

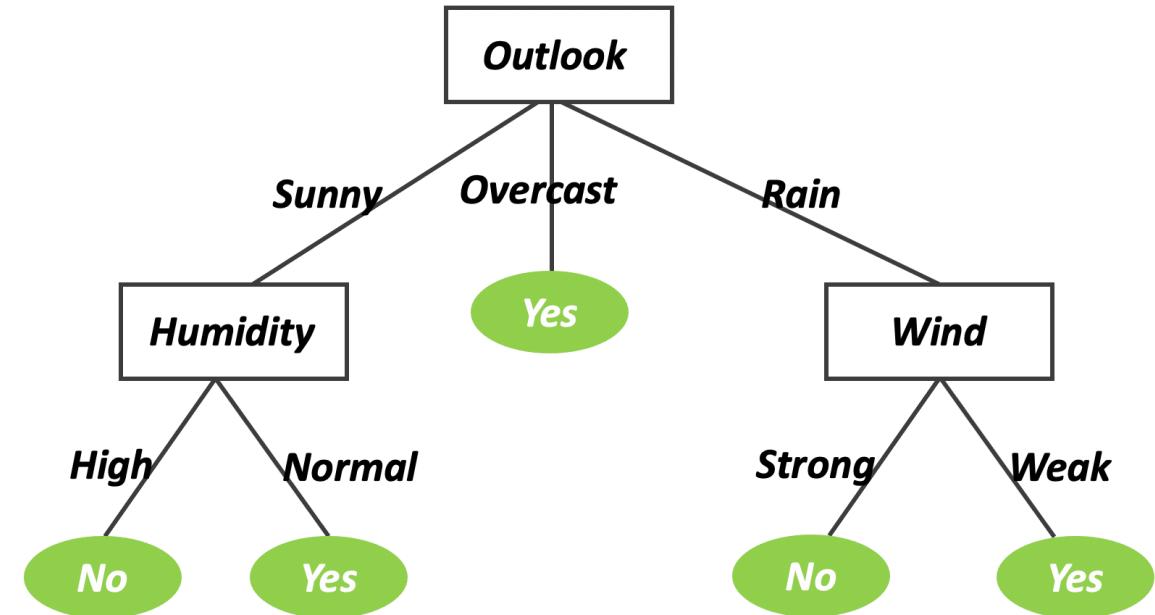
Rule-Based Prediction Algorithms

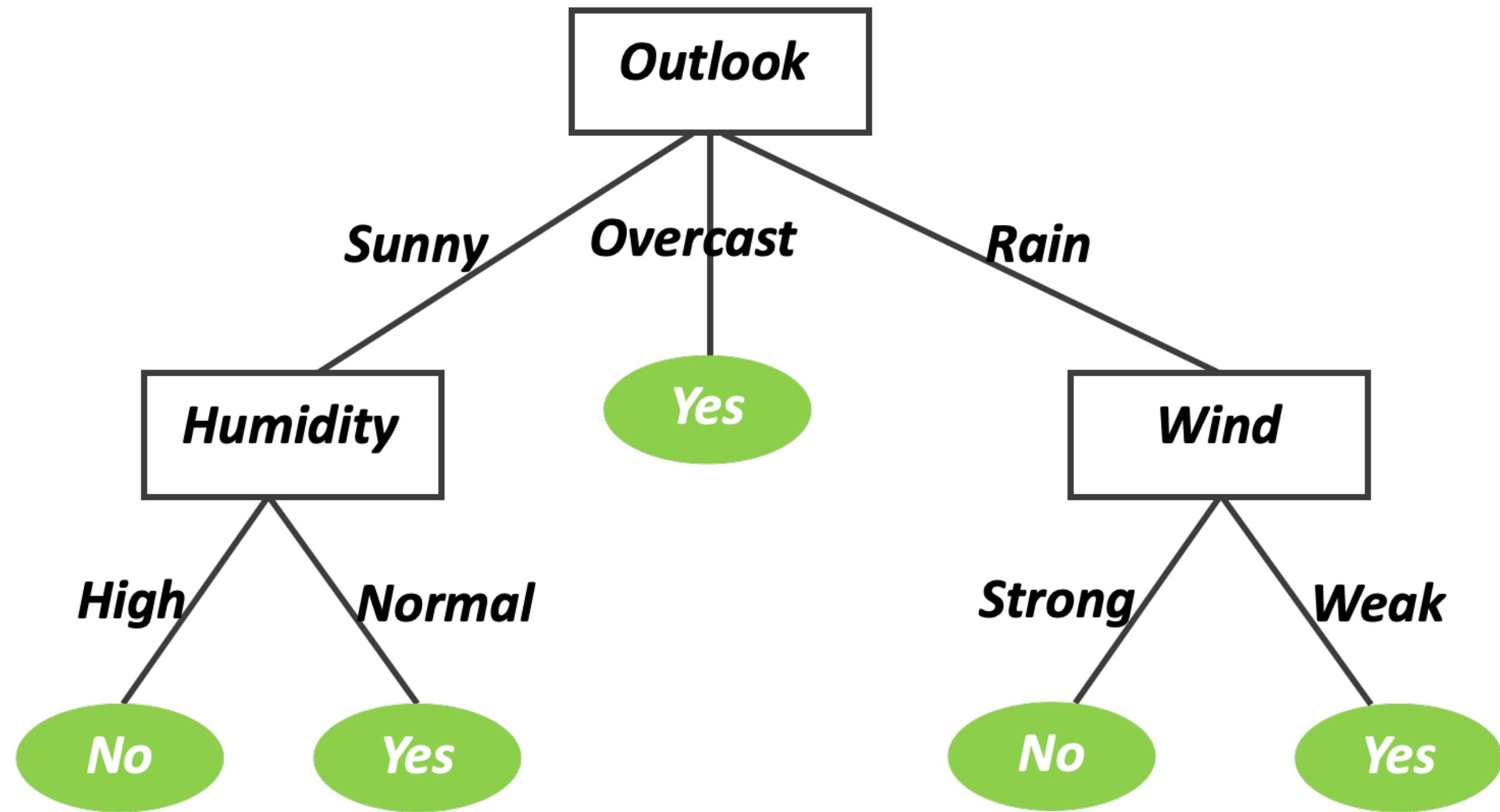
Decision Trees

Classification

Example: how to draw a decision tree using ID3 algorithm

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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





Steps in building the Decision Tree

- We want to know which of the attributes gives the maximum information out of the available attributes (independent variables/features).
- For that we will need to calculate the **information gain** for every attribute.
- The attribute that has the maximum information gain will be considered as a root node.
- Then we will continue building the tree.

To calculate the Information Gain of each attribute:

- Calculate the entropy (uncertainty) of the whole dataset.
- Calculate the entropy of the individual attribute values.
 - If we're referring to the **Outlook** attribute, the values will be *sunny*, *overcast*, and *rain*.
- Calculate the weighted sum of the entropy for each value
- The **information gain** for the attribute will be the entropy of the parent minus the weighted entropy of each attribute value.

Step 1: Consider the first attribute **Outlook**

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Outlook) = Sunny, Overcast, Rain

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{\text{Sunny}} \leftarrow [2+, 3-]$$

$$\text{Entropy}(S_{\text{Sunny}}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971$$

$$S_{\text{Overcast}} \leftarrow [4+, 0-]$$

$$\text{Entropy}(S_{\text{Overcast}}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} = 0$$

$$S_{\text{Rain}} \leftarrow [3+, 2-]$$

$$\text{Entropy}(S_{\text{Rain}}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \sum_{v \in \{\text{Sunny}, \text{Overcast}, \text{Rain}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \frac{5}{14} \text{Entropy}(S_{\text{Sunny}})$$

$$-\frac{4}{14} \text{Entropy}(S_{\text{Overcast}}) - \frac{5}{14} \text{Entropy}(S_{\text{Rain}})$$

$$\text{Gain}(S, \text{Outlook}) = 0.94 - \frac{5}{14} 0.971 - \frac{4}{14} 0 - \frac{5}{14} 0.971 = \underline{0.2464}$$

Step 2: attribute Temp

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Temp) = Hot, Mild, Cool

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Hot} \leftarrow [2+, 2-]$$

$$\text{Entropy}(S_{Hot}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1.0$$

$$S_{Mild} \leftarrow [4+, 2-]$$

$$\text{Entropy}(S_{Mild}) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} \leftarrow [3+, 1-]$$

$$\text{Entropy}(S_{Cool}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \sum_{v \in \{\text{Hot}, \text{Mild}, \text{Cool}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \frac{4}{14} \text{Entropy}(S_{Hot})$$

$$-\frac{6}{14} \text{Entropy}(S_{Mild}) - \frac{4}{14} \text{Entropy}(S_{Cool})$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.0289$$

