Creating Decision Trees using the Classification and Regression (CART) Algorithm

General Introduction

When we create decision trees, we need to select the root node and the decision nodes in an efficient way.

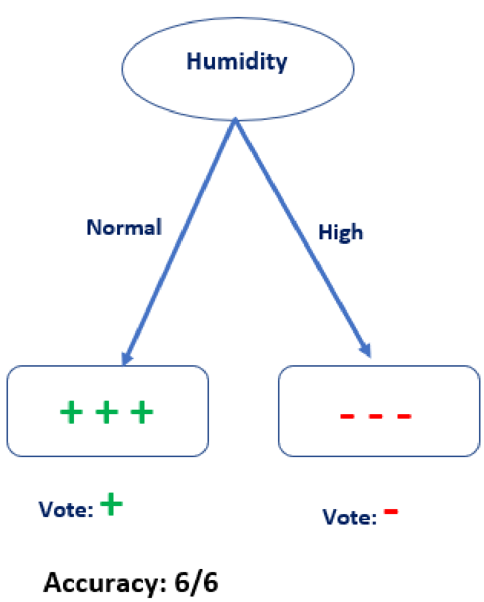
How can we determine the best split?

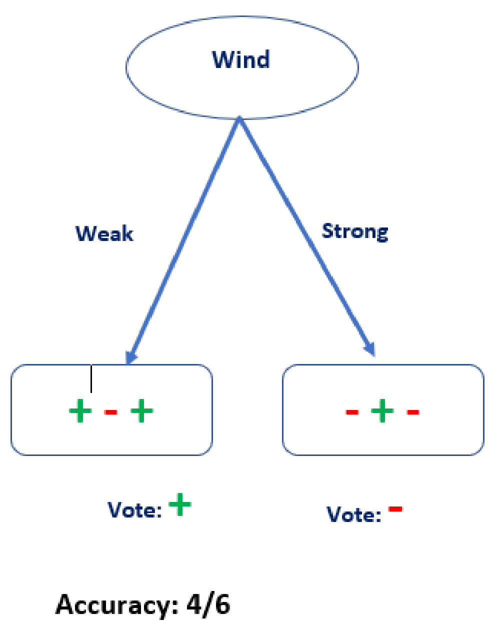
Suppose we have the following dataset related to playing tennis or not based on weather conditions.

We can use this information to build a decision tree to predict the decision based on the independent features.

How can we make the split? Which factor should we choose as a root node?

Humidity	Wind	Decision
Normal	Weak	Yes
High	Weak	No
Normal	Strong	Yes
High	Strong	No
High	Strong	No
Normal	Weak	Yes





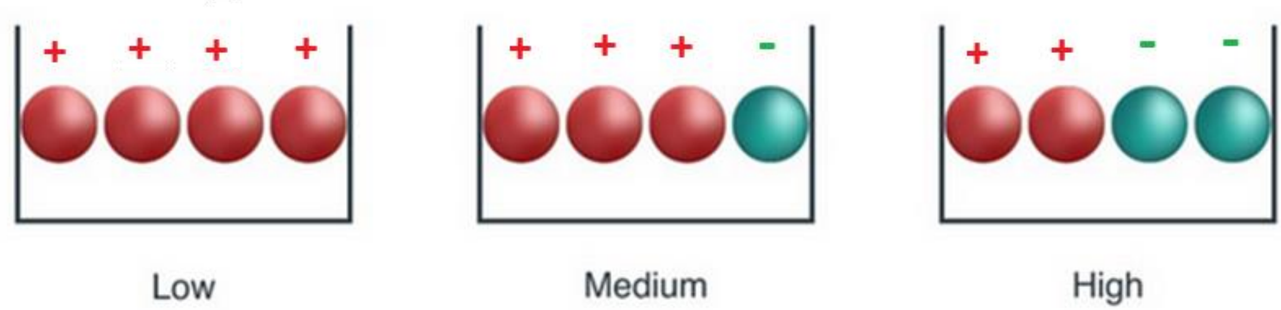
The Gini Index Method

The Gini index is a measure of inequality in sample. It has a value between 0 and 1.

