Decision Trees - CART

Indeed Inspiring Infotech

www.indeedinspiring.com

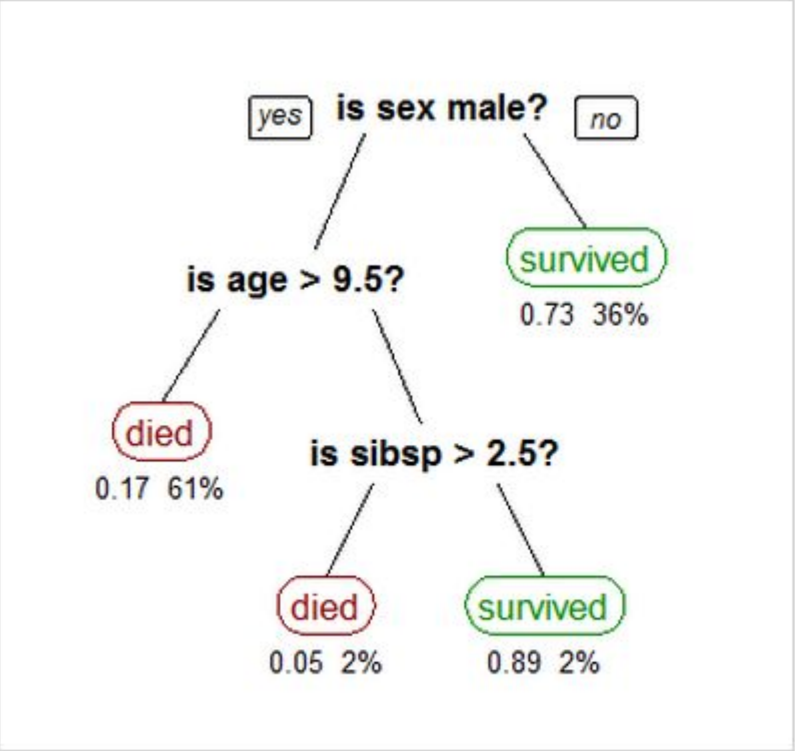
Introduction

Decision Trees are commonly used with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables).

The CART or Classification & Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees:

- **Classification Trees**: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.
- **Regression Trees**: where the target variable is continuous and tree is used to predict it's value.

- The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be.
- The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.
- A simple example of a decision tree is as given in figure.



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

The main elements of CART (and any decision tree algorithm) are:

- Rules for splitting data at a node based on the value of one variable;
- Stopping rules for deciding when a branch is terminal and can be split no more; and
- Finally, a prediction for the target variable in each terminal node.

Stopping Rules

Stopping rules control if the tree growing process should be stopped or not. The following stopping rules are used:

- If a node becomes pure; that is, all cases in a node have identical values of the dependent variable, the node will not be split.
- If all cases in a node have identical values for each predictor, the node will not be split.
- If the current tree depth reaches the user-specified maximum tree depth limit value, the tree growing process will stop.
- If the size of a node is less than the user-specified minimum node size value, the node will not be split.
- If the split of a node results in a child node whose node size is less than the user specified minimum child node size value, the node will not be split.

GINI Index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- Favors larger partitions.
- Uses squared proportion of classes.
- Perfectly classified, Gini Index would be zero.
- Evenly distributed would be 1 – (1/# Classes).
- You want a variable split that has a low Gini Index.
- The algorithm works as 1 – (P(class1)^2 + P(class2)^2 + … + P(classN)^2)
- The Gini index is used in the classic CART algorithm and is very easy to calculate.

Example

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

There are 14 instances of golf playing decisions based on outlook, temperature, humidity and wind factors.

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

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Outlook is a nominal feature. It can be sunny, overcast or rain. We will summarize the final decisions for outlook feature.

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

Then, we will calculate weighted sum of gini indexes for outlook feature.

Gini(Outlook) = (5/14) x 0.48 + (4/14) x 0 + (5/14) x 0.48 = 0.171 + 0 + 0.171 = **0.342**

Temperature

Similarly, temperature is a nominal feature and it could have 3 different values: Cool, Hot and Mild. Let’s summarize decisions for temperature feature.

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

We’ll calculate weighted sum of gini index for temperature feature

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 is a binary class feature. It can be high or normal.

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) x 0.489 + (7/14) x 0.244 = **0.367**

Wind

Wind is a binary class similar to humidity. It can be weak and strong.