Step 3: attribute Humidity

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Humidity) = High, Normal

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{\text{High}} \leftarrow [3+, 4-]$$

$$\text{Entropy}(S_{\text{High}}) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.9852$$

$$S_{\text{Normal}} \leftarrow [6+, 1-]$$

$$\text{Entropy}(S_{\text{Normal}}) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.5916$$

$$\text{Gain}(S, \text{Humidity}) = \text{Entropy}(S) - \sum_{v \in \{\text{High}, \text{Normal}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Humidity}) = \text{Entropy}(S) - \frac{7}{14} \text{Entropy}(S_{\text{High}})$$

$$-\frac{7}{14} \text{Entropy}(S_{\text{Normal}})$$

$$\text{Gain}(S, \text{Humidity}) = 0.94 - \frac{7}{14} 0.9852 - \frac{7}{14} 0.5916 = \underline{0.1516}$$

Step 4: attribute **Wind**

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Wind) = Strong, Weak

$$S = [9+, 5-]$$

$$\text{Entropy}(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$\text{Entropy}(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$\text{Entropy}(S_{Weak}) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.8113$$

$$\text{Gain}(S, \text{Wind}) = \text{Entropy}(S) - \sum_{v \in \{\text{Strong}, \text{Weak}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$\text{Gain}(S, \text{Wind}) = \text{Entropy}(S) - \frac{6}{14} \text{Entropy}(S_{Strong}) - \frac{8}{14} \text{Entropy}(S_{Weak})$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{6}{14} 1.0 - \frac{8}{14} 0.8113 = \underline{\underline{0.0478}}$$

Which attribute has the maximum Information Gain?

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

$$Gain(S, Outlook) = 0.2464$$

Will be the root node

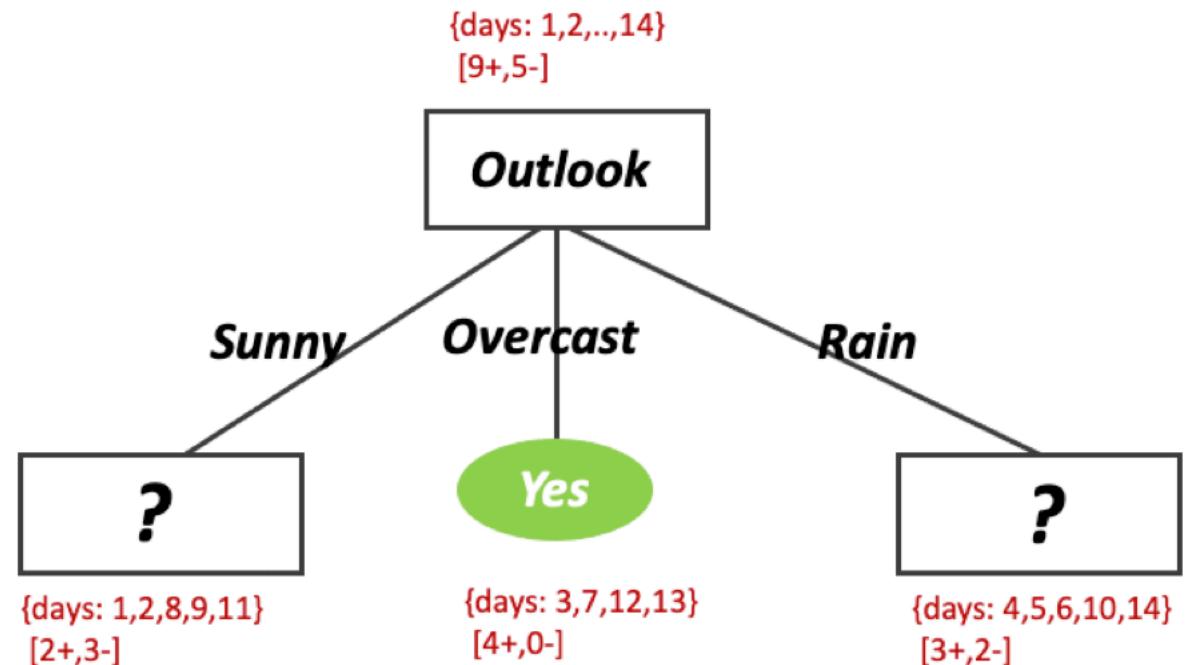
$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Humidity) = 0.1516$$

$$Gain(S, Wind) = 0.0478$$

The root node: Outlook

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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 (Sunny days):

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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



Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Outlook (Sunny days), Attribute: Temp

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Temp) = Hot, Mild, Cool

$$S_{Sunny} = [2+, 3-] \quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Hot} \leftarrow [0+, 2-] \quad Entropy(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [1+, 1-] \quad Entropy(S_{Mild}) = 1.0$$

$$S_{Cool} \leftarrow [1+, 0-] \quad Entropy(S_{Cool}) = 0.0$$

$$Gain(S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp) = Entropy(S) - \frac{2}{5}Entropy(S_{Hot})$$

$$-\frac{2}{5}Entropy(S_{Mild}) - \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S_{Sunny}, Temp) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = \underline{\underline{0.570}}$$

Outlook (Sunny days), Attribute: Humidity

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Humidity) = High, Normal

$$S_{Sunny} = [2+, 3-] \quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{High} \leftarrow [0+, 3-] \quad Entropy(S_{High}) = 0.0$$

$$S_{Normal} \leftarrow [2+, 0-] \quad Entropy(S_{Normal}) = 0.0$$

$$Gain(S_{Sunny}, \text{Humidity}) = Entropy(S) - \sum_{v \in \{\text{High, Normal}\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, \text{Humidity}) = Entropy(S) - \frac{3}{5}Entropy(S_{High}) - \frac{2}{5}Entropy(S_{Normal})$$

$$Gain(S_{Sunny}, \text{Humidity}) = 0.97 - \frac{3}{5}0.0 - \frac{2}{5}0.0 = \underline{\underline{0.97}}$$

Outlook (Sunny days), Attribute: Wind

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Wind) = Strong, Weak

$$S_{Sunny} = [2+, 3-] \quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Strong} \leftarrow [1+, 1-] \quad Entropy(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [1+, 2-] \quad Entropy(S_{Weak}) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$$

$$Gain(S_{Sunny}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

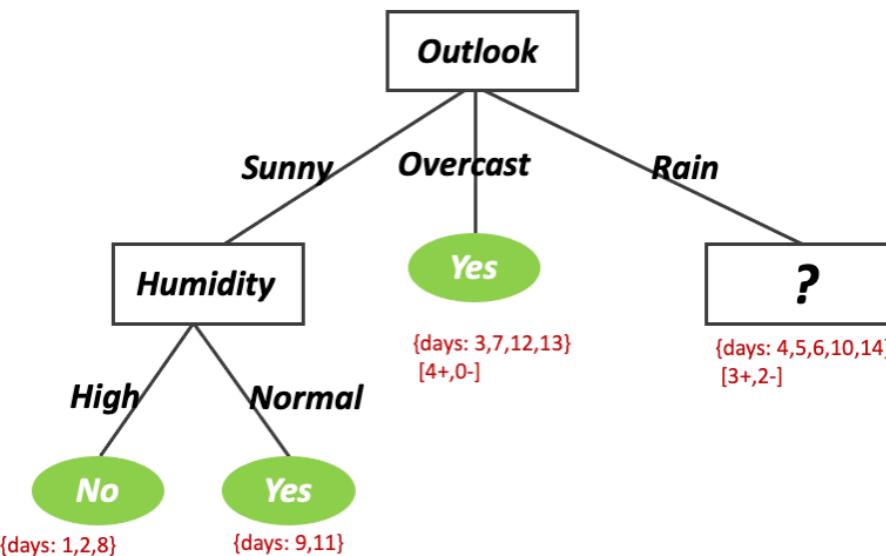
$$Gain(S_{Sunny}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

$$Gain(S_{Sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.9183 = 0.0192$$

Outlook (Sunny days): Gain of all the attributes

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

{days: 1,2,...14}
[9+,5-]



$$Gain(S_{sunny}, Temp) = 0.570$$

$$Gain(S_{sunny}, Humidity) = 0.97$$

Will be the node at this particular label

$$Gain(S_{sunny}, Wind) = 0.0192$$

Outlook (Rainy days):

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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



Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Outlook (Rainy days), Attribute: Temp