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

The Gini Index can be used to evaluate the split impurity when constructing classification trees.

Gini index of value 0 means sample is perfectly homogeneous, and all elements are similar, whereas Gini index of value 1 means maximal inequality among elements.



Example: Let’s start with a weather data set, which is quite famous in explaining decision tree algorithm, where target is to predict play or not (Yes or No) based on weather condition.

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	wtrong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Outlook

Outlook is a nominal feature. It can take three values, sunny, overcast and rain.

Outlook	Yes	No	# Instances
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

- Gini index (outlook=sunny) = 1-(2/5) ^{2}-(3/5) ^{2} = 1- 0.16–0.36 = 0.48
- Gini index(outlook=overcast) = 1- (4/4) ^{2}-(0/4) ^{2} = 1- 1- 0 = 0
- Gini index(outlook=rainfall) = 1- (3/5) ^{2} -(2/5) ^{2} = 1- 0.36- 0.16 = 0.48

Now, we will calculate the weighted sum of Gini index for the outlook features,

Gini(outlook) = (5/14) *0.48 + (4/14) *0 + (5/14) *0.48 = 0.342

Temperature

Similarly, temperature is also a nominal feature, it can take three values, hot, cold and mild.

Temperature	Yes	No	# Instances
hot	2	2	4
cool	3	1	4
mild	4	2	6

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

- Gini(temperature=hot) = 1-(2/4) ²-(2/4) ² = 0.5
- Gini(temperature=cool) = 1-(3/4) ²-(1/4) ² = 0.375
- Gini(temperature=mild) = 1-(4/6) ²-(2/6) ² = 0.445

Now, the weighted sum of Gini index for temperature features can be calculated as,

Gini(temperature)= (4/14) *0.5 + (4/14) *0.375 + (6/14) *0.445 =0.439

Humidity

Humidity	Yes	No	# Instances
high	3	4	7
Normal	6	1	7

Humidity is a binary class feature; it can take two values high and normal.

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

- Gini(humidity=high) = 1-(3/7) ^{2}-(4/7) ^{2} = 0.489
- Gini(humidity=normal) = 1-(6/7) ^{2}-(1/7) ^{2} = 0.244

Now, the weighted sum of Gini index for humidity features can be calculated as,

Gini(humidity) = (7/14) *0.489 + (7/14) *0.244=0.367

Wind

wind	Yes	No	# Instances
weak	6	2	8
strong	3	3	6

wind is a binary class feature; it can take two values weak and strong.

$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

- Gini(wind=weak) = 1-(6/8) ^{2}-(2/8) ^{2} = 0.375
- Gini(wind=strong) = 1-(3/6) ^{2}-(3/6) ^{2}= 0.5

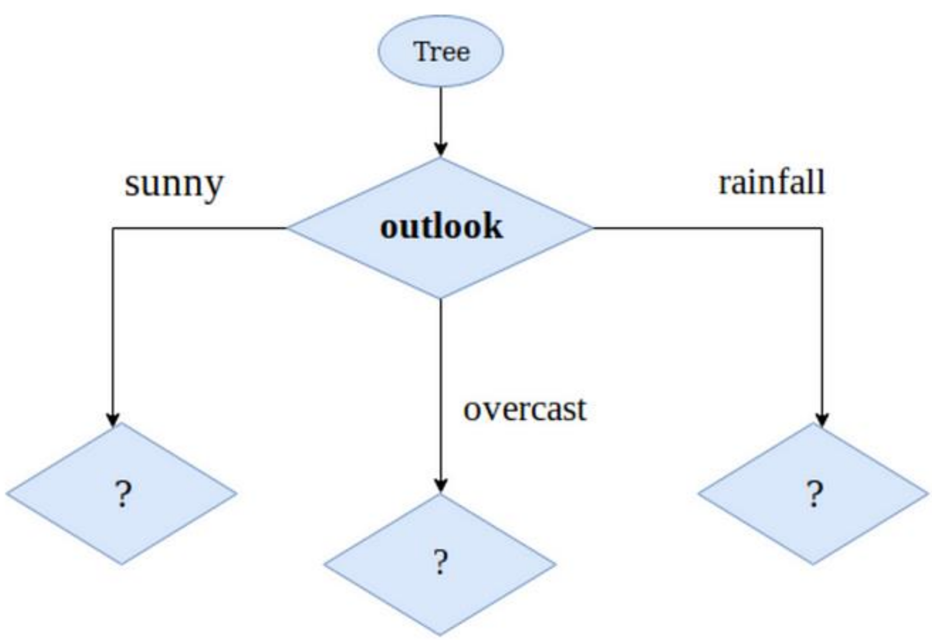
Now, the weighted sum of Gini index for wind features can be calculated as,

