

Supervised Learning with Classification

Decision Tree - Classification

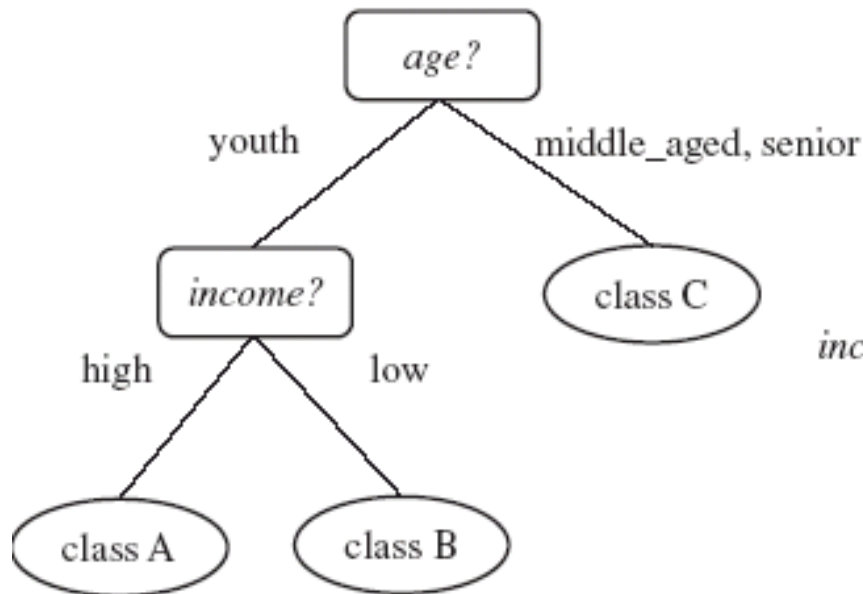
- Decision tree builds classification models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- The final result is a tree with **decision nodes** and **leaf nodes**.
- A decision node has two or more branches
- Leaf node represents a classification or decision.
- The topmost decision node in a tree which corresponds to the best predictor called **root node**.
- Decision trees can handle both categorical and numerical data.

Classification Model

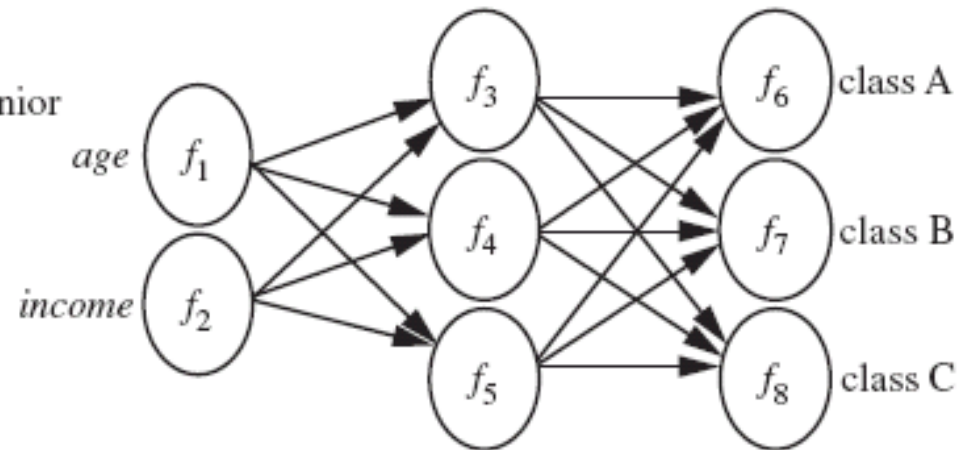
(a)

$\text{age}(X, \text{"youth"}) \text{ AND } \text{income}(X, \text{"high"}) \longrightarrow \text{class}(X, \text{"A"})$
 $\text{age}(X, \text{"youth"}) \text{ AND } \text{income}(X, \text{"low"}) \longrightarrow \text{class}(X, \text{"B"})$
 $\text{age}(X, \text{"middle_aged"}) \longrightarrow \text{class}(X, \text{"C"})$
 $\text{age}(X, \text{"senior"}) \longrightarrow \text{class}(X, \text{"C"})$

(b)



(c)



What is node impurity/purity in decision trees?