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Values (Temp) = Hot, Mild, Cool

$$S_{Rain} = [3+, 2-] \quad Entropy(S) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

$$S_{Hot} \leftarrow [0+, 0-] \quad Entropy(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [2+, 1-] \quad Entropy(S_{Mild}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$$

$$S_{Cool} \leftarrow [1+, 1-] \quad Entropy(S_{Cool}) = 1.0$$

$$Gain(S_{Rain}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp) = Entropy(S) - \frac{0}{5}Entropy(S_{Hot})$$

$$-\frac{3}{5}Entropy(S_{Mild}) - \frac{2}{5}Entropy(S_{Cool})$$

$$Gain(S_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}0.918 - \frac{2}{5}1.0 = \underline{\underline{0.0192}}$$

Outlook (Rainy days), Attribute: Humidity

Day	Outlook	Temp	Humidity	Wind	Play
					Tennis
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

Values (Humidity) = High, Normal

$$S_{Rain} = [3+, 2-] \quad Entropy(S_{Rain}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$S_{High} \leftarrow [1+, 1-] \quad Entropy(S_{High}) = 1.0$$

$$S_{Normal} \leftarrow [2+, 1-] \quad Entropy(S_{Normal}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183$$

$$Gain(S_{Rain}, \text{Humidity}) = Entropy(S) - \sum_{v \in \{\text{High, Normal}\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, \text{Humidity}) = Entropy(S) - \frac{2}{5} Entropy(S_{High})$$

$$-\frac{3}{5} Entropy(S_{Normal})$$

$$Gain(S_{Rain}, \text{Humidity}) = 0.97 - \frac{2}{5} 1.0 - \frac{3}{5} 0.918 = 0.0192$$

Outlook (Rainy days), Attribute: Wind

Day	Outlook	Temp	Humidity	Wind	Play
				Tennis	
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

Values (Wind) = Strong, Weak

$$S_{Rain} = [3+, 2-] \quad Entropy(S) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971$$

$$S_{Strong} \leftarrow [0+, 2-] \quad Entropy(S_{Strong}) = 0.0$$

$$S_{Weak} \leftarrow [3+, 0-] \quad Entropy(S_{Weak}) = 0.0$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5} Entropy(S_{Strong}) \\ - \frac{3}{5} Entropy(S_{Weak})$$

$$Gain(S_{Rain}, Wind) = 0.97 - \frac{2}{5} 0.0 - \frac{3}{5} 0.0 = \underline{\underline{0.97}}$$

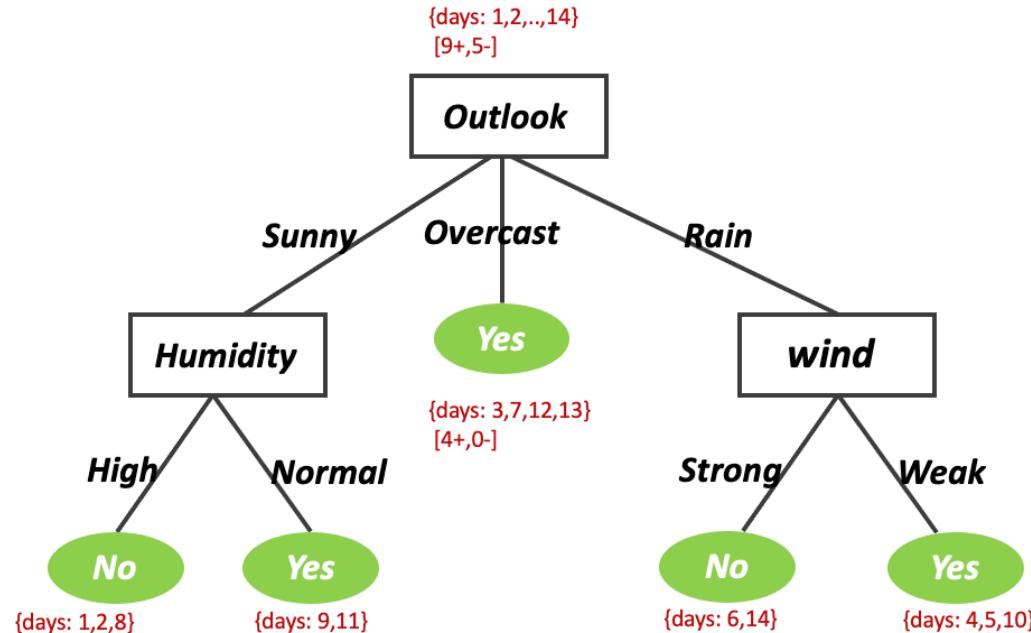
Outlook (Rainy days), Gain of all the attributes

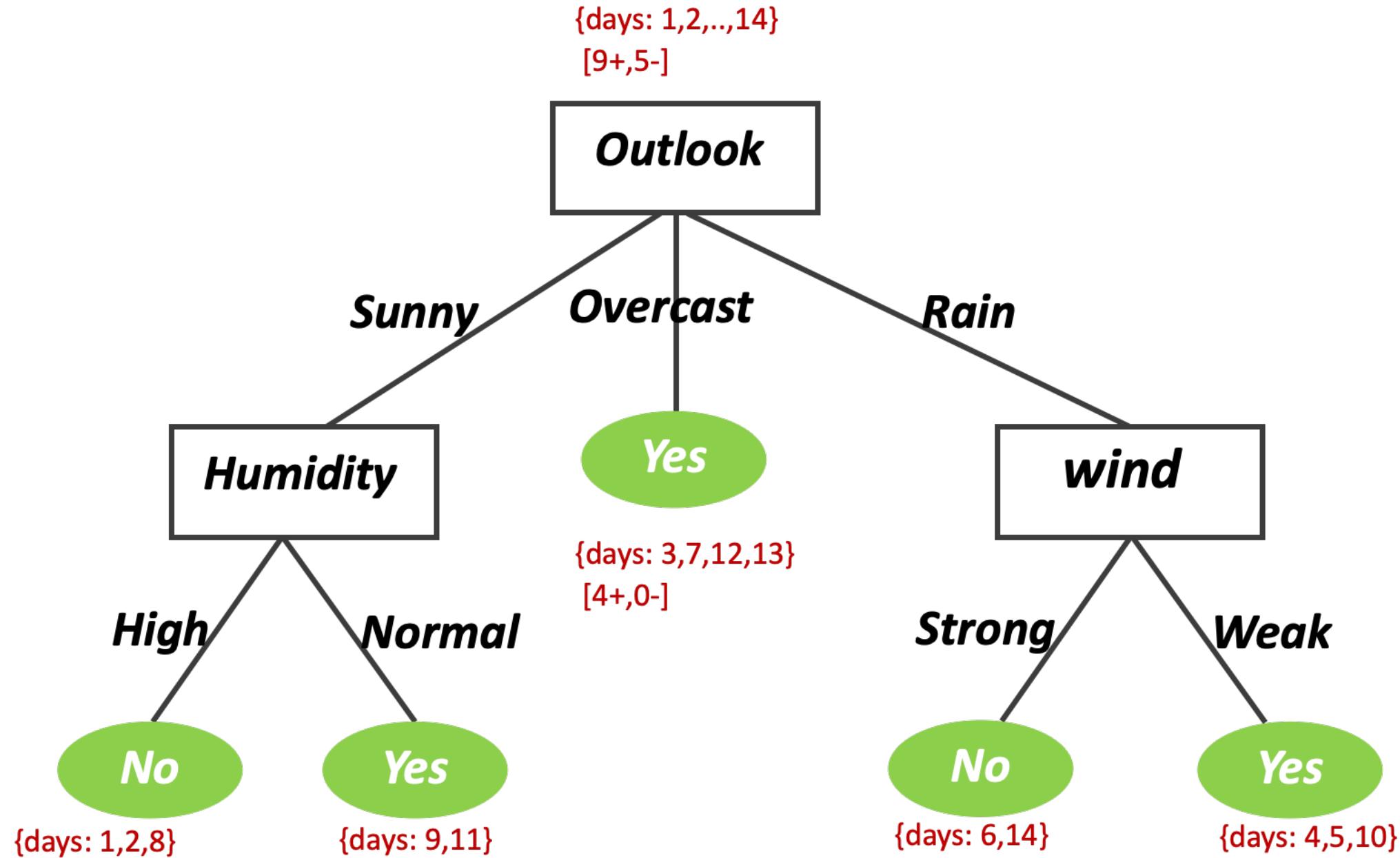
Day	Outlook	Temp	Humidity	Wind	Play
4	Rain	Mild	High	Weak	Tennis Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$Gain(S_{Rain}, Wind) = 0.97$ wind will be the node at this label.



