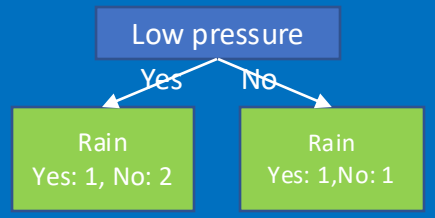
Let’s look at an example

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

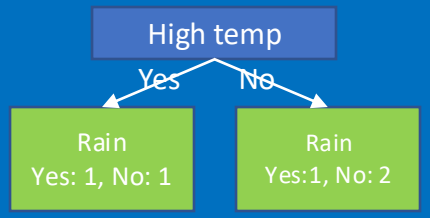
We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

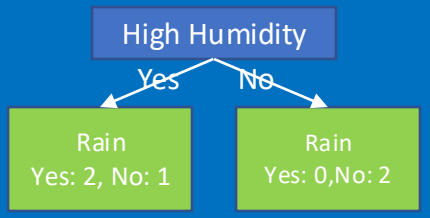
First let's look at “low pressure”



Second let's look at “High temp”



Third let's look at “High humidity”



In order to determine which predictor has the highest correlation, we use the metric called “Gini”

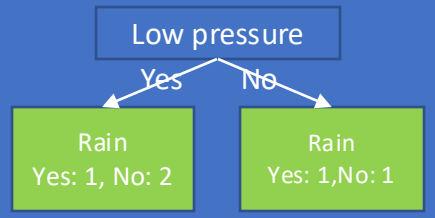
Let's look at an example

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

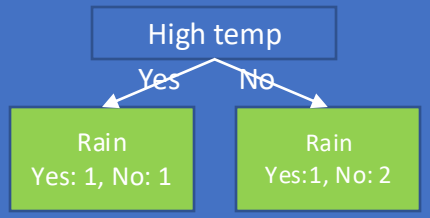
We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

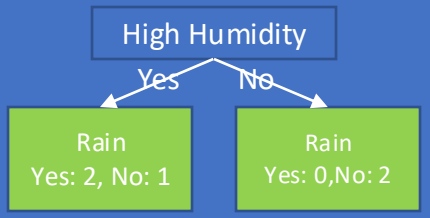
First let's look at "low pressure"



Second let's look at "High temp"



Third let's look at "High humidity"



In order to determine which predictor has the highest correlation, we use the metric called "Gini"

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

Where

P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

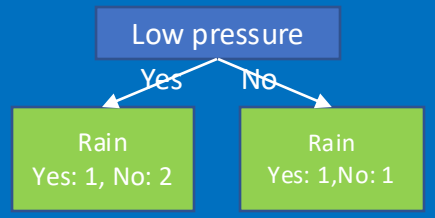
Let's look at an example

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

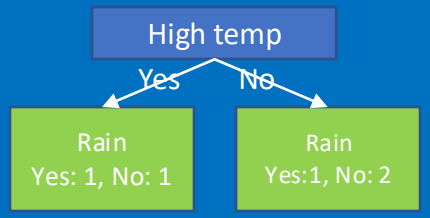
We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

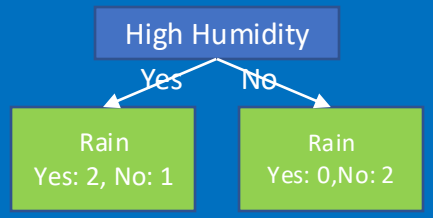
First let's look at "low pressure"



Second let's look at "High temp"



Third let's look at "High humidity"



In order to determine which predictor has the highest correlation, we use the metric called "Gini"

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

Or whatever the label is ? For example,

$$G = 1 - {P_{asian}}^2 - {P_{european}}^2 - {P_{maori}}^2$$

Where

P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

Let's look at an example

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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

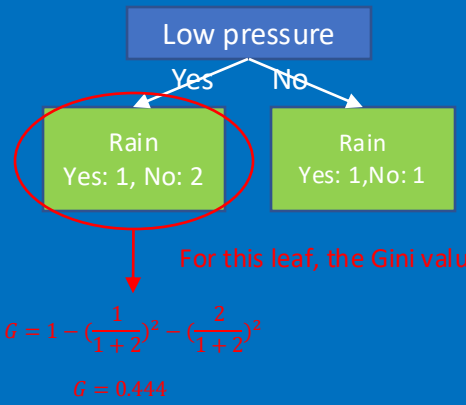
P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