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) = $1 – (3/6)^2 – (3/6)^2 = 1 – 0.25 – 0.25 = 0.5$

Weighted sum for Wind feature

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

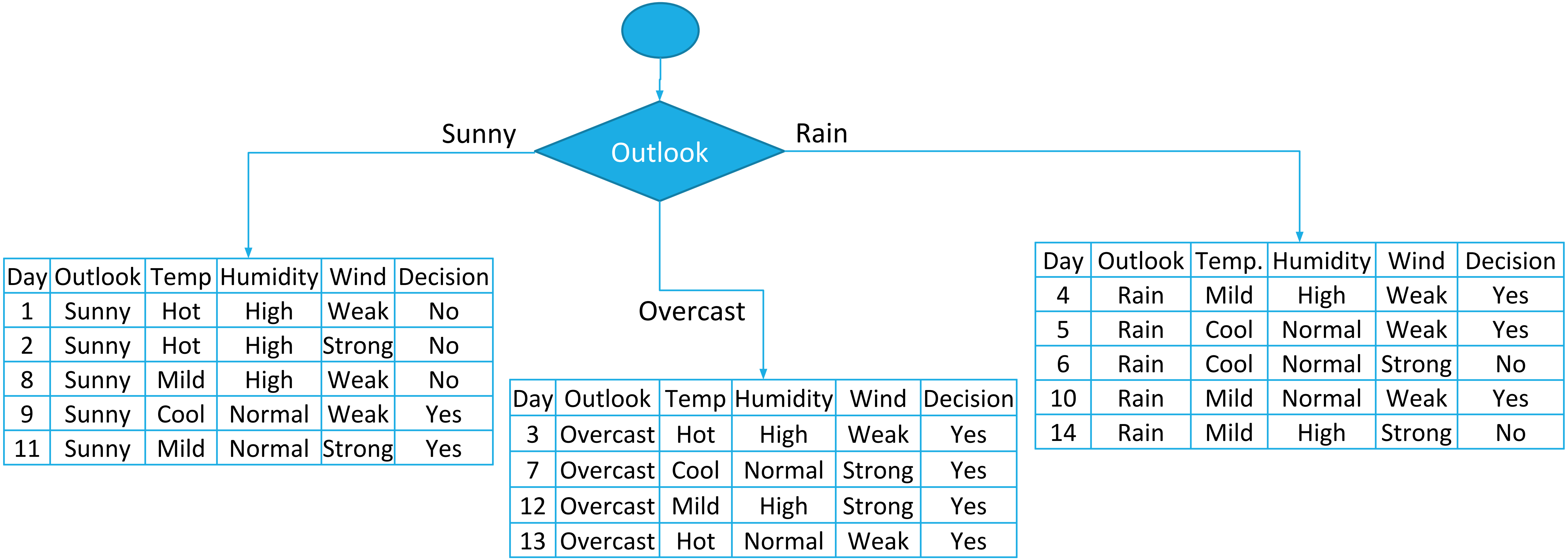
Time To decide

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
Humidity	0.367
Wind	0.428

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

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

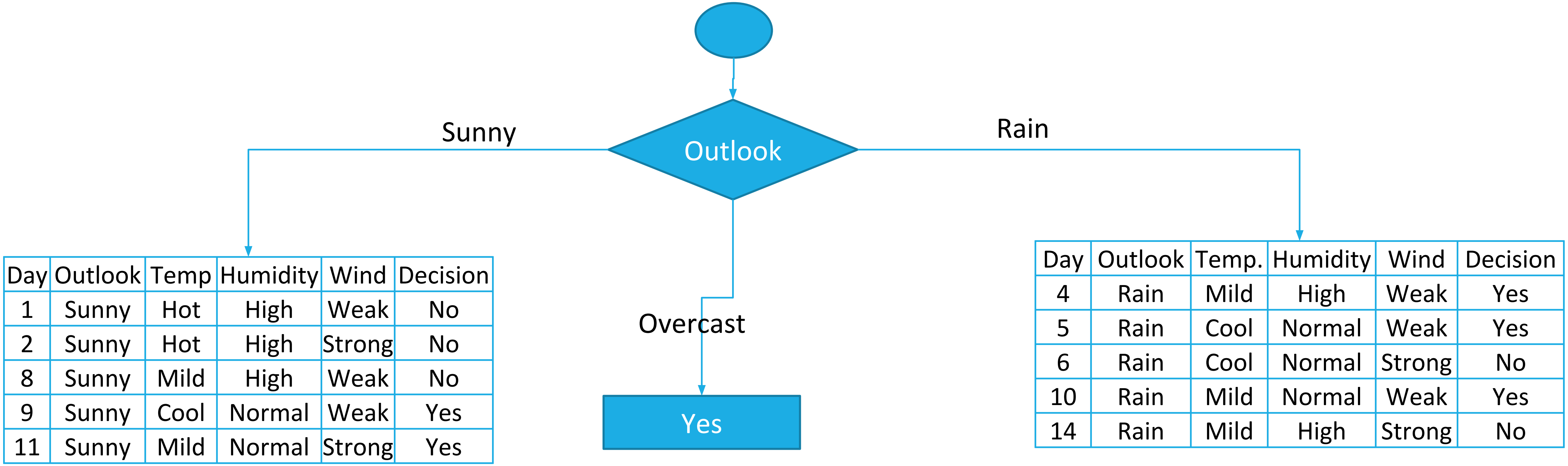


Day	Outlook	Temp	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	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

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