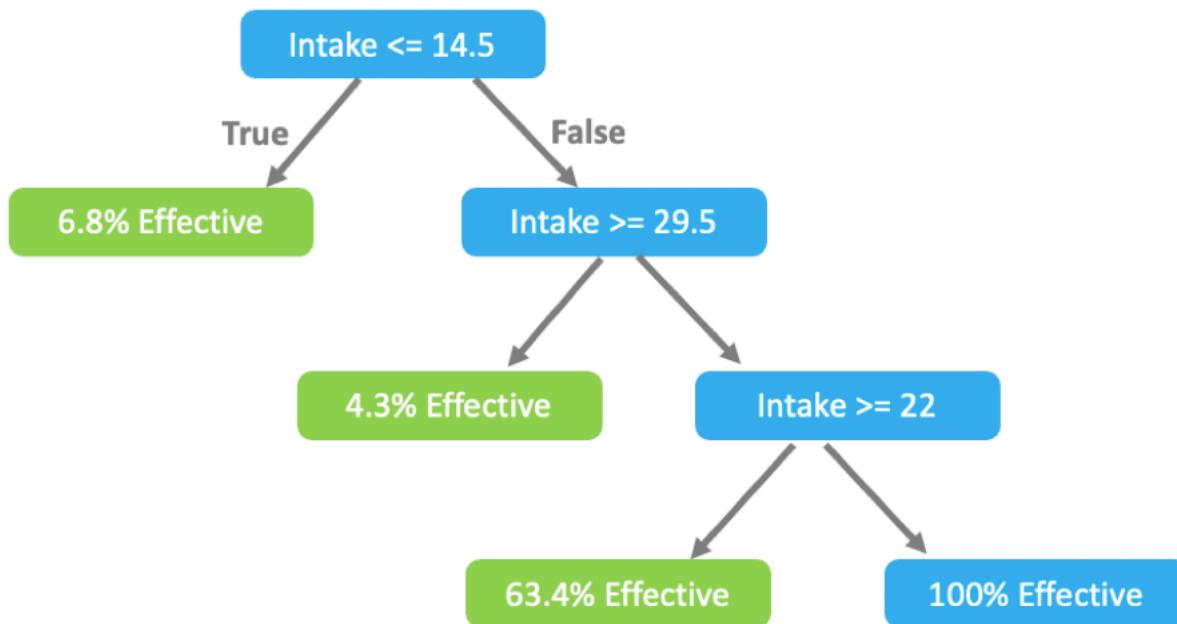


Decision Trees

Regression

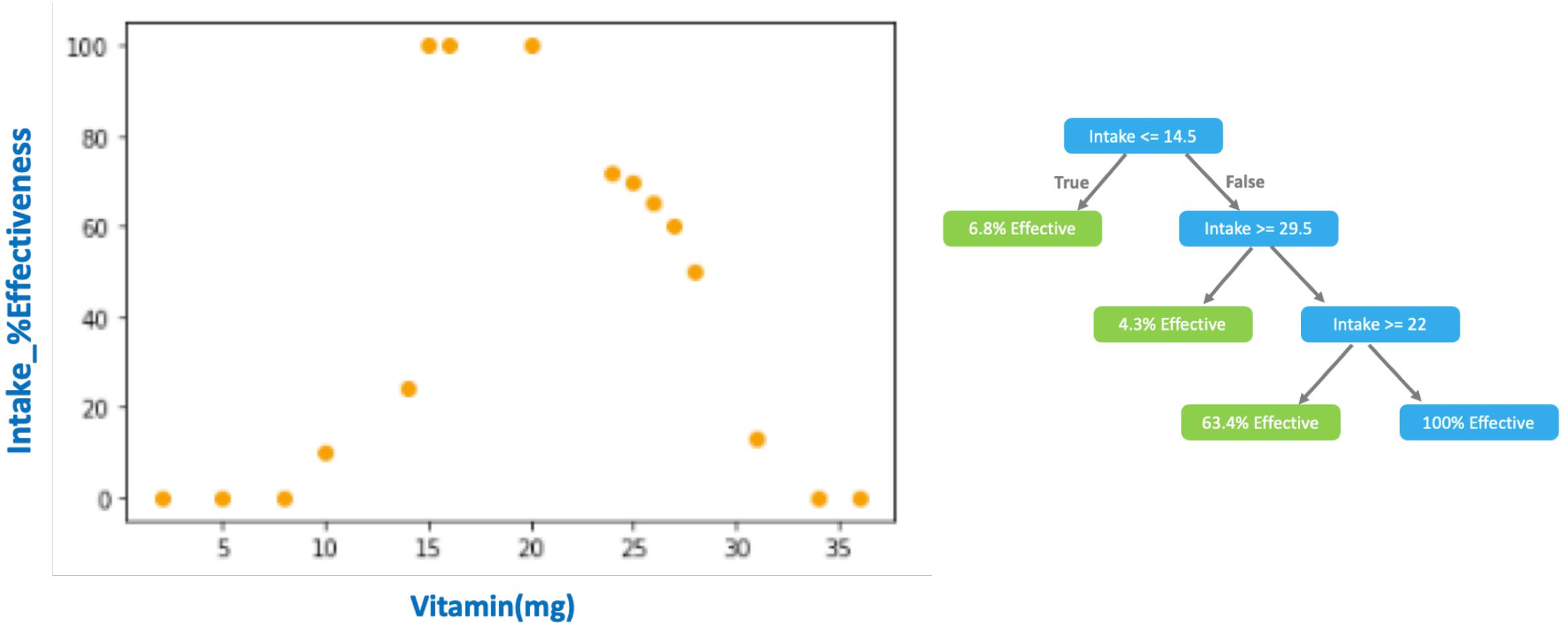
Example:

- Let's build a regression tree from scratch using the dataset given, using one predictor 'VitaminE(mg)'.
- We can build this regression tree:



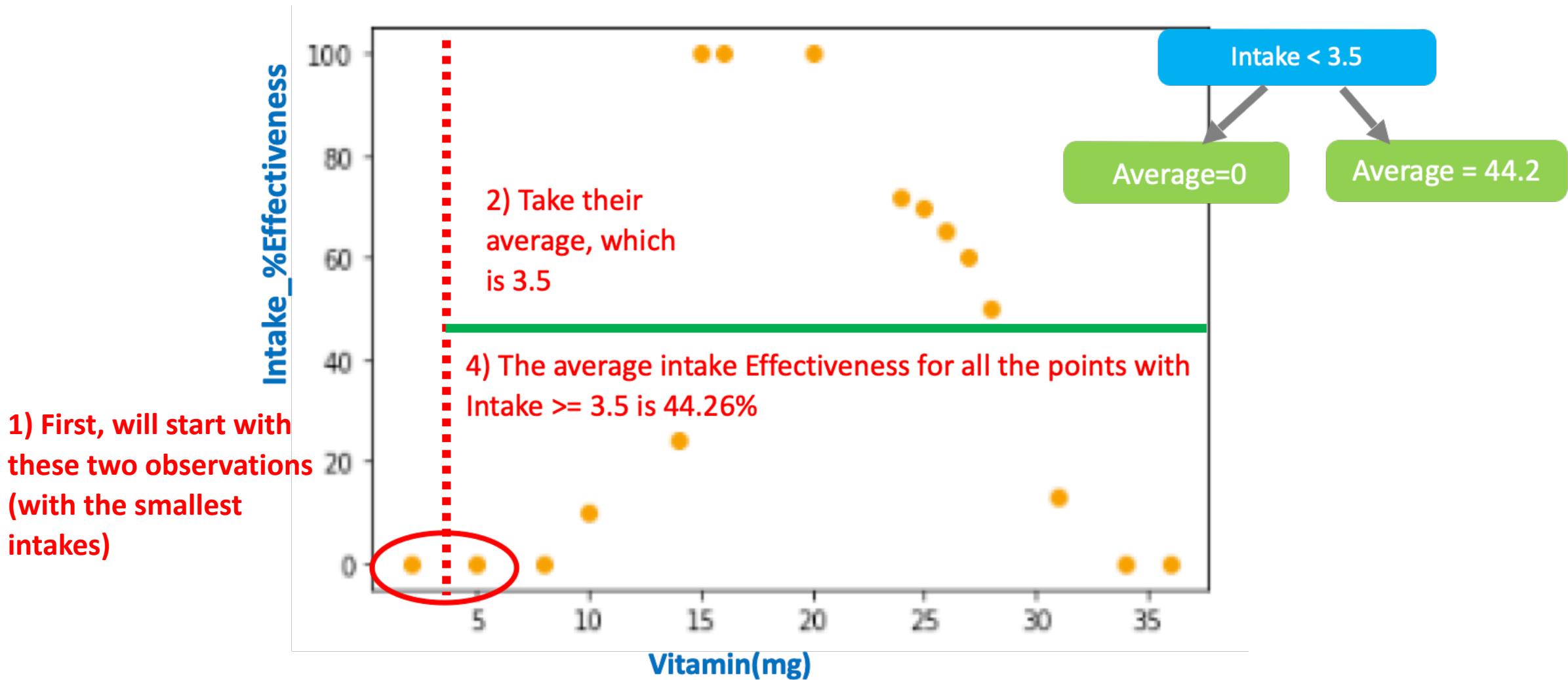
VitaminE(mg)	intake_%Effectiveness
2	0
5	0
8	0
10	10
14	24
15	100
16	100
20	100
24	72
25	70
26	65
27	60
28	50
31	13
34	0
36	0

- Regression trees are built from the top down.
- Why start by asking if intake < 14.5?