- The decision tree is a greedy algorithm that performs a recursive binary partitioning of the feature space.
- The tree predicts the same label for each bottommost (leaf) partition.
- Each partition is chosen greedily by selecting the *best split* from a set of possible splits.

Consider an example as the set of atoms in a metallic ball

- If all of the ball's atoms were gold - you would say that the ball is purely gold, and that its purity level is highest (and its impurity level is lowest).
- Similarly, if all of the examples in the set were of the same class, then the set's purity would be highest.
- If $\frac{1}{3}$ of the atoms were gold, $\frac{1}{3}$ silver, and $\frac{1}{3}$ iron - you would say that for a ball made of 3 kinds of atoms, its purity is lowest.
- Similarly, if the examples are split evenly between all of the classes, then the set's purity is lowest.
- So the purity of a set of examples is the homogeneity of its examples - with regard to their classes.
- Gini index is one of the popular measures of impurity

ISSUES IN DECISION TREE LEARNING

1. Overfitting the Data
2. Incorporating Continuous valued attributes
3. Handling training examples with missing attribute values
4. Handling attributes with different costs
5. Alternative measures for selecting attributes

CART Algorithm

- CART Algorithm is an abbreviation of **C**lassification **A**nd **R**egression **T**rees.
- Rather than general trees that could have multiple branches, CART makes use binary tree, which has only two branches from each node.
- CART use Gini Impurity as the criterion to split node, not Information Gain.
- CART supports numerical target variables, which enables itself to become a Regression Tree that predicts continuous values.
- Just like the ID3 and C4.5 algorithms that rely on Information Gain as the criterion to split nodes, the CART algorithm makes use another criterion called Gini to split the nodes.

CART Algorithm

- In CART algorithm it is intuitively using the Gini coefficient for a similar purpose. That is, the larger Gini coefficient means the larger impurity of the node.
- Similar to ID3 and C4.5 using Information Gain to select the node with more uncertainty, the Gini coefficient will guide the CART algorithm to find the node with larger uncertainty (i.e. impurity) and then split it.
- Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.
- It means an attribute with lower Gini index should be preferred.
- Sklearn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.

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

Example 3

PlayTennis: training examples

↓(Target attribute)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Outlook

Outlook is a nominal feature. It can be sunny, overcast or rain. I 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

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

$$\text{Gini}(\text{Outlook}=\text{Overcast}) = 1 - (4/4)^2 - (0/4)^2 = 0$$

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

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

$$\text{Gini}(\text{Outlook}) = (5/14) \times 0.48 + (4/14) \times 0 + (5/14) \times 0.48 = 0.171 + 0 + 0.171 = 0.342$$

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

$$\text{Gini}(\text{Temp}=\text{Hot}) = 1 - (2/4)^2 - (2/4)^2 = 0.5$$

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

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

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

$$\text{Gini}(\text{Temp}) = (4/14) \times 0.5 + (4/14) \times 0.375 + (6/14) \times 0.445 = 0.142 + 0.107 + 0.190 = 0.439$$

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

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

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

Weighted sum for humidity feature will be calculated next

$$\text{Gini}(\text{Humidity}) = (7/14) \times 0.489 + (7/14) \times 0.244 = 0.367$$

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

$$\text{Gini}(\text{Wind}=\text{Weak}) = 1 - (6/8)^2 - (2/8)^2 = 1 - 0.5625 - 0.0625 = 0.375$$