Gini(wind) = (8/14) *0.375 + (6/14) *0.5=0.428

Decision for root node

Features	Gini Index
outlook	0.342
temperature	0.439
humidity	0.367
wind	0.428

From the table, we can see that Gini index for outlook feature is lowest. So, we get our root node.



Let’s now focus on sub data on sunny outlook features.

We need to find the Gini index for temperature, humidity and wind feature respectively.

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

Gini index for temperature on sunny outlook

Temperature	Yes	No	# Instances
hot	0	2	2
cool	1	1	1
mild	1	1	2

- Gini (outlook=sunny & temperature=hot) = $1-(0/2)^2-(2/2)^2 = 0$
- Gini (outlook=sunny & temperature=cool) = $1-(1/1)^2-(0/1)^2 = 0$
- Gini (outlook=sunny & temperature=mild) = $1-(1/2)^2-(1/2)^2 = 0.5$

Now, the weighted sum of Gini index for temperature on sunny outlook features can be calculated as,

Gini (outlook=sunny & temperature) = (2/5) *0 + (1/5) *0+ (2/5) *0.5 =0.2

Gini Index for humidity on sunny outlook

Humidity	Yes	No	# Instances
high	0	3	3
Normal	2	0	2

- Gini (outlook=sunny & humidity=high) = $1-(0/3)^2-(3/3)^2 = 0$
- Gini (outlook=sunny & humidity=normal) = $1-(2/2)^2-(0/2)^2 = 0$

Now, the weighted sum of Gini index for humidity on sunny outlook features can be calculated as,

Gini (outlook = sunny & humidity) = $(3/5) *0 + (2/5) *0=0$

Gini Index for wind on sunny outlook

wind	Yes	No	# Instances
weak	1	2	3
strong	1	1	2

- Gini (outlook=sunny & wind=weak) = $1-(1/3)^2-(2/3)^2 = 0.44$
- Gini (outlook=sunny & wind=strong) = $1-(1/2)^2-(1/2)^2 = 0.5$