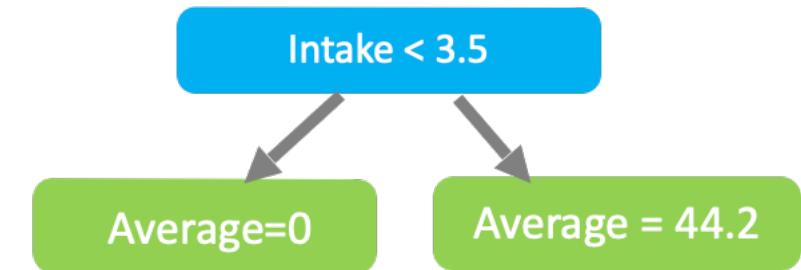
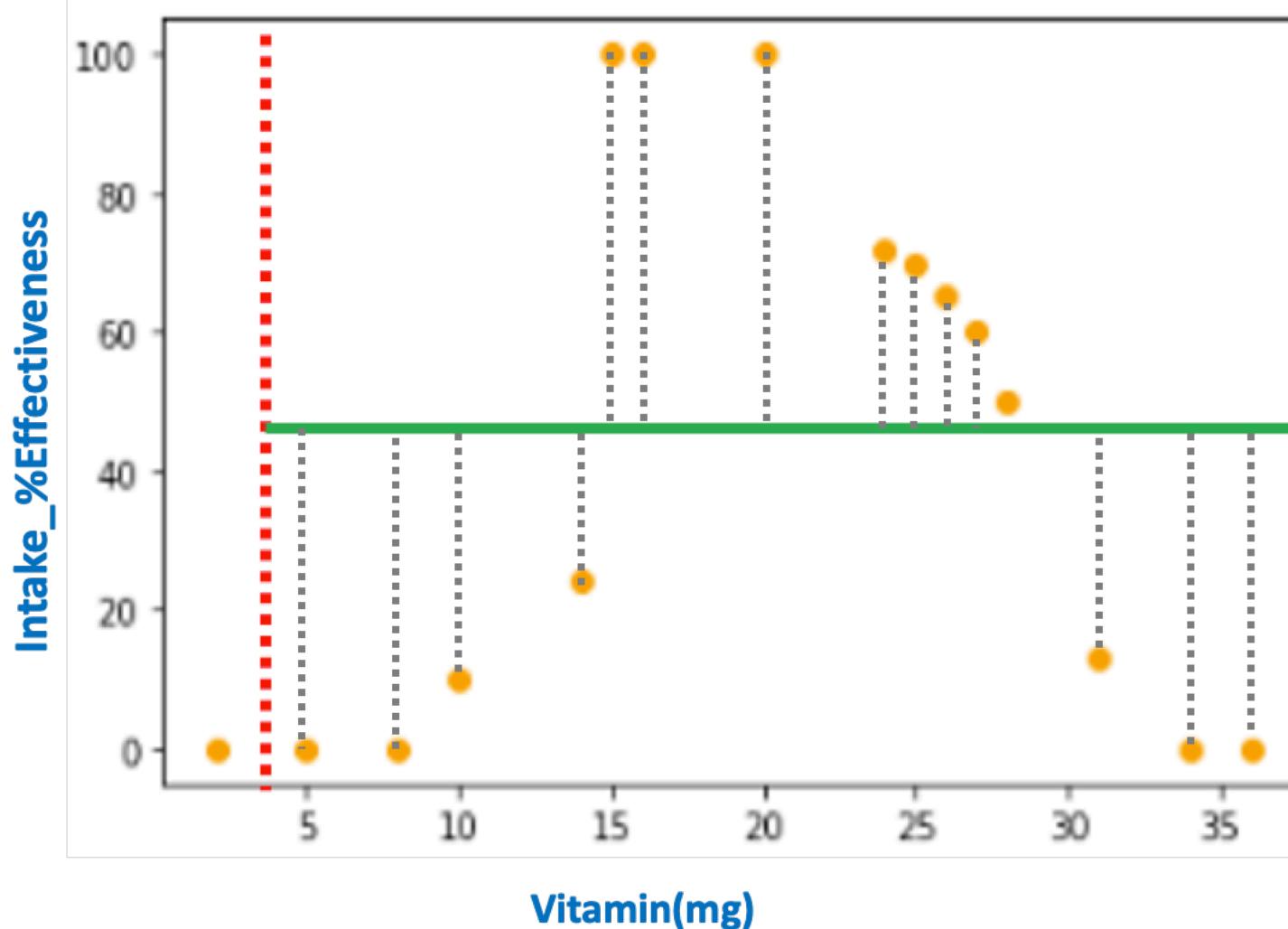


Step 1:

3) Start with a simple tree



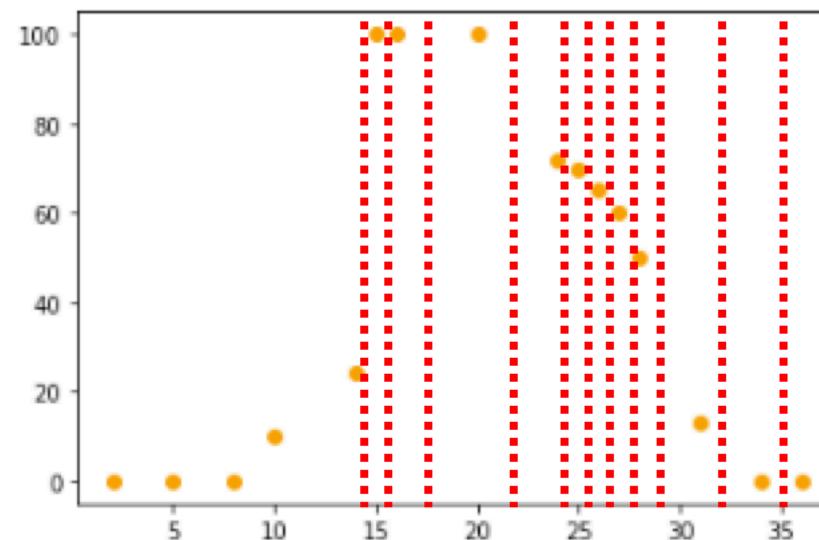
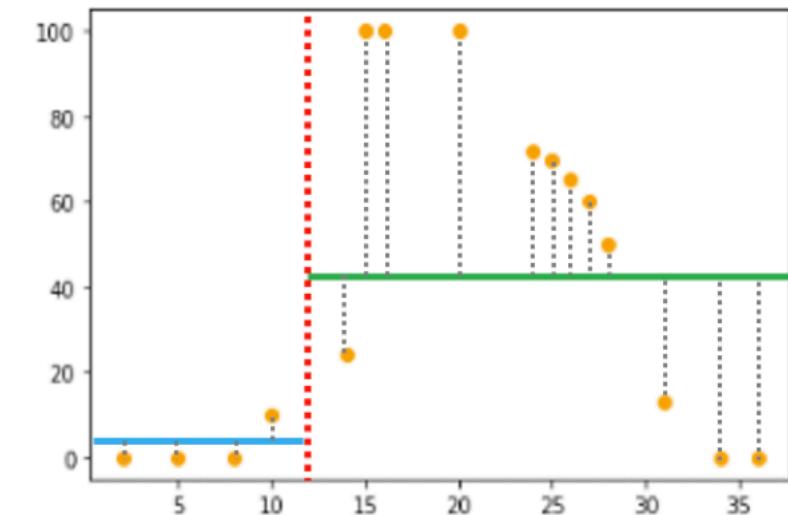
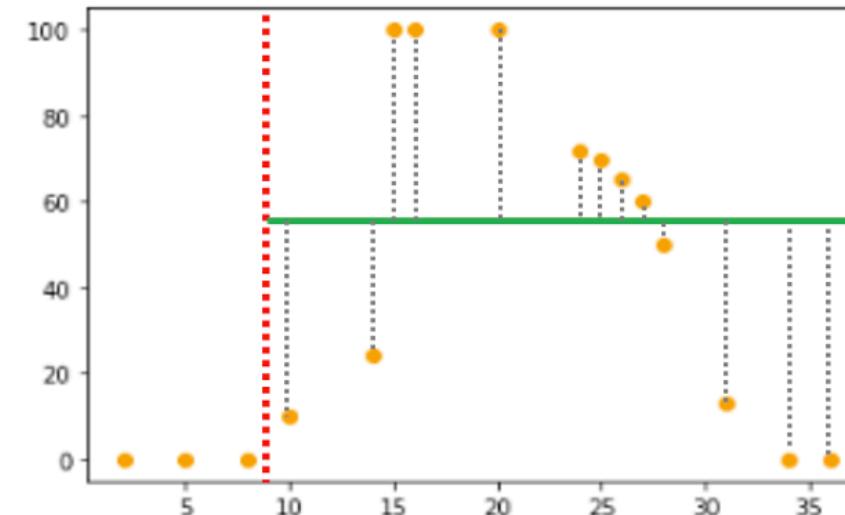
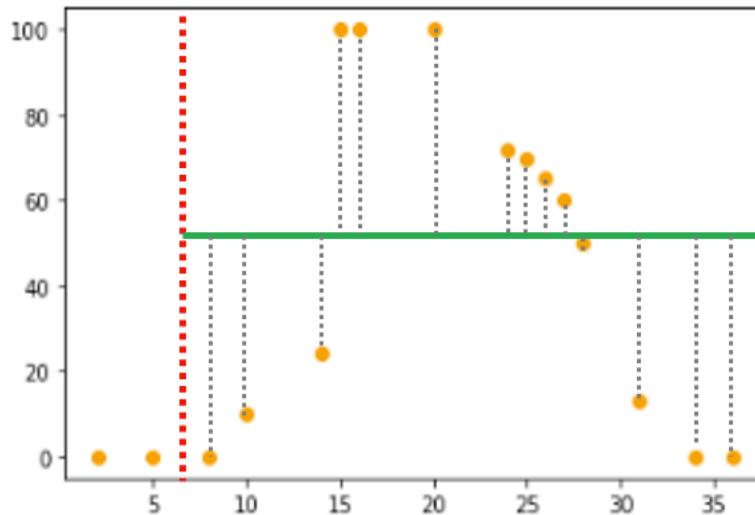
How good is this tree with predictions?