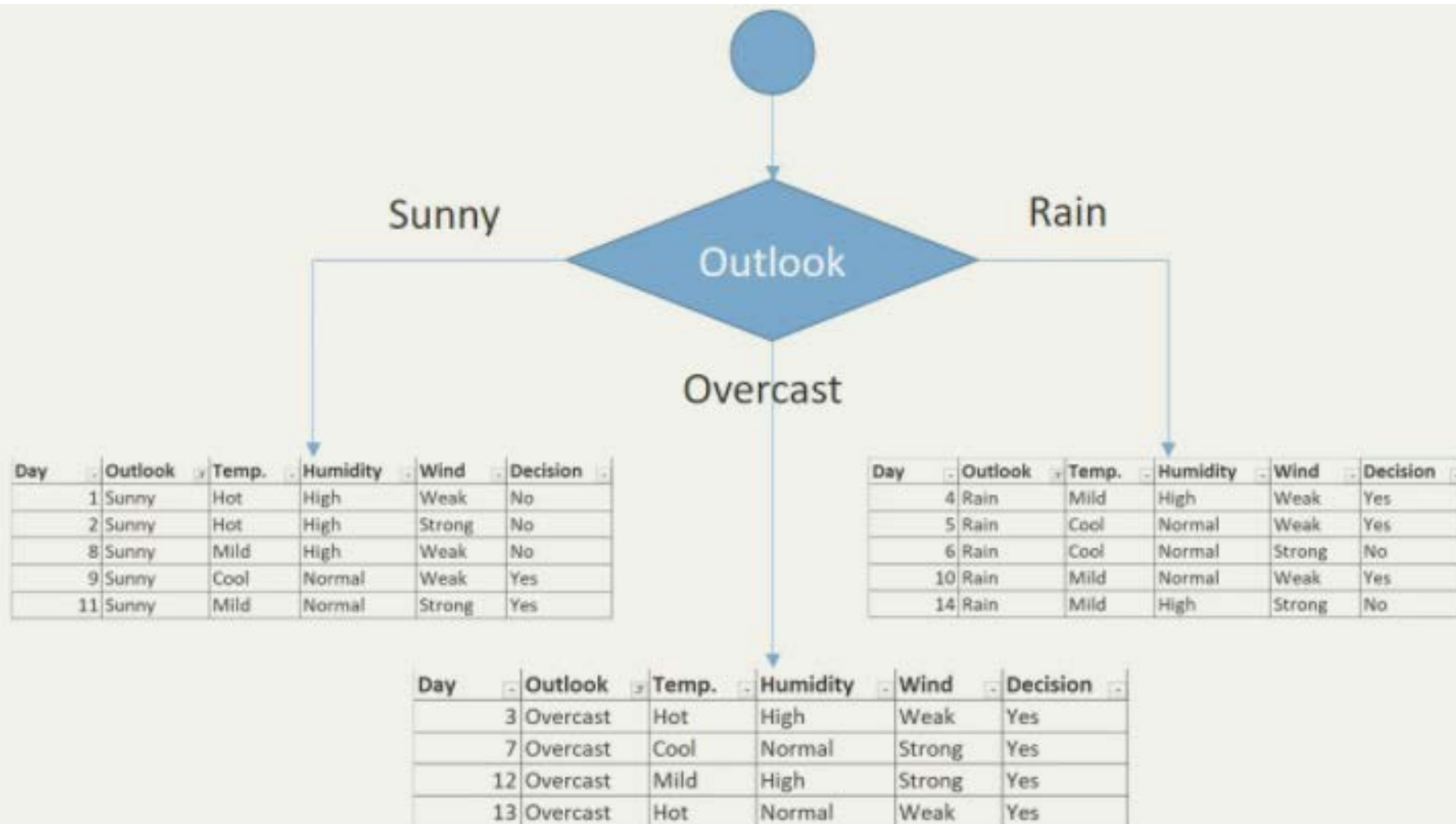
$$\text{Gini}(\text{Wind}=\text{Strong}) = 1 - (3/6)^2 - (3/6)^2 = 1 - 0.25 - 0.25 = 0.5$$

$$\text{Gini}(\text{Wind}) = (8/14) \times 0.375 + (6/14) \times 0.5 = 0.428$$

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





Tree is over for overcast outlook leaf

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

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

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

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

$$\text{Gini}(\text{Outlook}=\text{Sunny and Temp.}) = (2/5) \times 0 + (1/5) \times 0 + (2/5) \times 0.5 = 0.2$$

