Decision Tree-CART Algorithm

Vipin Venugopal
Amrita School of AI

Algorithm

There are many algorithms there to build a decision tree.

They are

CART (Classification and Regression Trees) — This makes use of Gini impurity as the metric.

ID3 (Iterative Dichotomiser 3) — This uses entropy and information gain as metric.

Gini Index

- ❑ Many alternative measures to Information Gain
- ❑ Most popular altermative: Gini index

  # used in e.g., in CART (Classification And Regression Trees)

  # impurity y)

$$Gini(S) = 1 - \sum_i p_i^2$$

average Gini index (instead of average entropy / information)

$$Gini(S, A) = \sum_i \frac{|S_i|}{|S|} \cdot Gini(S_i)$$

- ❑ Gini Gain

● could be defined analogously to information gain

● but typically avg. Gini index is minimized instead of maximizing Gini gain

A Step by Step CART Decision Tree Example

Make a Decision tree that predicts whether tennis will be played on the day?

Data set

For instance, the following table informs about decision making factors to play tennis at outside for previous 14 days.

CART Algorithm for Classification

Here is the approach for most decision tree algorithms at their most simplest. The tree will be constructed in a top-down approach as follows:

Step 1: Start at the root node with all training instances
Step 2: Select an attribute on the basis of splitting criteria (Gain Ratio or other impurity metrics, discussed below)
Step 3: Partition instances according to selected attribute recursively

Partitioning stops when:
- There are no examples left
- All examples for a given node belong to the same class
- There are no remaining attributes for further partitioning – majority class is the leaf

DATA SET

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Outlook

Outlook	Yes	No	Number of instances
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

$$Gini(Outlook = Sunny) = 1 - (2/5)^2 - (3/5)^2$$
$$= 1 - 0.16 - 0.36 = 0.48$$

$$Gini(Outlook = Overcast) = 1 - (4/4)^2 - (0/4)^2 = 0$$

$$Gini(Outlook = Rain) = 1 - (3/5)^2 - (2/5)^2$$
$$= 1 - 0.36 - 0.16 = 0.48$$

Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x0.48

= 0.171 + 0 + 0.171 = 0.342

Temperature

Temperature	Yes	No	Number of instances
Hot	2	2	4
Cool	3	1	4
Mild	4	2	6

$$Gini(Temp = Hot) = 1 - (2/4)^2 - (2/4)^2 = 0.5$$

$$Gini(Temp = Cool) = 1 - (3/4)^2 - (1/4)^2$$
$$= 1 - 0.5625 - 0.0625 = 0.375$$

$$Gini(Temp = Mild) = 1 - (4/6)^2 - (2/6)^2$$
$$= 1 - 0.444 - 0.111 = 0.445$$

$$Gini(Temp)$$
$$= (4/14)\,x\,0.5 + (4/14)\,x\,0.375$$
$$+ (6/14)\,x\,0.445 = 0.142 + 0.107 + 0.190$$
$$= 0.439$$

Humidity

Humidity	Yes	No	Number of instances
High	3	4	7
Normal	6	1	7

$$Gini(Humidity = High) = 1 - (3/7)^2 - (4/7)^2$$
$$= 1 - 0.183 - 0.326 = 0.489$$

$$Gini(Humidity = Normal) = 1 - (6/7)^2 - (1/7)^2$$
$$= 1 - 0.734 - 0.02 = 0.244$$