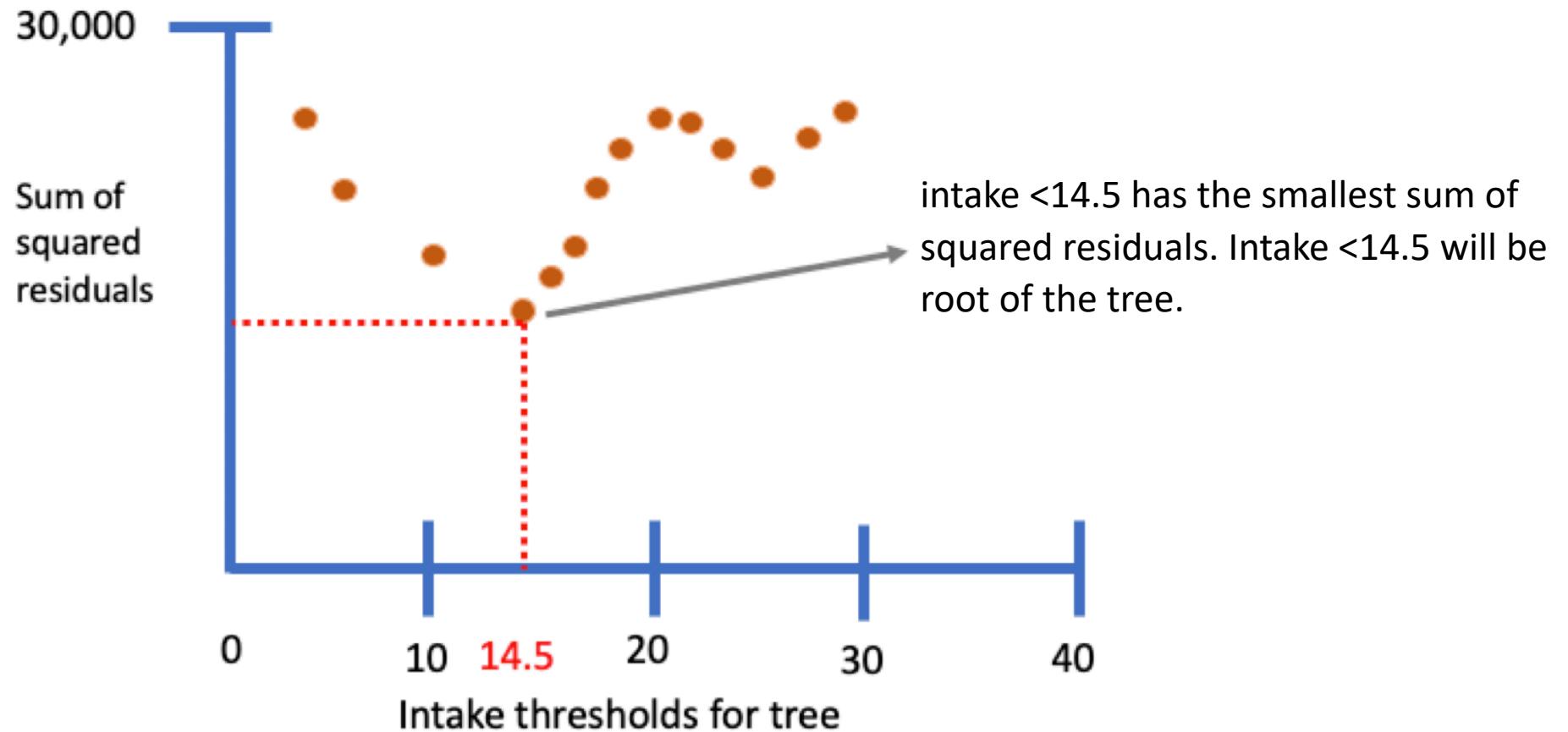
5) Take the sum of squared residuals:

$$\begin{aligned} & (0-0)^2 + (0-44.2)^2 + (0-44.2)^2 + (10-44.2)^2 \\ & + (24-44.2)^2 + (100-44.2)^2 + \dots + (0-44.2)^2 \\ & = 21,861 \end{aligned}$$

Next: Shift the threshold over to the average intake for the next two points. Repeat until the sum of squared residuals for all the remaining thresholds have been calculated.