Gini of humidity for sunny outlook

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

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

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

$$\text{Gini}(\text{Outlook}=\text{Sunny and Humidity}) = (3/5) \times 0 + (2/5) \times 0 = 0$$

Gini of wind for sunny outlook

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

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

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

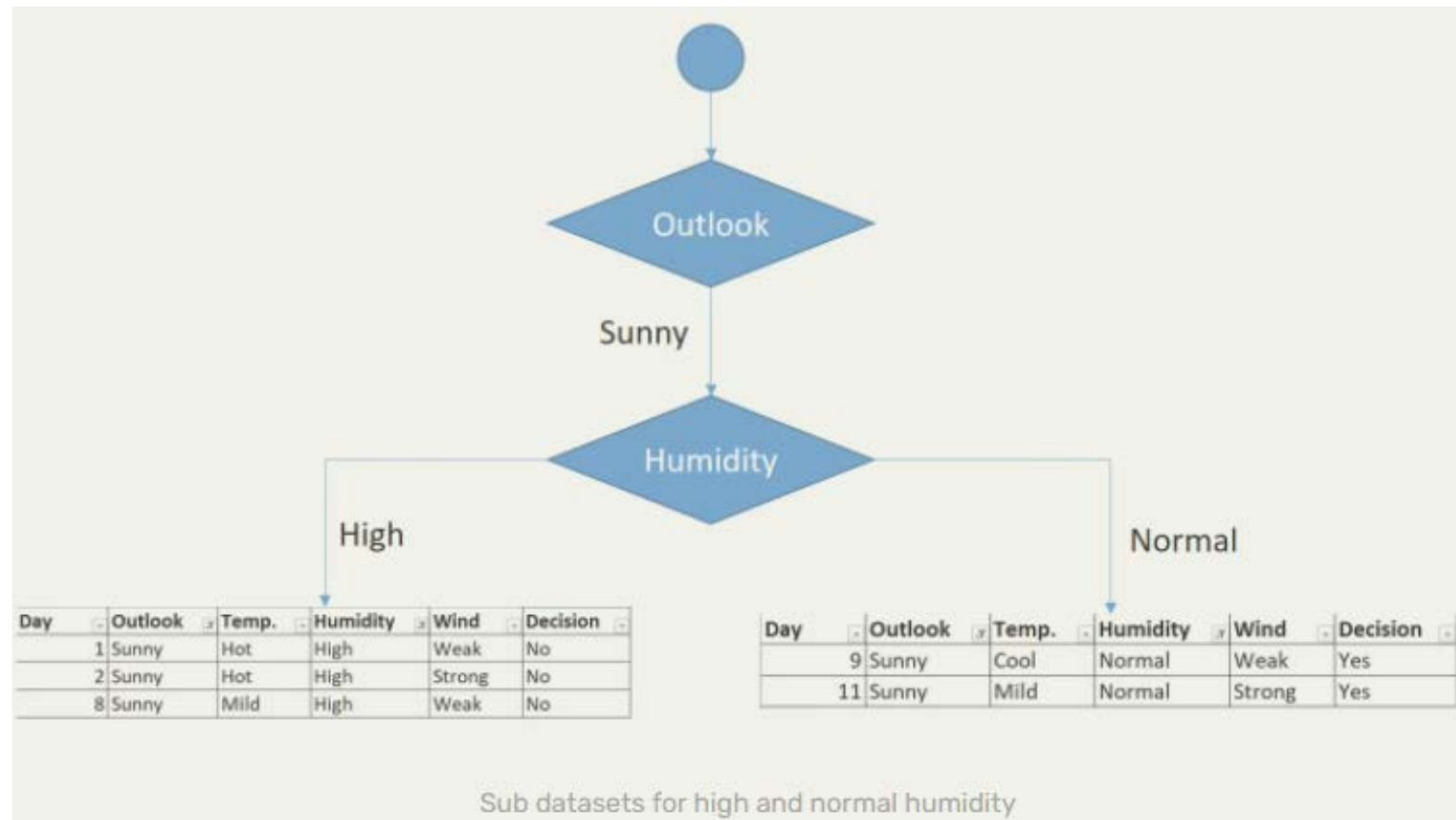
$$\text{Gini}(\text{Outlook}=\text{Sunny and Wind}) = (3/5) \times 0.266 + (2/5) \times 0.2 = 0.466$$