Weighted sum for humidity feature will be calculated next

$$Gini(Humidity) = (7/14) * 0.489 + (7/14) * 0.244$$
$$= 0.36$$

Wind

Wind	Yes	No	Number of instances
Weak	6	2	8
Strong	3	3	6

$$Gini(Wind = Weak) = 1 - (6/8)^2 - (2/8)^2$$
$$= 1 - 0.5625 - 0.062 = 0.375$$

$$Gini(Wind = Strong) = 1 - (3/6)^2 - (3/6)^2$$
$$= 1 - 0.25 - 0.25 = 0.5$$

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

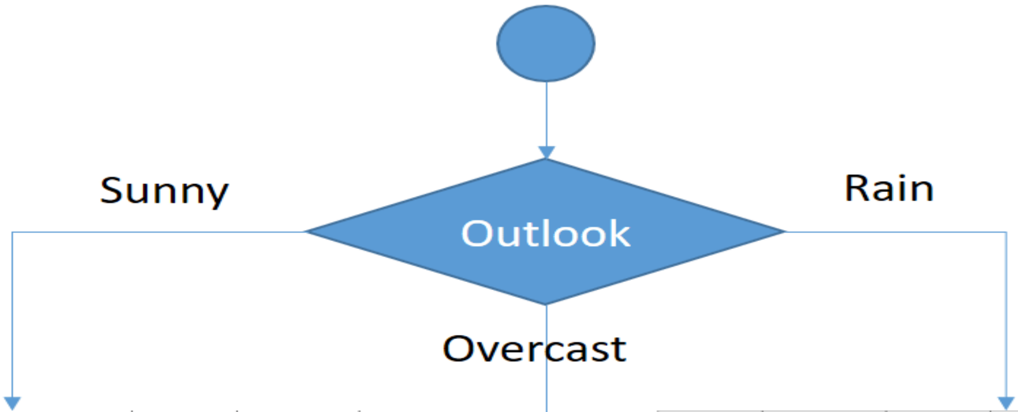
Time to decide

We’ve calculated gini index values for each feature. The winner will be outlook feature because its cost is the lowest.

Feature	Gini index
Outlook	**0.342**
Temperature	**0.439**

We’ll put outlook decision at the top of the tree.

First decision would be outlook feature



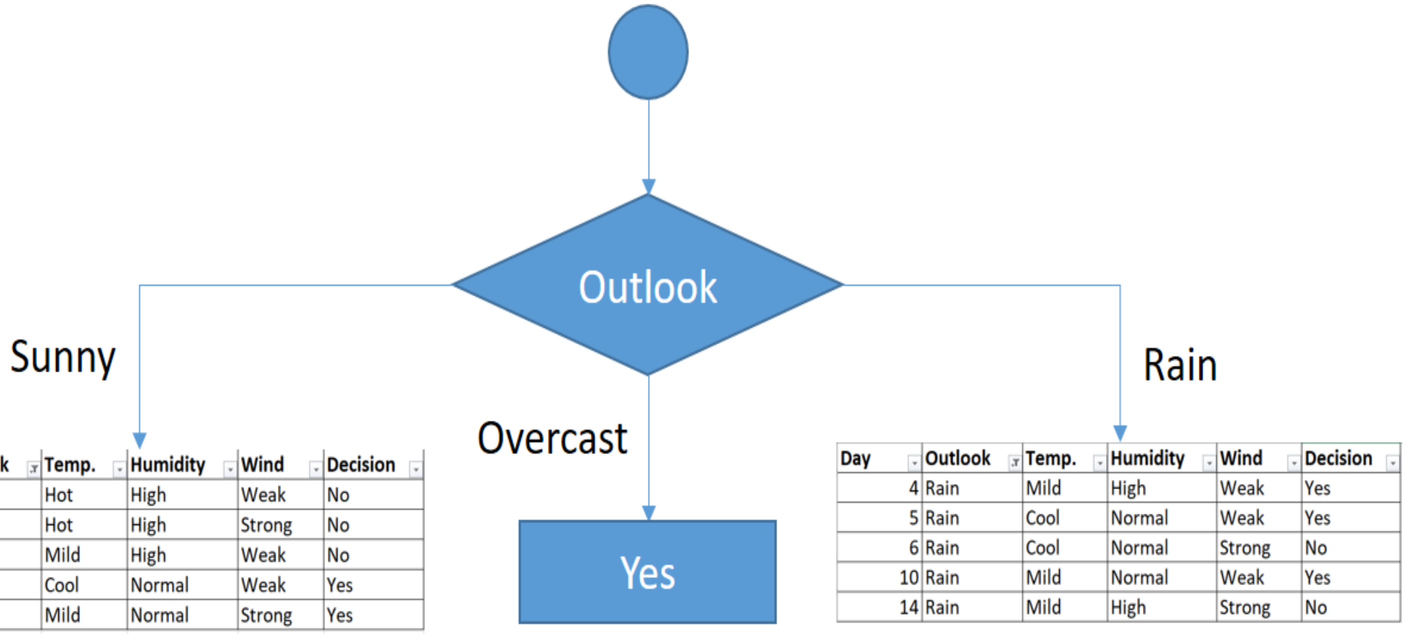
Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	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

Day	Outlook	Temp.	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

Tree is over for overcast outlook leaf

You might realize that sub dataset in the overcast leaf has only yes decisions. This means that overcast leaf is over.