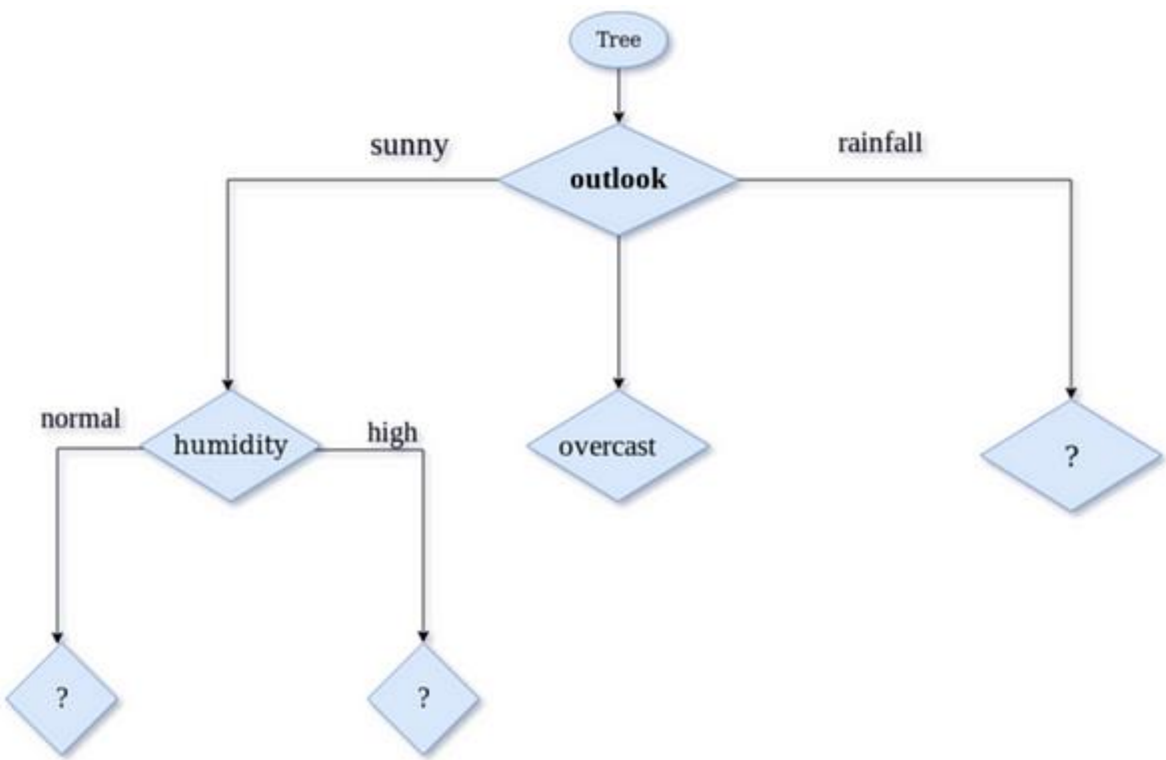
Now, the weighted sum of Gini index for wind on sunny outlook features can be calculated as,

Gini (outlook = sunny and wind) = (3/5) *0.44 + (2/5) *0.5=0.266+0.2= 0.466

Decision on sunny outlook factor

Features	Gini Index
temperature	0.2
humidity	0
wind	0.466

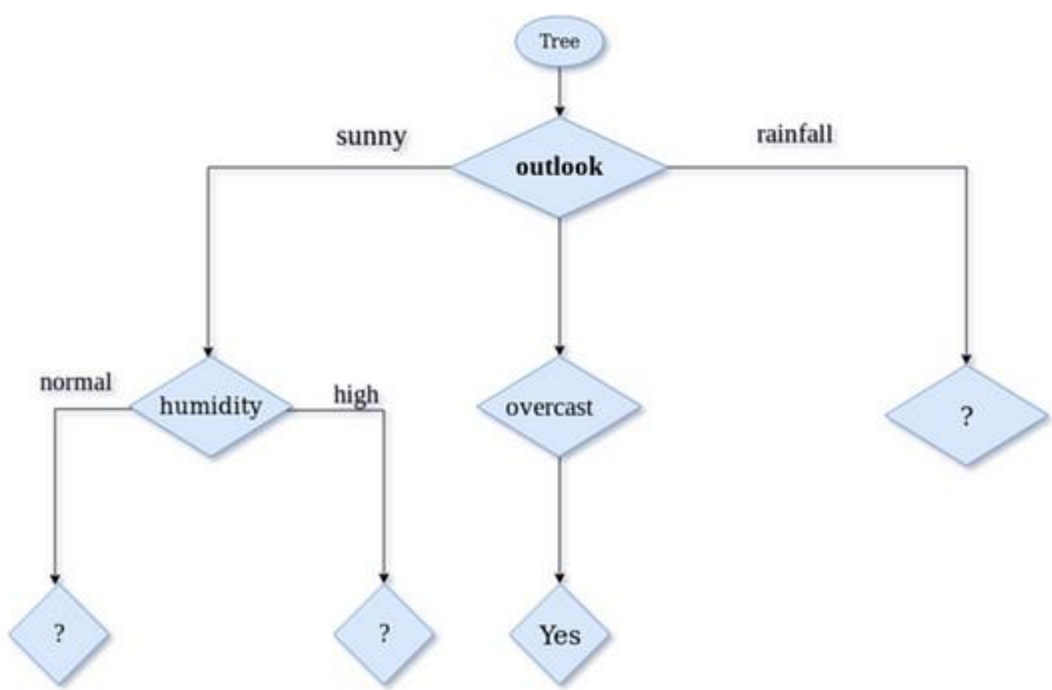
Humidity has the lowest value. So the next node will be humidity.