Decision for sunny outlook

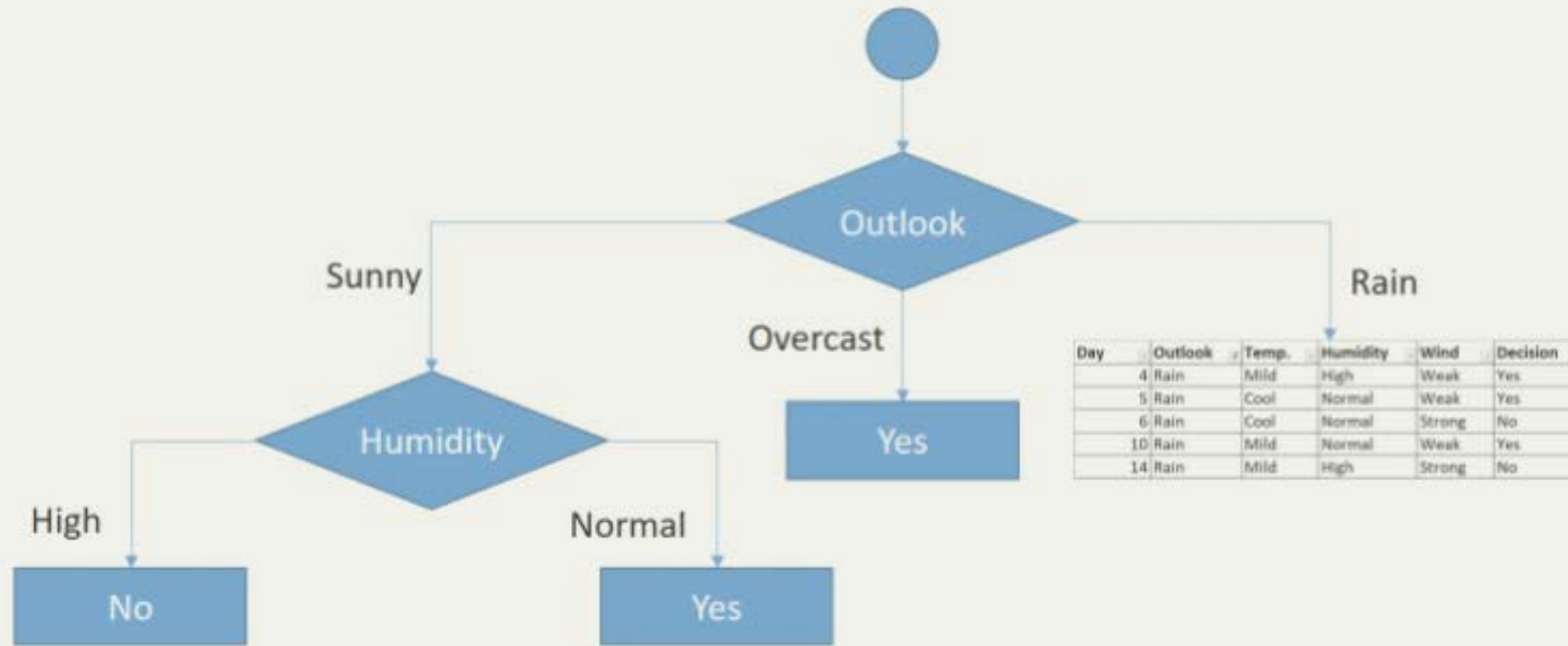
We've calculated gini index scores for feature when outlook is sunny. The winner is humidity because it has the lowest value.