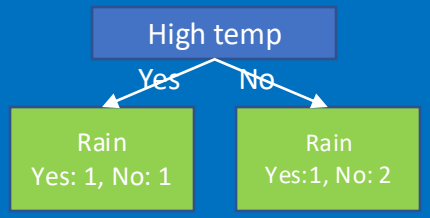
We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

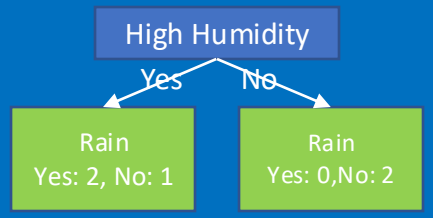
First let's look at "low pressure"



Second let's look at "High temp"



Third let's look at "High humidity"



Let's look at an example

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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

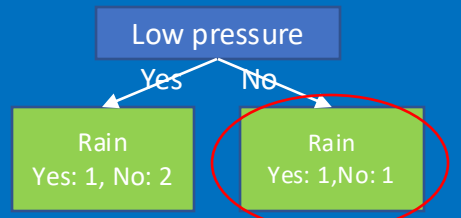
P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

First let's look at "low pressure"



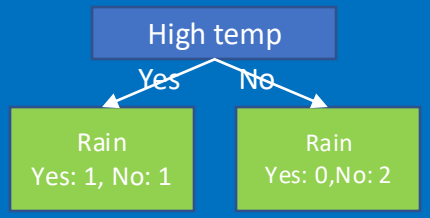
$G = 0.444$

For this leaf, the Gini value is

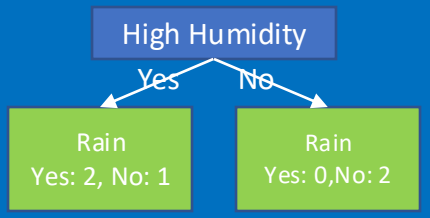
$$G = 1 - (\frac{1}{1+1})^2 - (\frac{1}{1+1})^2$$

$$G = 0.5$$

Second let's look at "High temp"



Third let's look at "High humidity"



Let's look at an example

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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

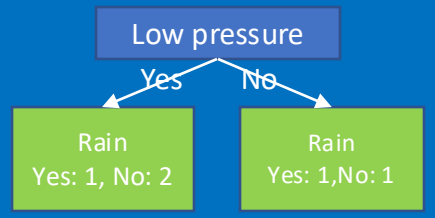
P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

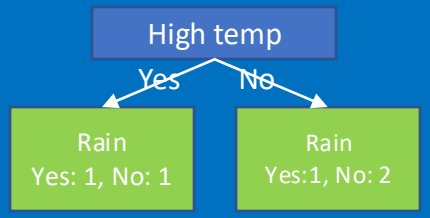
In order to do so, we check the correlation for each predictor individually

First let's look at "low pressure"

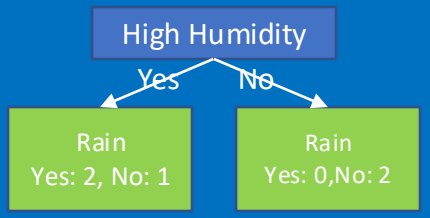


$G = 0.444$ $G = 0.5$

Second let's look at "High temp"



Third let's look at "High humidity"



So, the average Gini for "Low pressure" can be calculated by

$$G = 0.444 \times \left(\frac{3}{5}\right) + 0.5 \times \left(\frac{2}{5}\right)$$

$$G = 0.4664$$

Thus, the average Gini for "Low pressure" is 0.4664

Let's look at an example

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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

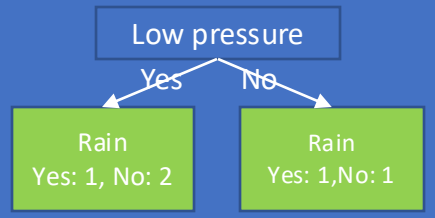
P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

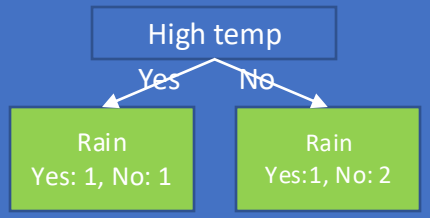
In order to do so, we check the correlation for each predictor individually

First let's look at "low pressure"