Now, Let’s focus on sub data for overcast outlook feature.

Day	outlook	temperature	humidity	wind	decision
3	overcast	hot	high	weak	Yes
7	overcast	cool	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes

Looking at the table above, we can see that all the decision for overcast outlook feature is always ‘Yes’. Then Gini index for each feature is 0, which means it is a leaf node.

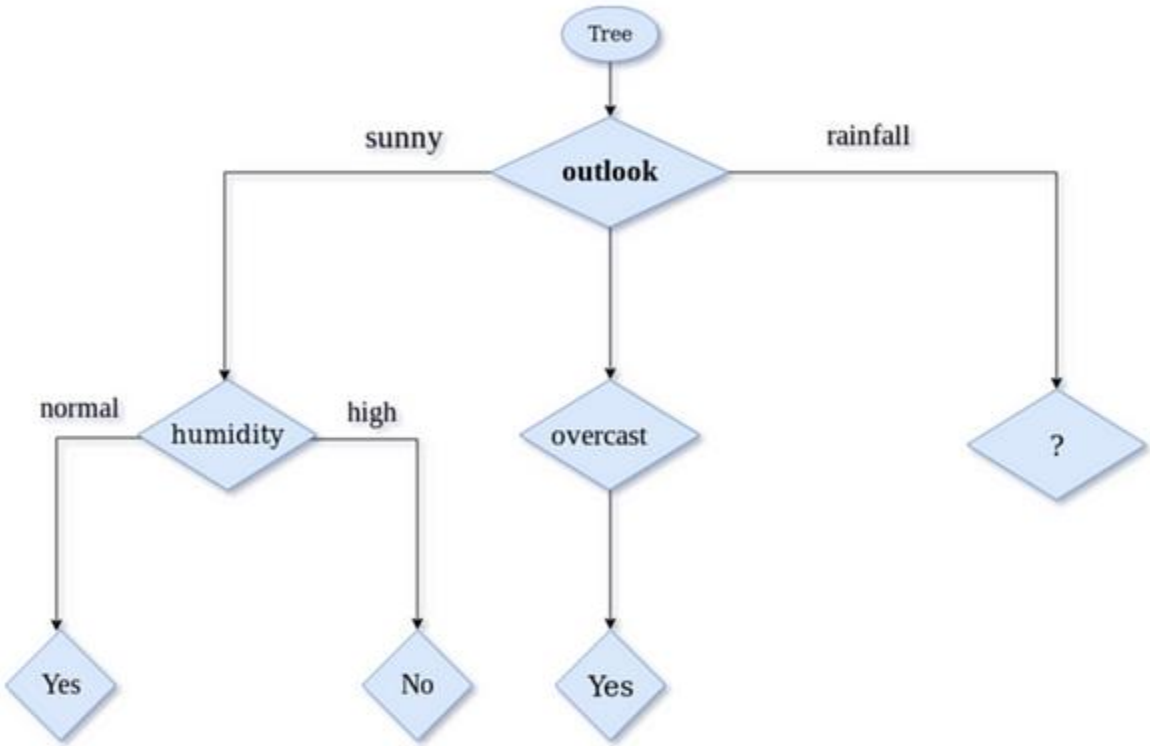


Now, Let’s focus on sub data for high and normal humidity feature.

Day	outlook	temperature	humidity	wind	decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No

Day	outlook	temperature	humidity	wind	decision
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

From the given two table, the decision is always ‘No’ when humidity is ‘high’ and decision is always ‘Yes’ when humidity is ‘normal’. So we got leaf node. now decision tree can be viewed as,


Tree
sunny
outlook
rainfall
normal
humidity
high
overcast
?
Yes
No
Yes

Now, let's focus on sub data for rainfall outlook features.

We need to find the Gini index for temperature, humidity and wind feature respectively.