Feature	Gini index
Temperature	0.2
Humidity	0
Wind	0.466

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



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.



Decisions for high and normal humidity

Now, we need to focus on rain outlook.

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 temprature for rain outlook

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

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

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

$$\text{Gini}(\text{Outlook}=\text{Rain and Temp.}) = (2/5) \times 0.5 + (3/5) \times 0.444 = 0.466$$

Gini of humidity for rain outlook

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

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

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

$$\text{Gini}(\text{Outlook}=\text{Rain and Humidity}) = (2/5) \times 0.5 + (3/5) \times 0.444 = 0.466$$

Gini of wind for rain outlook

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

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

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

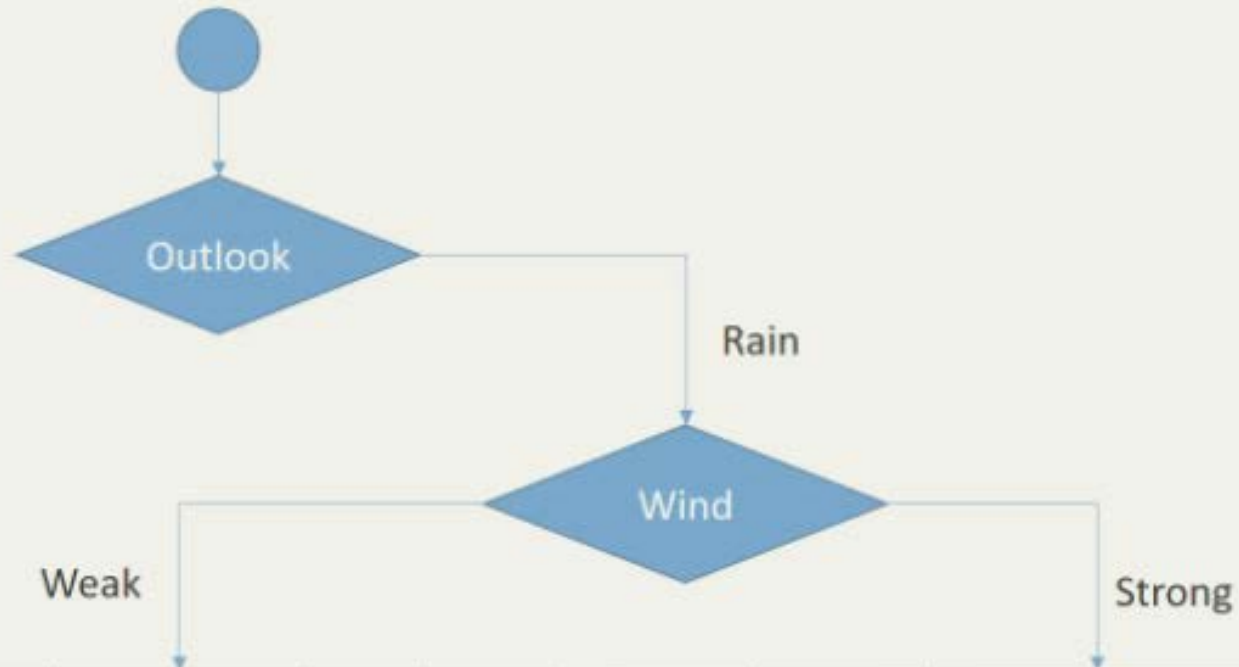
$$\text{Gini}(\text{Outlook}=\text{Rain and Wind}) = (3/5) \times 0 + (2/5) \times 0 = 0$$