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) = $1 – (1/2)^2 – (1/2)^2 = 1 – 0.25 – 0.25 = 0.5$

Weighted Gini Index,

Gini(Outlook = Sunny and Temp.) = (2/5)x0 + (1/5)x0 + (2/5)x0.5 = 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) = $1 – (2/2)^2 – (0/2)^2 = 0$

Weighted Gini Index,

Gini(Outlook=Sunny and Humidity) = (3/5)x0 + (2/5)x0 = 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

Weighted Gini Index,

Gini(Outlook=Sunny and Wind) = (3/5)x0.266 + (2/5)x0.2 = 0.466

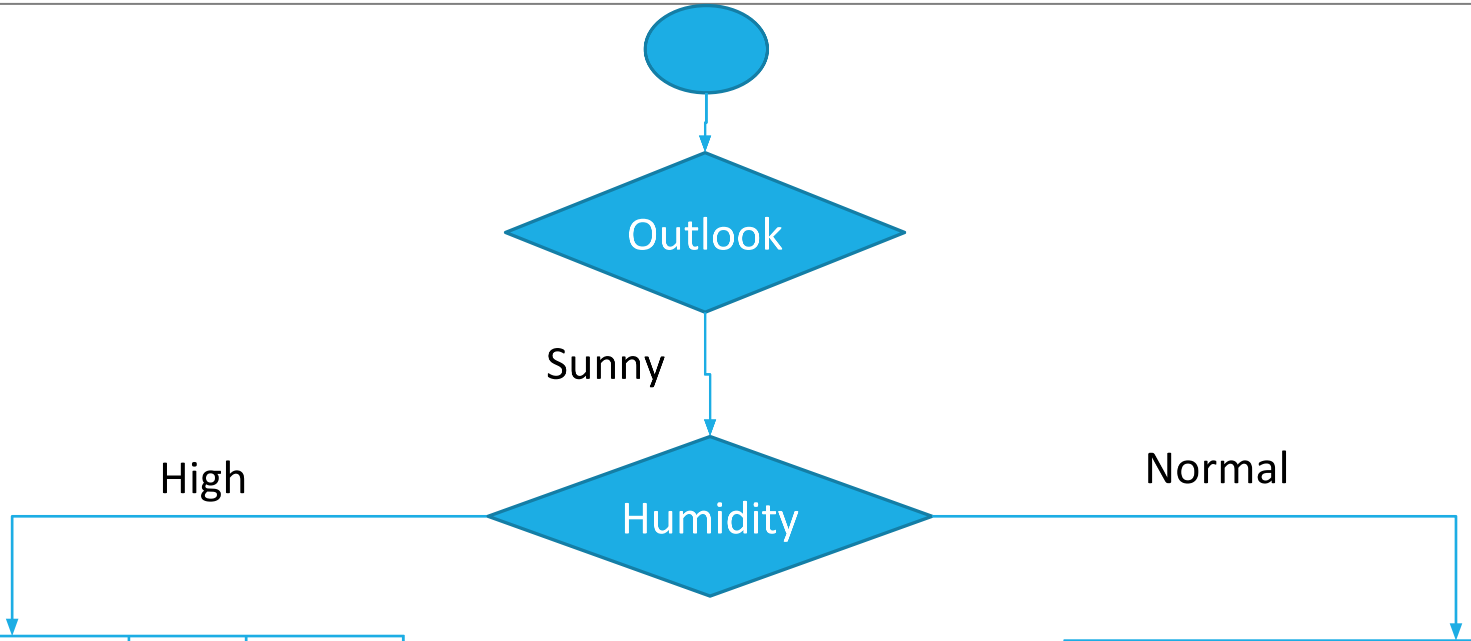
Decision for sunny outlook

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.