Day	Outlook	Temp.	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

Day	Outlook	Temp.	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

Sub Datasets

We will apply same principles to those sub datasets in the following steps.
Focus on the sub dataset for sunny outlook. We need to find the gini index scores for temperature, humidity and wind features respectively.

Day	Outlook	Temp.	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 of temperature for sunny outlook

Temperature	Yes	No	Number of instances
Hot	0	2	2
Cool	1	0	1
Mild	1	1	2

$$Gini(Outlook = Sunny\ and\ Temp. = Hot)$$
$$= 1 - (0/2)^2 - (2/2)^2 = 0$$

$$Gini(Outlook = Sunny\ and\ Temp. = Cool)$$
$$= 1 - (1/1)^2 - (0/1)^2 = 0$$

$$Gini(Outlook = Sunny\ and\ Temp. = Mild)$$
$$= 1 - (1/2)^2 - (1/2)^2 = 1 - 0.25 - 0.25 = 0.5$$

$$Gini(Outlook = Sunny\ and\ Temp.)$$
$$= (2/5) * 0 + (1/5) * 0 + (2/5) * 0.5 = 0.2$$

Gini of humidity for sunny outlook

Humidity	Yes	No	Number of instances
High	0	3	3
Normal	2	0	2

$$Gini(Outlook = Sunny\ and\ Humidity = High)$$
$$= 1 - (0/3)^2 - (3/3)^2 = 0$$

$$Gini(Outlook = Sunny\ and\ Humidity = Normal)$$
$$= 1 - (2/2)^2 - (0/2)^2 = 0$$

•

$$Gini(Outlook = Sunny\ and\ Humidity)$$
$$= (3/5) * 0 + (2/5) * 0 = 0$$

Gini of wind for sunny outlook

Wind	Yes	No	Number of instances
Weak	1	2	3
Strong	1	1	2

$$Gini(Outlook = Sunny\ and\ Wind = Weak)$$
$$= 1 - (1/3)^2 - (2/3)^2 = 0.266$$

$$Gini(Outlook = Sunny\ and\ Wind = Strong)$$
$$= 1 - (1/2)^2 - (1/2)^2 = 0.2$$

.