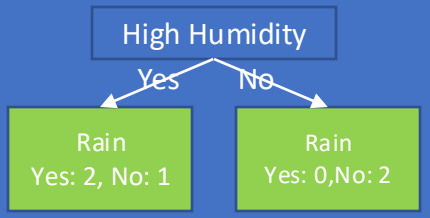


$G = 0.444$ $G = 0.5$

Second let's look at "High temp"



Third let's look at "High humidity"



So, the average Gini for "Low pressure" can be calculated by

$$G = 0.444 \times \left(\frac{3}{5}\right) + 0.5 \times \left(\frac{2}{5}\right)$$

$$G = 0.4664$$

Total 5 leafs, 3 of them from Yes (low pressure), and 2 from No

Thus, the average Gini for "Low pressure" is 0.4664

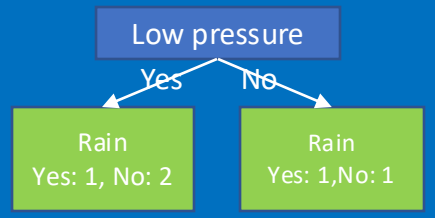
Let's look at an example

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

We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

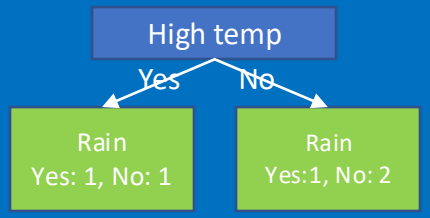
In order to do so, we check the correlation for each predictor individually

First let's look at "low pressure"

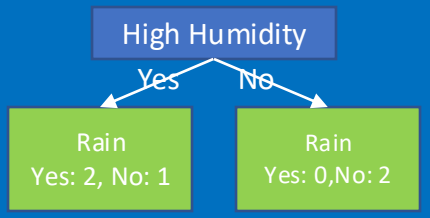


$G = 0.444$ $G = 0.5$

Second let's look at "High temp"



Third let's look at "High humidity"



So, the average Gini for "Low pressure" can be calculated by

$$G = 0.444 \times \left(\frac{3}{5}\right) + 0.5 \times \left(\frac{2}{5}\right)$$

$$G = 0.4664$$

Thus, the average Gini for "Low pressure" is 0.4664

Similarly, we can the average Gini for "High Temp" as 0.512

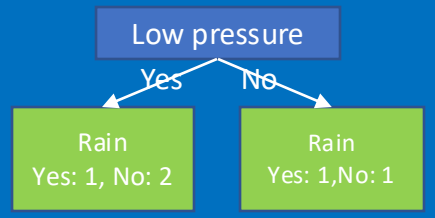
Let's look at an example

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

We need to determine whether "low pressure", "high temperature" or "high humidity" should be on the top of the tree (root)

In order to do so, we check the correlation for each predictor individually

First let's look at "low pressure"



$G = 0.444$ $G = 0.5$

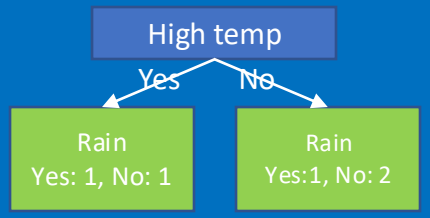
So, the average Gini for "Low pressure" can be calculated by

$$G = 0.444 \times \left(\frac{3}{5}\right) + 0.5 \times \left(\frac{2}{5}\right)$$

$$G = 0.4664$$

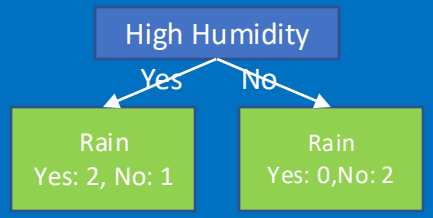
Thus, the average Gini for "Low pressure" is 0.4664

Second let's look at "High temp"



Similarly, we can the average Gini for "High Temp" as 0.512

Third let's look at "High humidity"



Similarly, we can the average Gini for "High Humidity" as 0.733

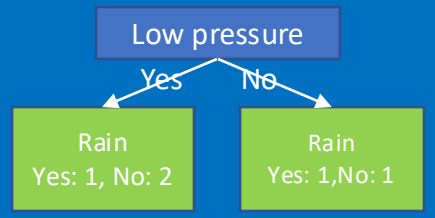
Let’s look at an example

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

We need to determine whether “low pressure”, “high temperature” or “high humidity” should be on the top of the tree (root)