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

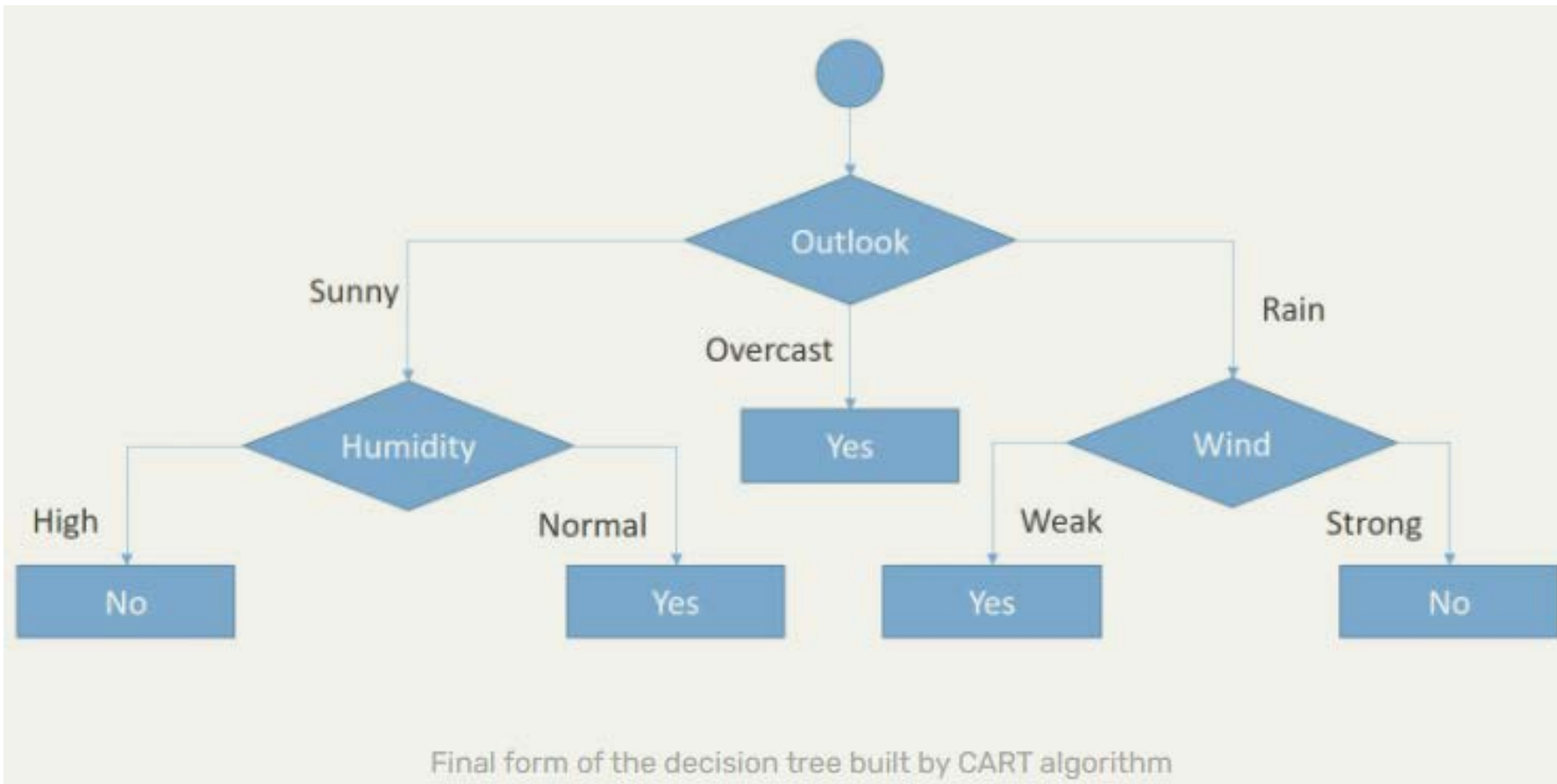
Put the wind feature for rain outlook branch and monitor the new sub data sets.



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.



Thank you