$$Gini(Outlook = Sunny\ and\ Wind)$$
$$= (3/5) * 0.266 + (2/5) * 0.2 = 0.466$$

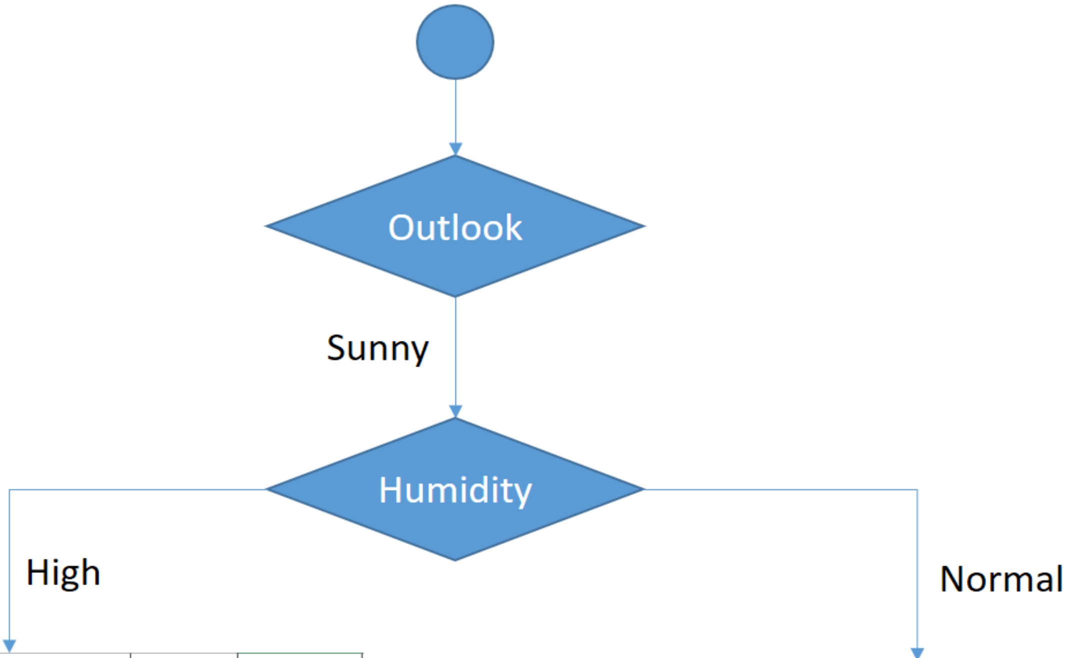
Decision for sunny outlook

We’ve calculated gini index scores for feature when outlook is sunny. The winner is humidity because it has the lowest value.

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

We’ll put humidity check at the extension of sunny outlook.

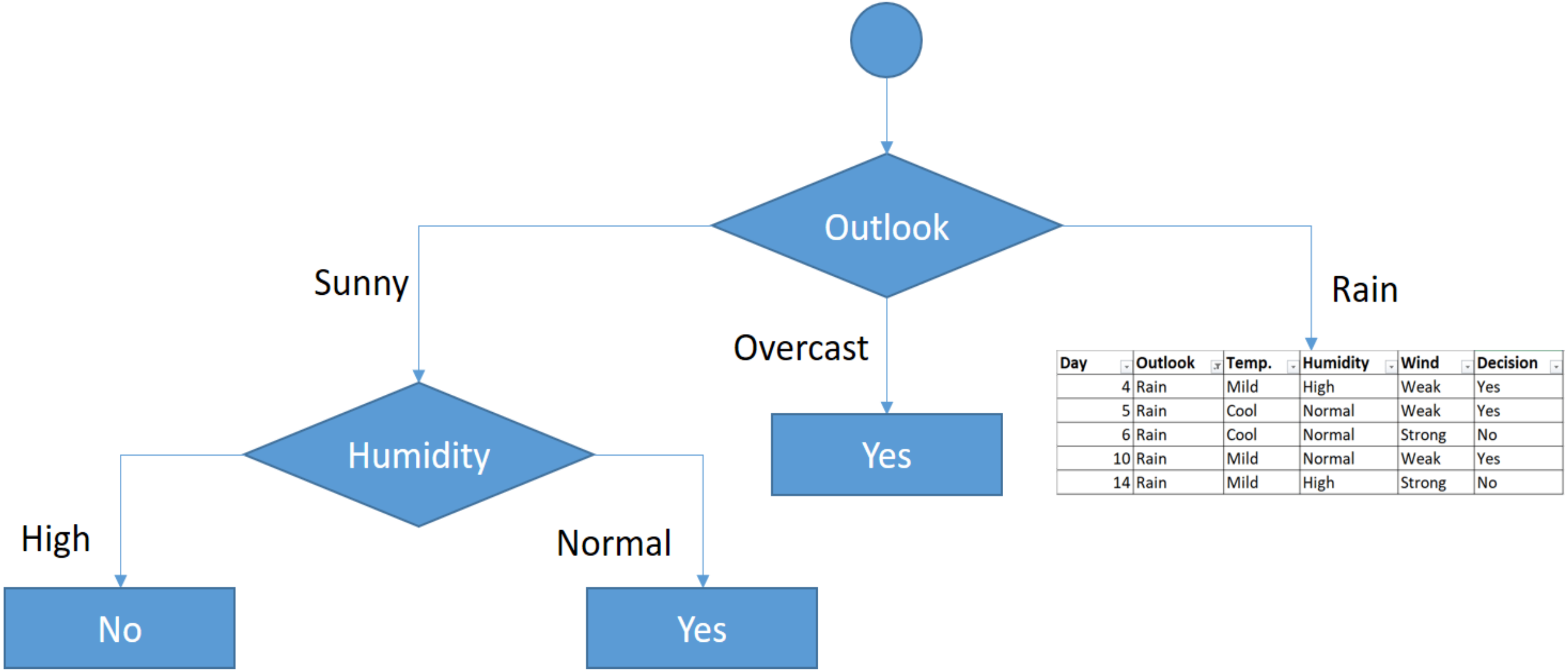
Sub datasets for high and normal humidity



Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Decisions for high and normal humidity



Day	Outlook	Temp.	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

Rain outlook

Now, we need to focus on rain outlook.