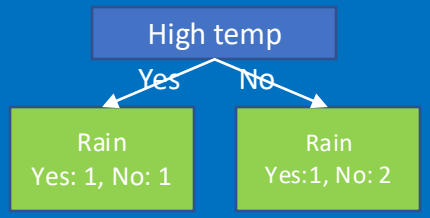
In order to do so, we check the correlation for each predictor individually

First let's look at “low pressure”



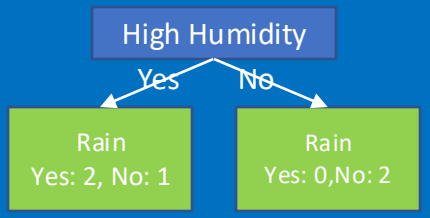
0.4664

Second let's look at “High temp”



0. 512

Third let's look at “High humidity”



0. 733

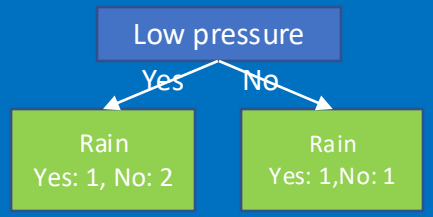
These numbers are called Gini impurity

Let’s look at an example

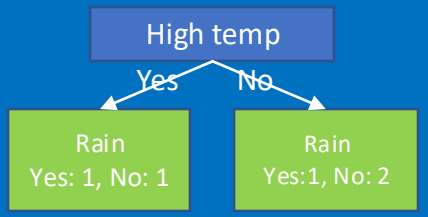
Low pressure	High Temperature	High humidity	Wind Speed	Rain
No	No	No	10.0	No
Yes	Yes	Yes	30.0	Yes
Yes	Yes	No	20.0	No
Yes	No	Yes	50.0	No
No	No	Yes	70.0	Yes

Now let’s look at a more complicated example, in predictor, we have wind speed, which is not marked as “Yes/No”, instead it has a series of values.

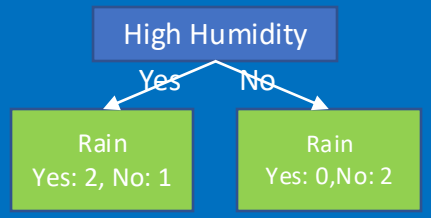
We want to see how the Gini impurity can be calculated



0.4664



0. 512



0. 733

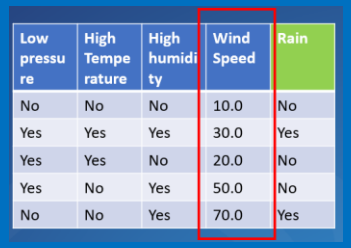
Let’s look at an example

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

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the “wind speed” from smallest to biggest

Let’s look at an example



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

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the “wind speed” from smallest to biggest

15.0

25.0

40.0

60.0

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

Second step, calculate all adjacent wind speed

Let's look at an example

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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the "wind speed" from smallest to biggest

15.0

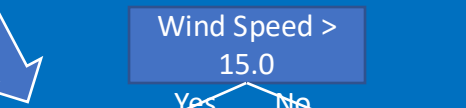
25.0

40.0

60.0

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

Second step, calculate all adjacent average wind speed



$$G = 1 - (\frac{2}{2+2})^2 - (\frac{2}{2+2})^2$$

$$G = 0.5$$

Then we create tree for each adjacent averaged wind speed, e.g., for wind speed of 15.0, we have average Gini as ?

Let's look at an example

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

$$G = 1 - P_{yes}{}^2 - P_{no}{}^2$$

P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the "wind speed" from smallest to biggest

15.0

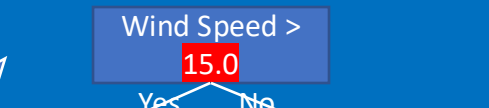
25.0

40.0

60.0

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

Second step, calculate all adjacent average wind speed



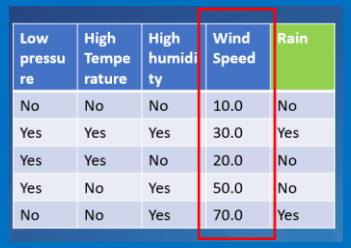
0.5

$$G = 1 - (\frac{0}{0+1})^2 - (\frac{1}{0+1})^2$$

$$G = 0.0$$

Then we create tree for each adjacent averaged wind speed, e.g., for wind speed of 15.0, we have average Gini as ?