Day	outlook	temperature	humidity	wind	Decision
4	rain	mild	high	weak	Yes
5	rain	cool	normal	weak	Yes
6	rain	cool	normal	strong	No
10	rain	mild	normal	weak	Yes
14	rain	mild	high	strong	No

Gini Index for temperature for rainfall outlook

temperature	Yes	No	# Instances
cool	1	1	2
mild	2	1	3

Gini (outlook=rainfall and temp.=Cool) = 1 — (1/2)2 — (1/2)2 = 0.5

Gini (outlook=rainfall and temp.=Mild) = 1 — (2/3)2 — (1/3)2 = 0.444

Gini (outlook=rainfall and temp.) = (2/5) *0.5 + (3/5) *0.444 = 0.466

Gini Index for humidity for rainfall outlook

humidity	Yes	No	# Instances
high	1	1	2
normal	2	1	3

- Gini (outlook=rainfall and humidity=high) = 1 — (1/2)2 — (1/2)2 = 0.5
- Gini (outlook=rainfall and humidity=normal) = 1 — (2/3)2 — (1/3)2 = 0.444
- Gini (Outlook=rainfall and humidity) = (2/5) *(0.5 + (3/5) *0.444 = 0.466

Gini Index for wind for rainfall outlook feature

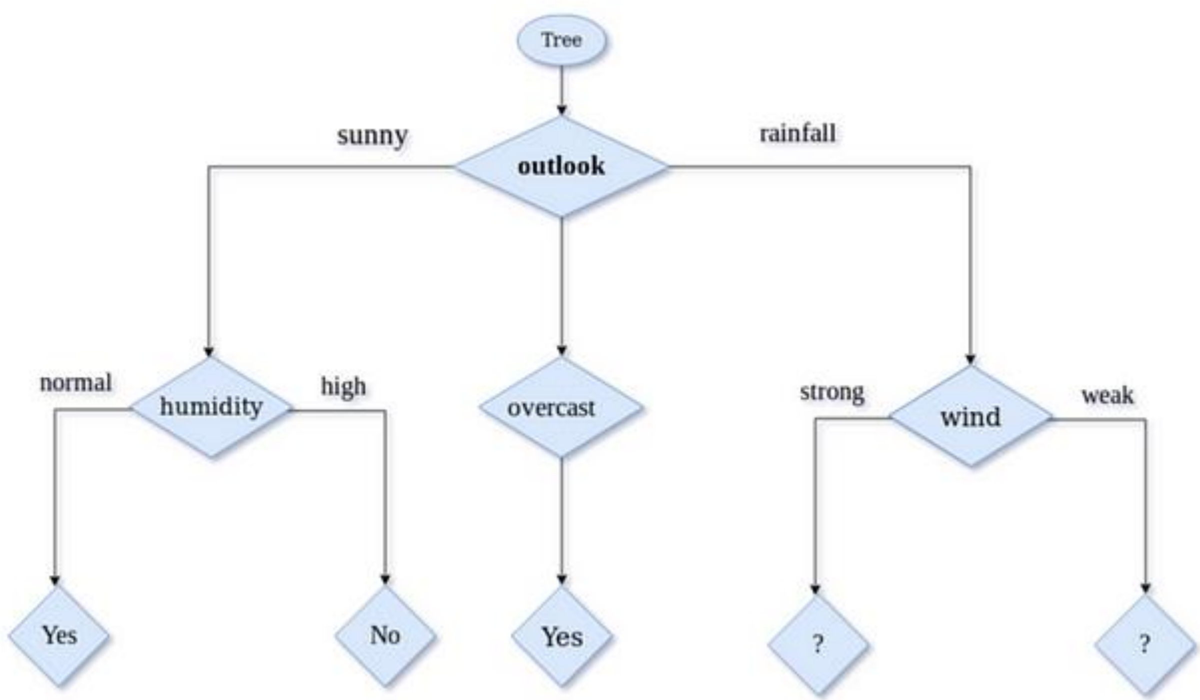
wind	Yes	No	# Instances
weak	3	0	3
strong	0	2	2

- Gini (outlook=rainfall and wind=weak) = 1 — (3/3)2 — (0/3)2 = 0
- Gini (outlook=rainfall and wind=strong) = 1 — (0/2)2 — (2/2)2 = 0
- Gini (outlook=rainfall and wind) = (3/5) *0 + (2/5) *0 = 0

Decision on rainfall outlook factor

Features	Gini Index
temperature	0.466
humidity	0.466
wind	0

We have calculated the Gini index of all the features when the outlook is rainfall. You can infer that wind has lowest value. so next node will be wind.