Day	Outlook	Temp.	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 of temprature for rain outlook

Temperature	Yes	No	Number of instances
Cool	1	1	2
Mild	2	1	3

$$Gini(Outlook = Rain\ and\ Temp. = Cool)$$
$$= 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(Outlook = Rain\ and\ Temp. = Mild)$$
$$= 1 - (2/3)^2 - (1/3)^2 = 0.444$$

$$Gini(Outlook = Rain\ and\ Temp.)$$

- $$= (2/5)x0.5 + (3/5)x0.444 = 0.466$$

Gini of humidity for rain outlook

Humidity	Yes	No	Number of instances
High	1	1	2
Normal	2	1	3

$$Gini(Outlook = Rain\ and\ Humidity = High)$$
$$= 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(Outlook = Rain\ and\ Humidity = Normal)$$
$$= 1 - (2/3)^2 - (1/3)^2 = 0.444$$

- $$Gini(Outlook = Rain\ and\ Humidity)$$
$$= (2/5) * 0.5 + (3/5) * 0.444 = 0.466$$

Gini of wind for rain outlook

Wind	Yes	No	Number of instances
Weak	3	0	3
Strong	0	2	2

$$Gini(Outlook = Rain\ and\ Wind = Weak)$$
$$= 1 - (3/3)^2 - (0/3)^2 = 0$$

$$Gini(Outlook = Rain\ and\ Wind = Strong)$$
$$= 1 - (0/2)^2 - (2/2)^2 = 0$$

$$Gini(Outlook = Rain\ and\ Wind)$$
$$= (3/5) * 0 + (2/5) * 0 = 0$$

Decision for rain outlook

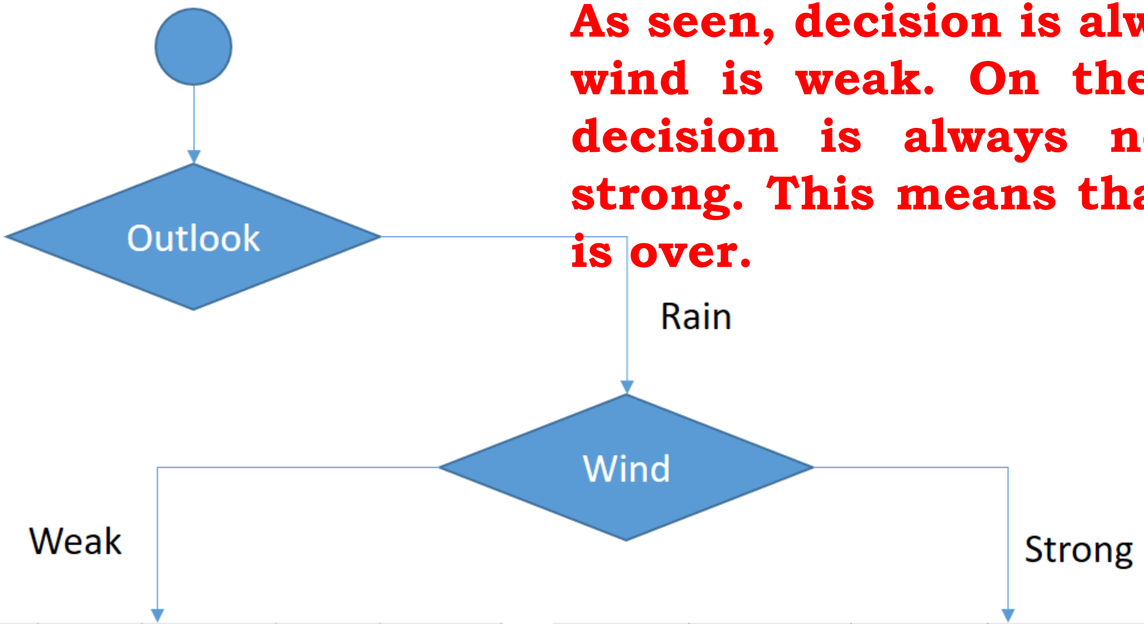
The winner is wind feature for rain outlook because it has the minimum gini index score in features.