Let's look at an example



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

$$G = 1 - {P_{yes}}^2 - {P_{no}}^2$$

P_{yes} is the probability of "yes" in a leaf

P_{no} is the probability of "no" in a leaf

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the "wind speed" from smallest to biggest

15.0

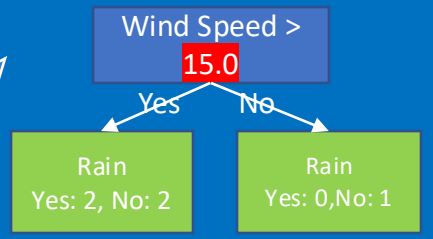
25.0

40.0

60.0

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

Second step, calculate all adjacent average wind speed



0.5 0.0

$$G = 0.5 \times \left(\frac{4}{5}\right) + 0.0 \times \left(\frac{1}{5}\right)$$

$$G = 0.4$$

Then we create tree for each adjacent averaged wind speed, e.g., for wind speed of 15.0, we have average Gini as 0.4

Let's look at an example

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

Wind Speed	Rain
10.0	No
20.0	No
30.0	Yes
50.0	No
70.0	Yes

First step, we sort the "wind speed" from smallest to biggest

	Wind Speed	Rain	
15.0	10.0	No	$G = 0.4$
25.0	20.0	No	$G = 0.3$
40.0	30.0	Yes	$G = 0.5$
60.0	50.0	No	$G = 0.45$
	70.0	Yes	

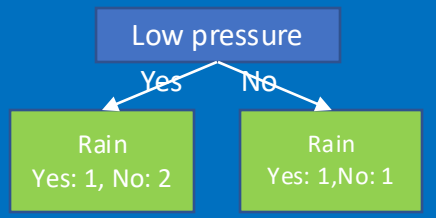
Second step, calculate all adjacent average wind speed

We can have the Gini impurity for all the adjacent average wind speed

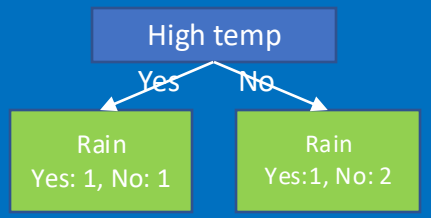
In this case, the threshold of 25.0 gives the smallest Gini, so it is picked up to represent "wind speed"

Let’s look at an example

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

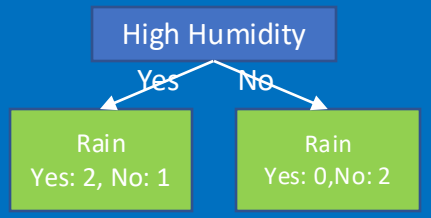


0.4664

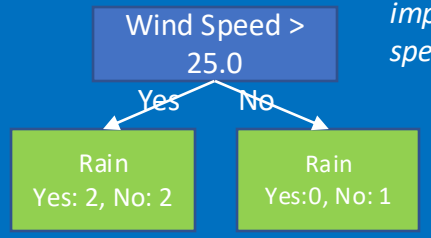


0. 512

Representative Gini impurity from “wind speed”



0. 733

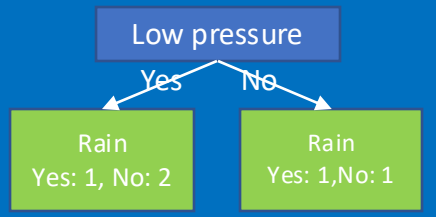


0. 3

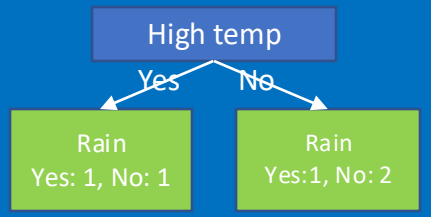
So apparently now we have Gini impurity for all predictors

Let’s look at an example

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

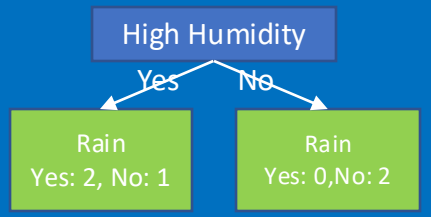


0.4664

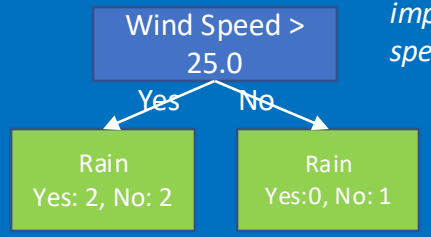


0. 512

Representative Gini impurity from “wind speed”