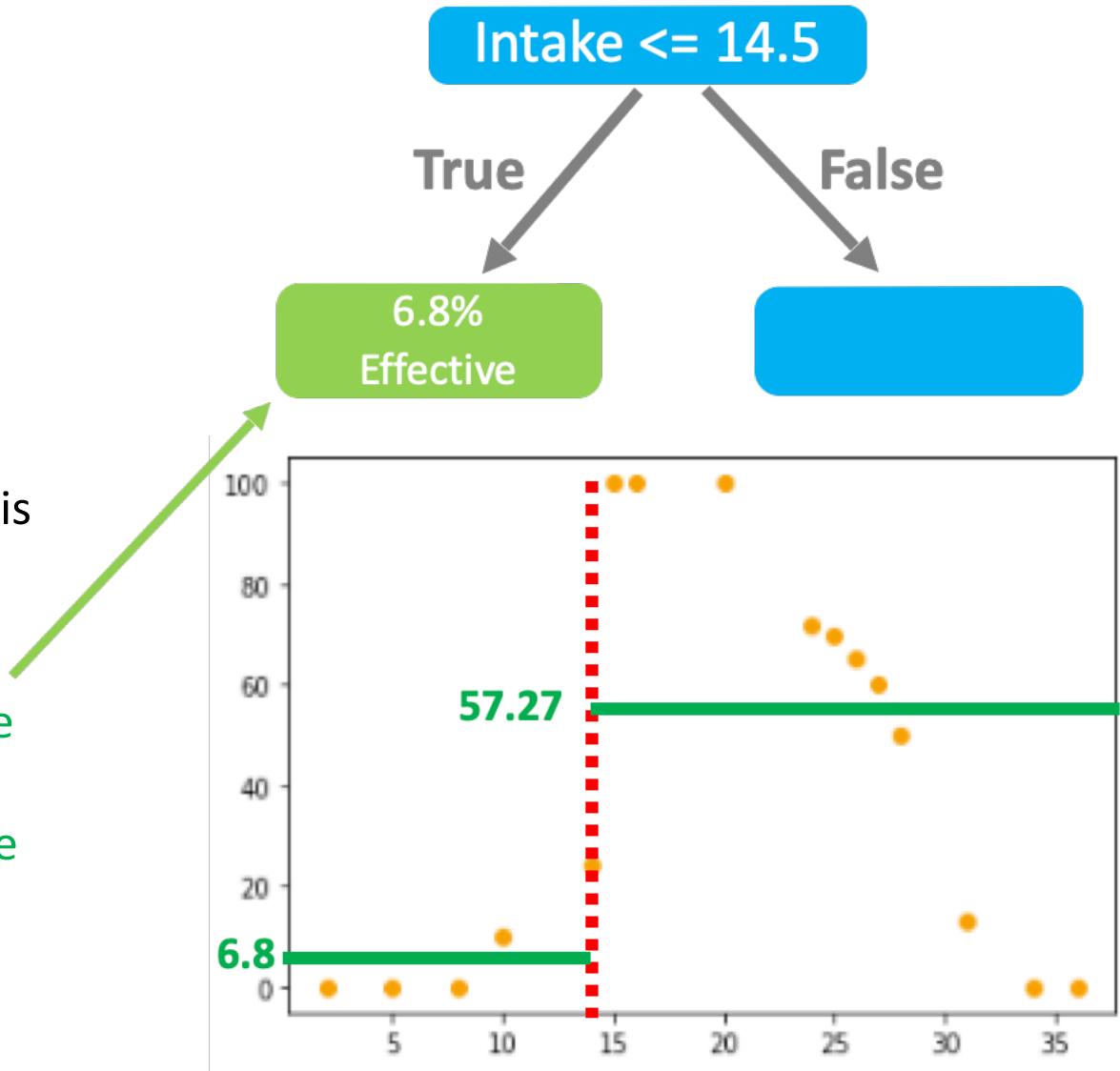


Graph the sum of squared residuals for all the thresholds:



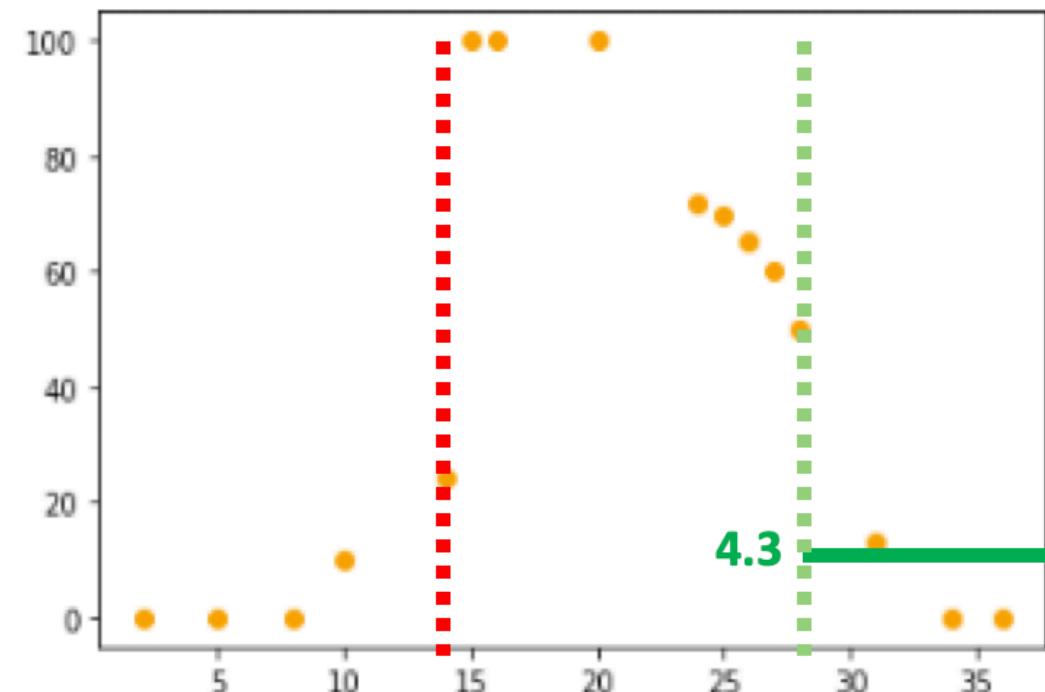
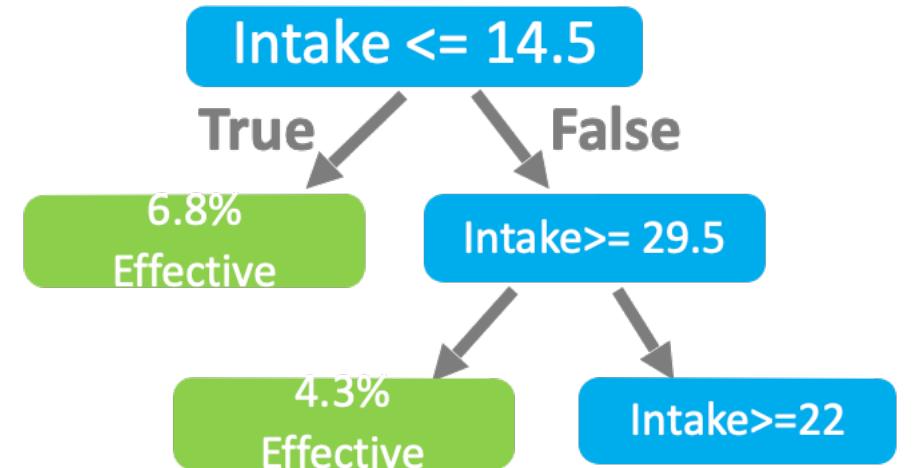
Root of the tree:

- Is there a way to prevent overfitting the training data?
- Yes. One approach is to only split observations when there are more than some minimum number.
- Since there are not many observations in this dataset, set the minimum to 6.
- Since there are only 5 observations with $\text{intake} \leq 14.5$, the observations in this node will not be split (it will become a **leaf**). The output for the leaf will be the average intake effectiveness for the 5 observations with $\text{intake} \leq 14.5$, 6.8%.



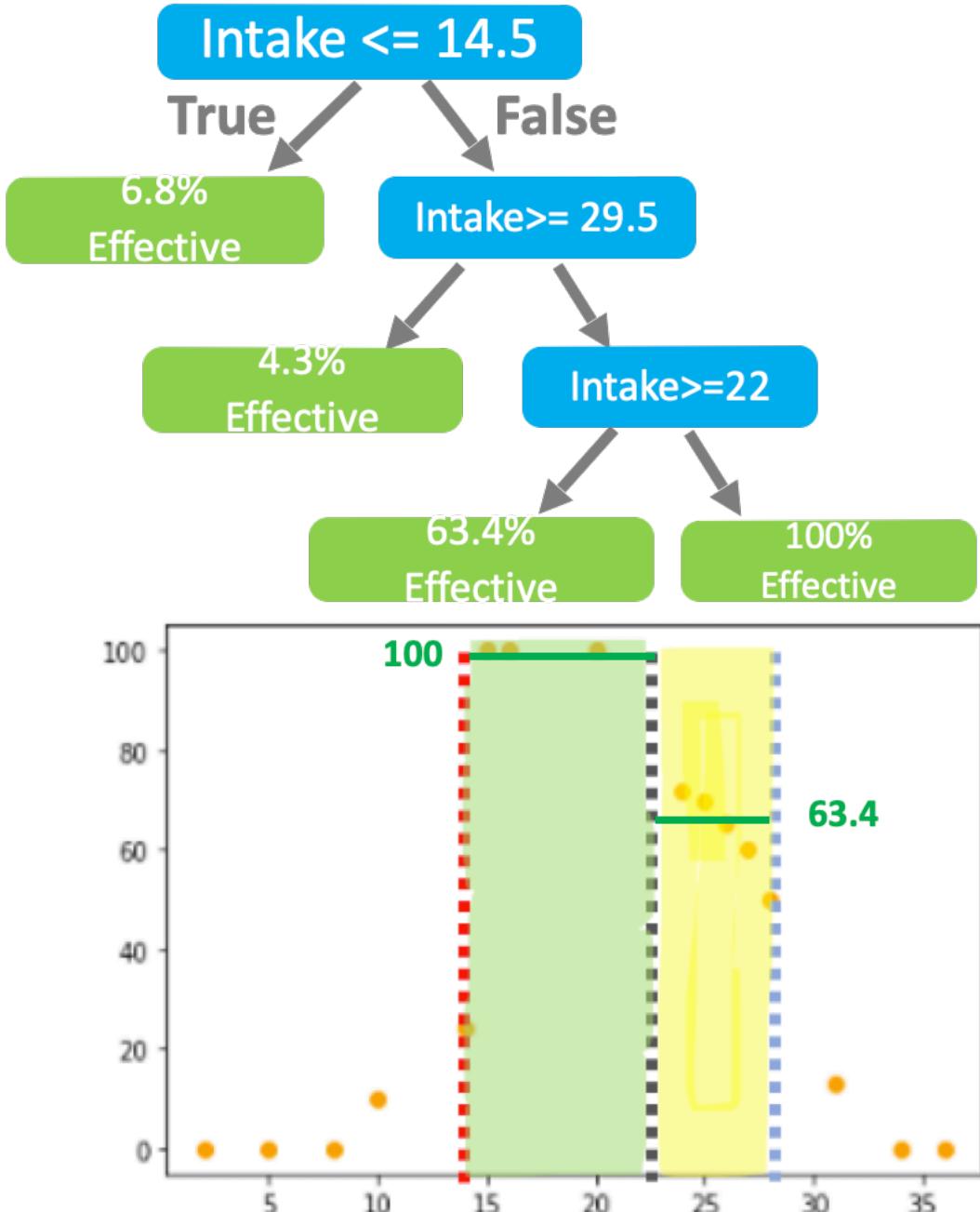
Example Continue:

- We have 11 observations with intake > 14.5 and we can split them into two groups.
- Find the threshold that gives the smallest sum of squared residuals: intake ≥ 29.5
- There are only 3 observations with intake ≥ 29.5 , so that becomes a leaf. Compute the average of the three observations – 4.3% - and that is the output.
- There are 8 observations with intakes between 14.5 and 29. Split them into two groups by finding the threshold that gives the minimum sum of squared residuals.



Example Continue:

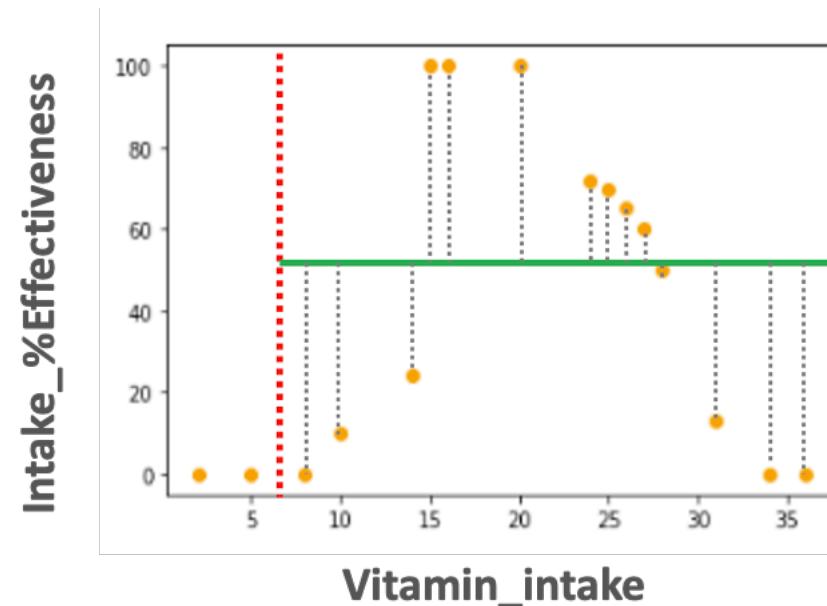
- The threshold between 14.5 and 29 with the minimum sum of squared residuals is 22.
- The average effectiveness for observations with intake between 14.5 and 22 is 100%.
- The average effectiveness for observations with intake between 22 and 29.5 is 63.4%.
- Since no leaf has more than 6 observations in it, the tree is complete. Each leaf corresponds to the average **Intake Effectiveness** from a different cluster of observations.



Build a tree with multiple predictors

- Decision trees can also predict **Intake Effectiveness** with multiple predictors.
- Start by using the first predictor, **vitamin_intake**, to predict **intake_%Effectiveness**.
- Just as before, try different thresholds for intake and calculate the sum of squared residuals at each step.
- The best threshold becomes a candidate for the root.

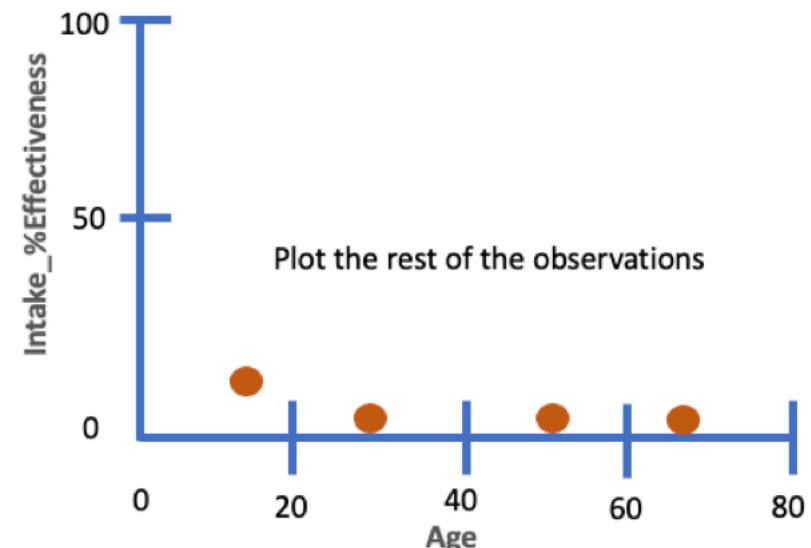
vitamin_intake	Age	Sex	intake_%Effectiveness
2	27	Female	0
5	69	Male	0
8	51	Female	0
10	16	Male	10
etc...			



Build a tree with multiple predictors

- Focus on **age** to predict **intake_%Effectiveness**.
- Try different thresholds for age, calculating the sum of squared residuals at each step.
- Pick the threshold that gives the minimum sum of squared residuals.
- The best threshold becomes another candidate for the root.

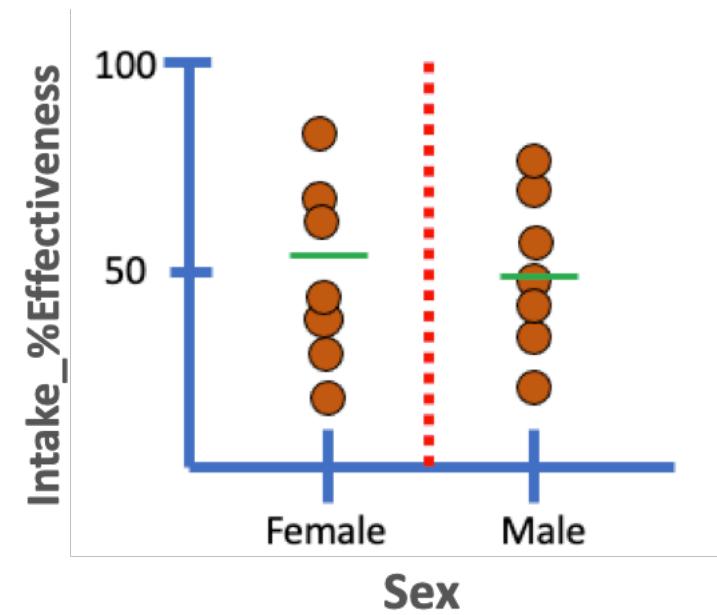
vitamin_intake	Age	Sex	intake_%Effectiveness
2	27	Female	0
5	69	Male	0
8	51	Female	0
10	16	Male	10
etc..			



Build a tree with multiple predictors

- Focus on **Sex** to predict **intake_%Effectiveness**.
- With **Sex** there is only one threshold to try. Use that threshold to calculate the sum of squared residuals
- That threshold becomes another candidate for the root.

vitamin_intake	Age	Sex	intake_%Effectiveness
2	27	Female	0
5	69	Male	0
8	51	Female	0
10	16	Male	10
etc..			



Last step to build a tree with multiple predictors

- Compare the sum of squared residuals for each candidate and pick the candidate with the lowest value
- Since $\text{Age} > 50$ had the lowest sum of squared residuals, it becomes the root of the tree.
- Grow the tree, comparing the lowest sum of squared residuals from each predictor.
- As before, when a leaf has fewer observations than the minimum value (6 in this case), don't split it further.