Feature	Gini index
Temperature	0.466
Humidity	0.466
Wind	0

Put the wind feature for rain outlook branch and monitor the new sub data sets.

Sub data sets for weak and strong wind and rain outlook

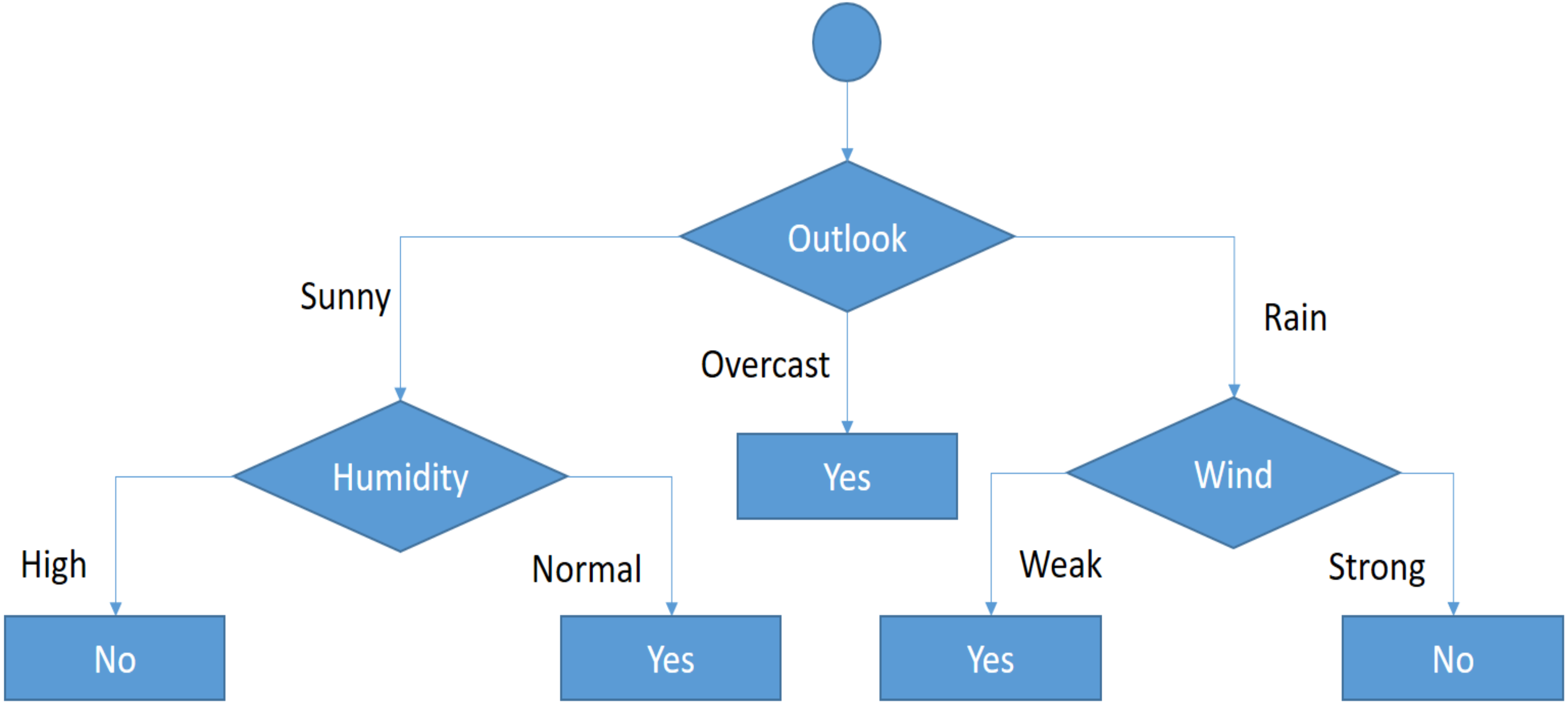
As seen, decision is always yes when wind is weak. On the other hand, decision is always no if wind is strong. This means that this branch is over.



Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

Final form of the decision tree built by CART algorithm



Final form of the decision tree built by CART algorithm