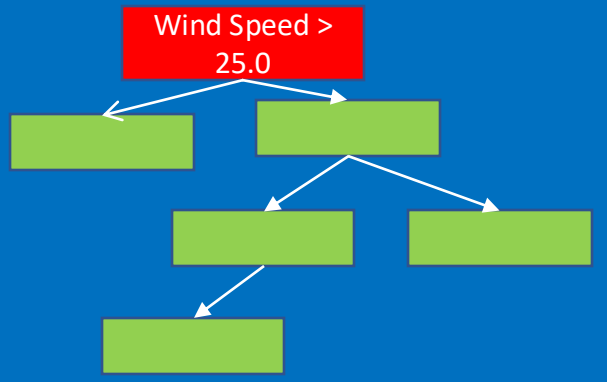
0. 733



0. 3

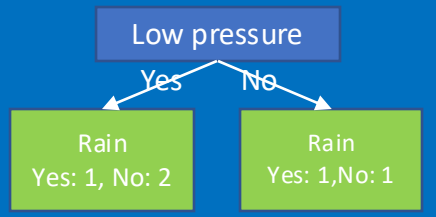
So apparently now we have Gini impurity for all predictors

Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.

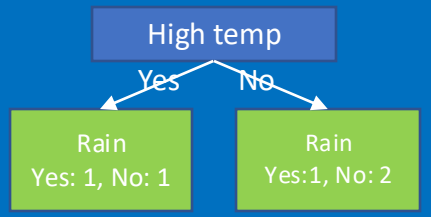


Let's look at an example

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

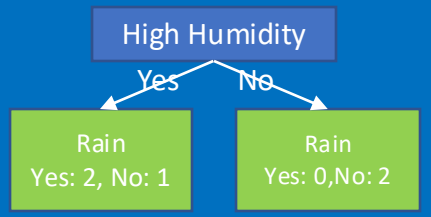


0.4664

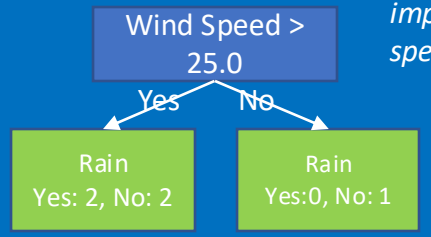


0. 512

Representative Gini impurity from "wind speed"



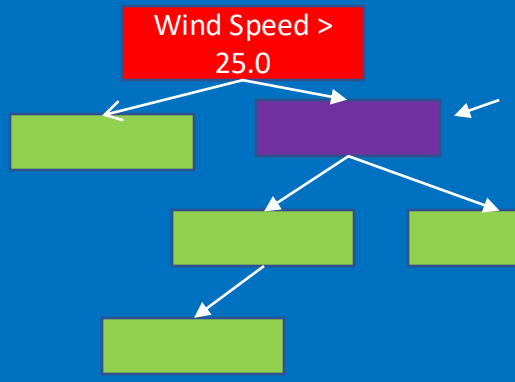
0. 733



0. 3

So apparently now we have Gini impurity for all predictors

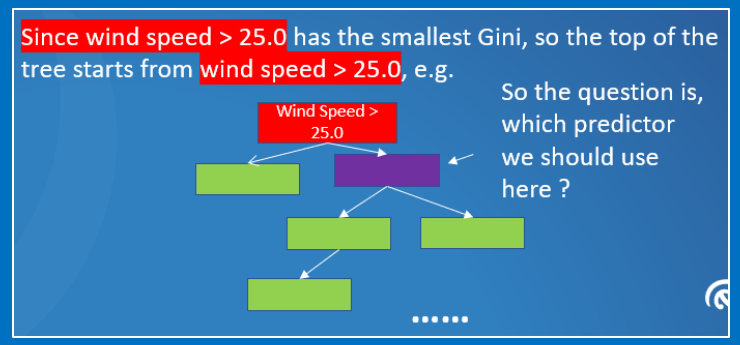
Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.



So the question is, which predictor we should use here ?