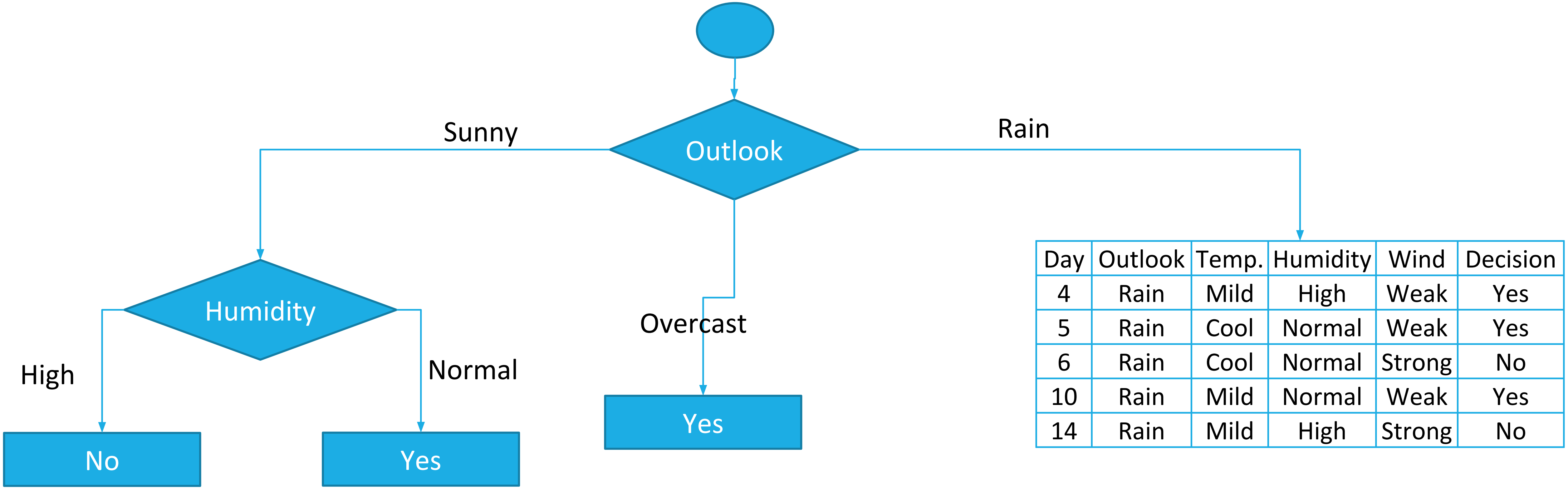
We’ll put humidity check at the extension of sunny outlook.



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

As seen, decision is always no for high humidity and sunny outlook. On the other hand, decision will always be yes for normal humidity and sunny outlook. This branch is over.



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

We'll calculate gini index scores for temperature, humidity and wind features when outlook is rain.

Gini of temperature 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) = $1 – (2/3)^2 – (1/3)^2 = 0.444$

Weighted Gini Index,

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

Weighted Gini Index,

Gini(Outlook=Rain and Humidity) = (2/5)x0.5 + (3/5)x0.444 = 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) = $1 – (0/2)^2 – (2/2)^2 = 0$