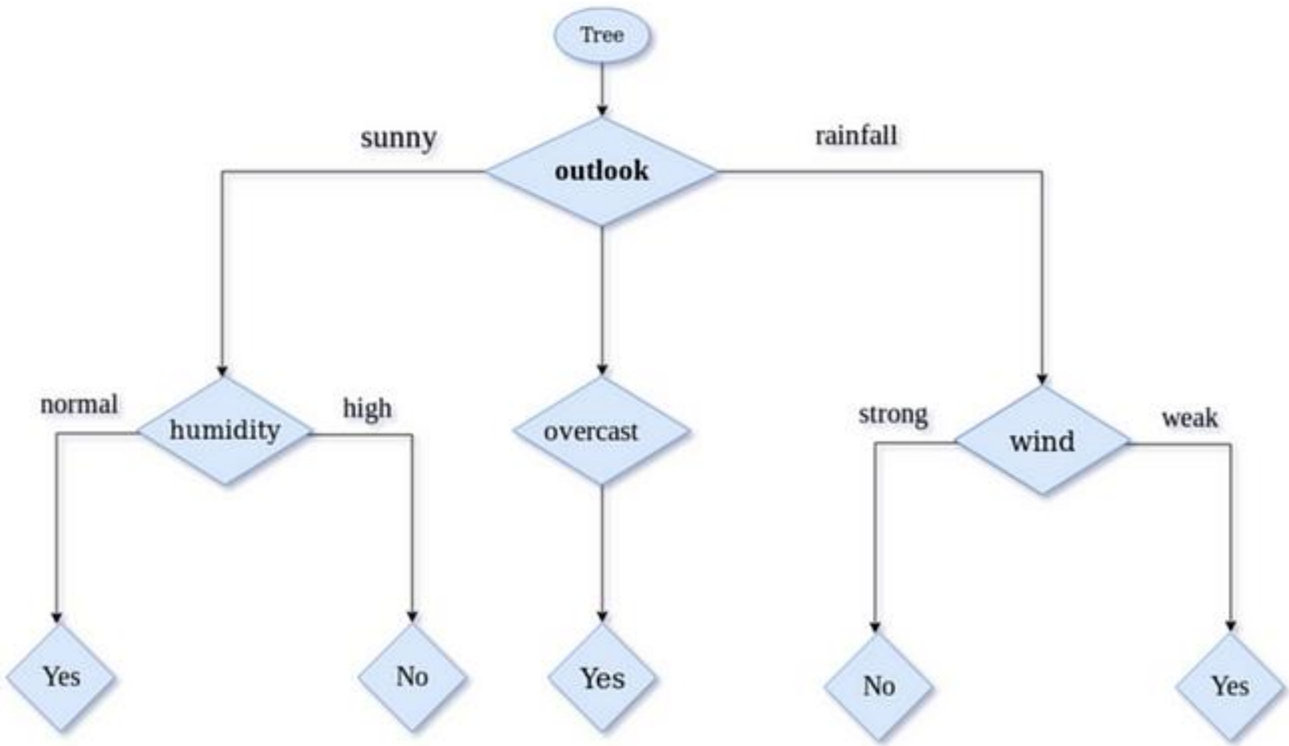


Now, let's focus on sub data strong and weak for wind rainfall feature.

Day	outlook	temperature	humidity	wind	decision
6	rainfall	cool	normal	strong	No
14	rainfall	mild	high	strong	No

Day	outlook	temperature	humidity	wind	decision
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes

From the above two table, the decision is always 'No' when wind is 'strong' and decision is always 'Yes' when wind is 'weak'. So we got leaf node.



References:

https://medium.com/@singhakshay.etw69/decision-tree-algorithm-cart-e61032794927