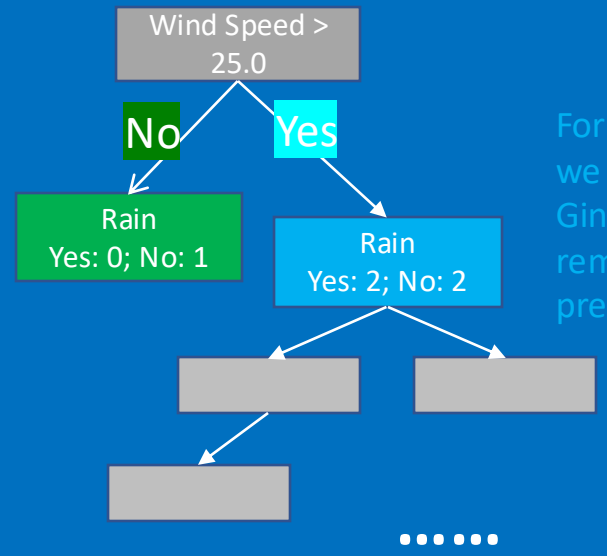
.....

Let's look at an example



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

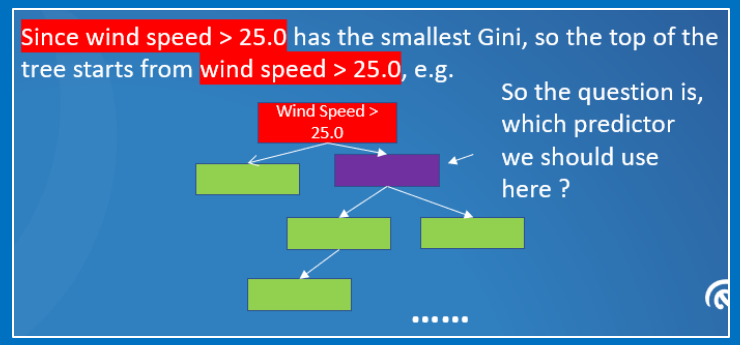
Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.



There is no need to split this branch since we've got "0/1" situation here (impure)

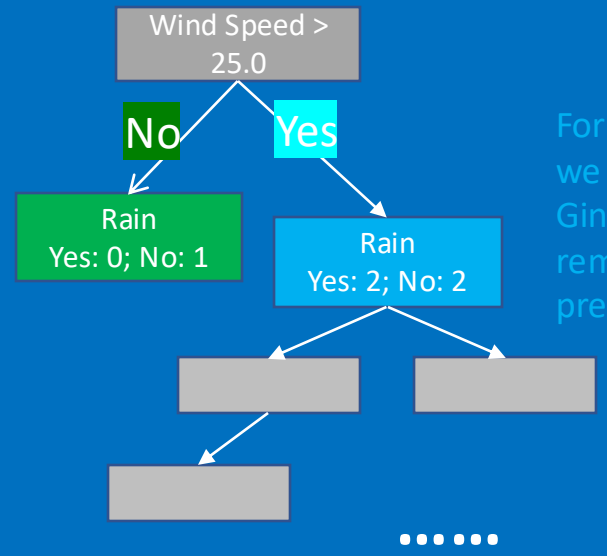
For this section, we calculate the Gini for all remaining predictors

Let's look at an example



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

Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.



There is no need to split this branch since we've got "0/1" situation here

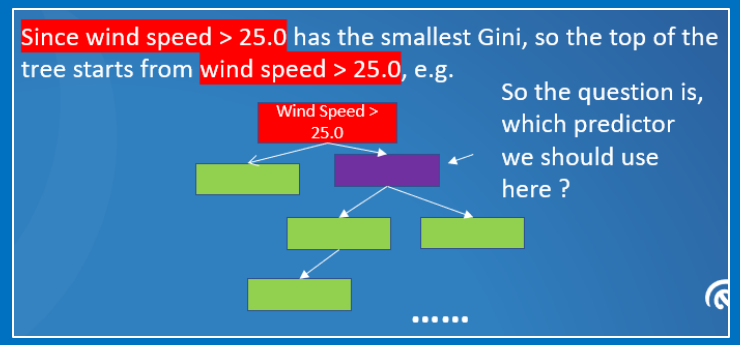
For this section, we calculate the Gini for all remaining predictors