Weighted Gini Index,

Gini(Outlook=Rain and Wind) = (3/5)x0 + (2/5)x0 = 0

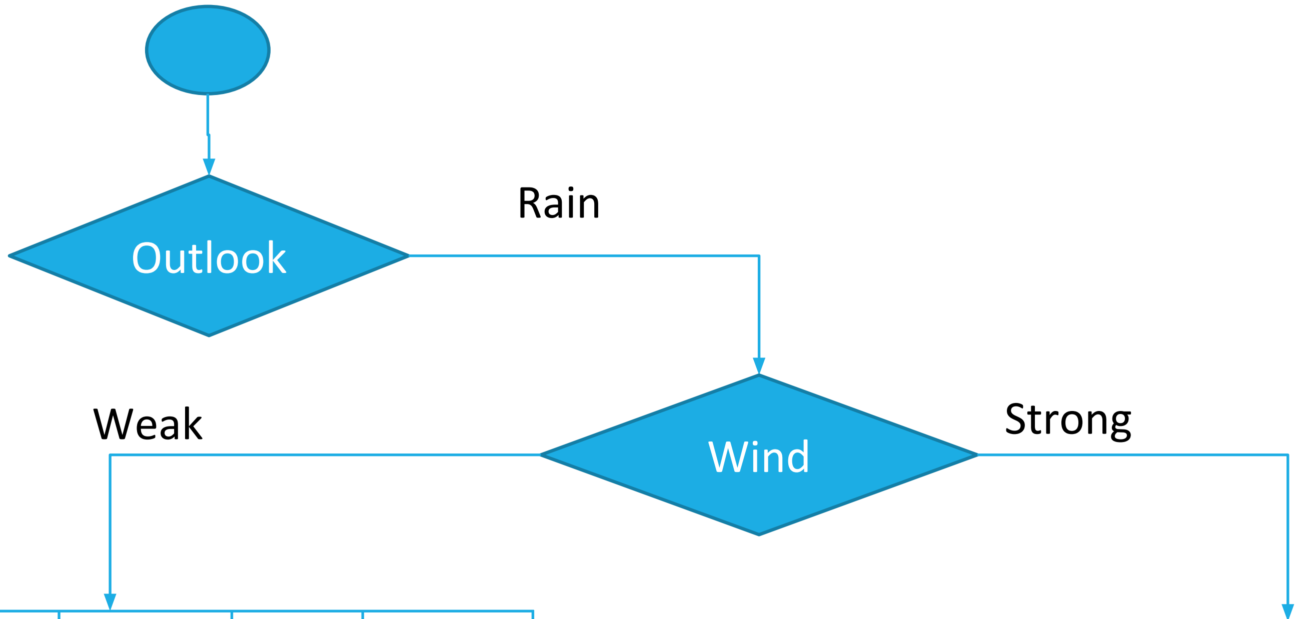
Decision for rain outlook

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.

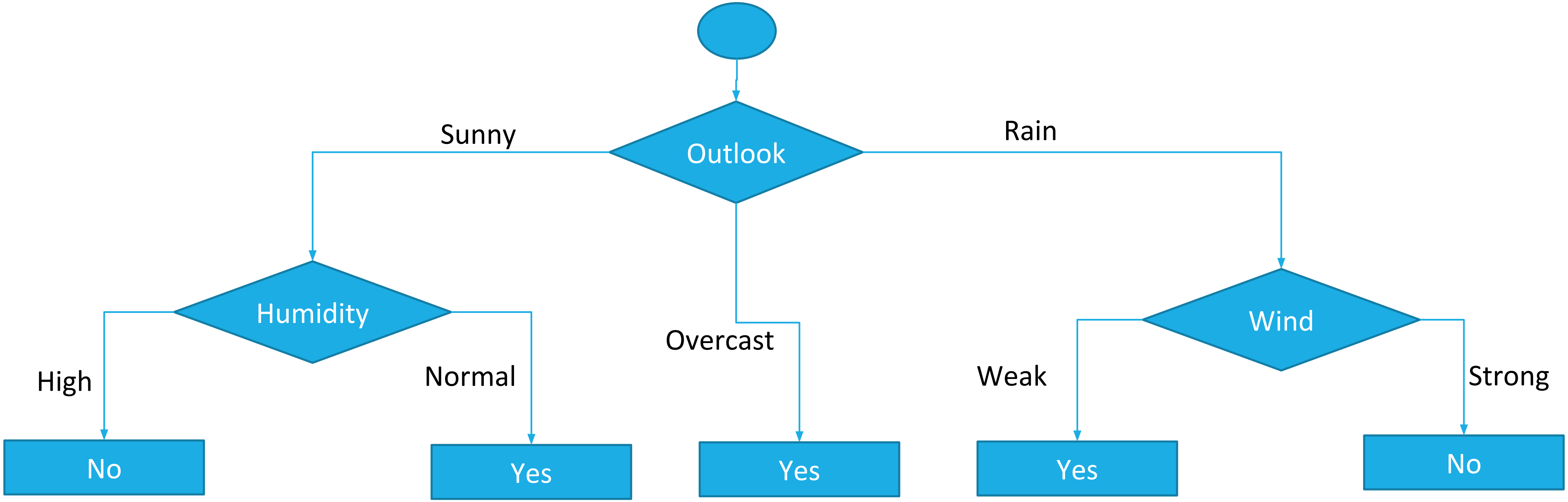
Put the wind feature for rain outlook branch and monitor the new sub data sets.



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

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.



Properties

- CART is nonparametric and therefore does not rely on data belonging to a particular type of distribution.
- CART is not significantly impacted by outliers in the input variables.
- You can relax stopping rules to "overgrow" decision trees and then prune back the tree to the optimal size. This approach minimizes the probability that important structure in the data set will be overlooked by stopping too soon.
- CART incorporates both testing with a test data set and cross-validation to assess the goodness of fit more accurately.
- CART can use the same variables more than once in different parts of the tree. This capability can uncover complex interdependencies between sets of variables.
- CART can be used in conjunction with other prediction methods to select the input set of variables.