Low pressure: 0.35

High temp: 0.2

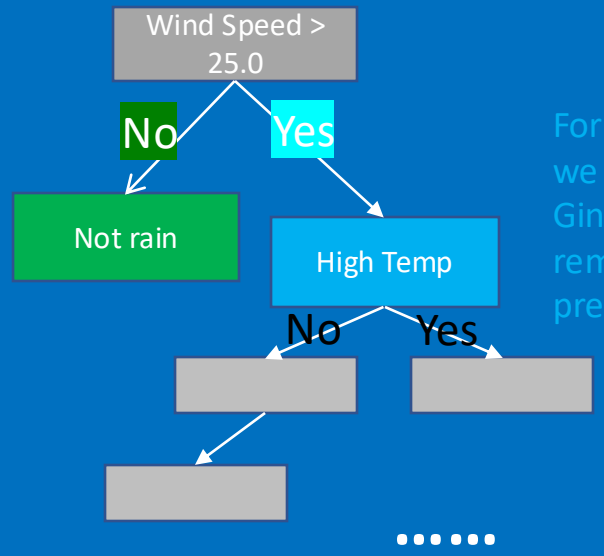
High humidity: 0.45

Let's look at an example



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

Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.



For this section, we calculate the Gini for all remaining predictors

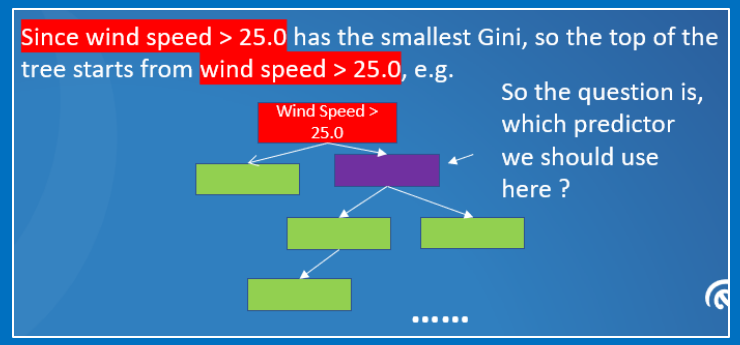
Low pressure: 0.35

High temp: 0.2

High humidity: 0.45

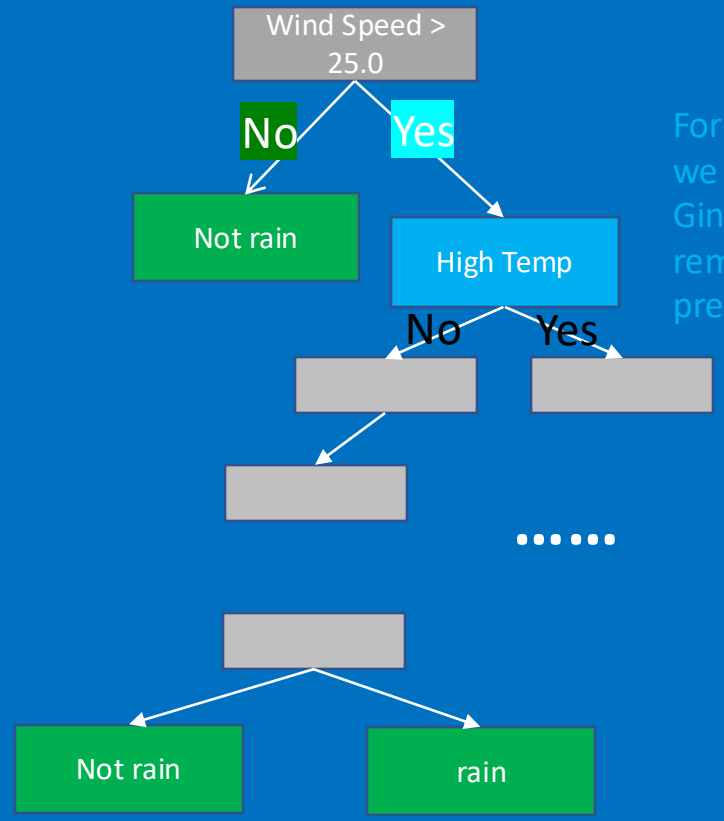
High Temp has the smallest Gini so it is chosen as the secondary split here

Let's look at an example



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

Since wind speed > 25.0 has the smallest Gini, so the top of the tree starts from wind speed > 25.0, e.g.



For this section, we calculate the Gini for all remaining predictors

Low pressure: 0.35

High temp: 0.2

High humidity: 0.45

High Temp has the smallest Gini so it is chosen as the secondary split here

By doing this over and over (going through all predictors) until we reach to the level that we are not able to split anymore (e.g., "0/1" or "impure" situation)