Decision-trees for classification

Reflections on AI
AI Ethics

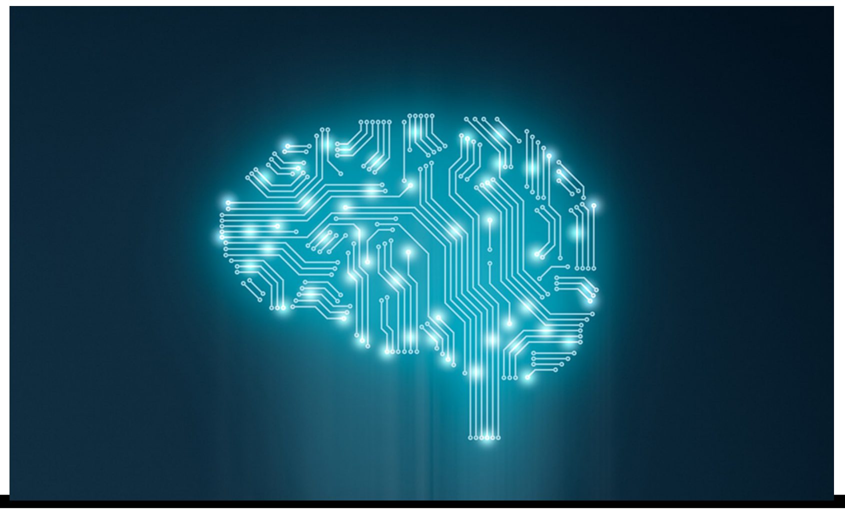
02.12.2024

Outline

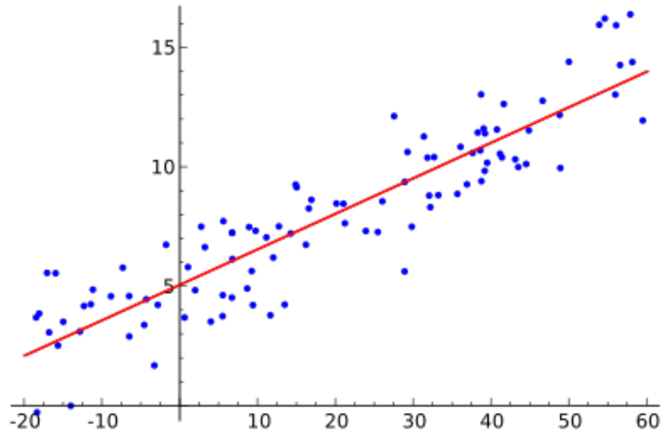
- Hour 1
  - Introduction on AI reflections/ethics
  - Decision-tree continued

- Hour 2: Reflections on AI
  - Societal challenges
  - How could AI help/hurt?

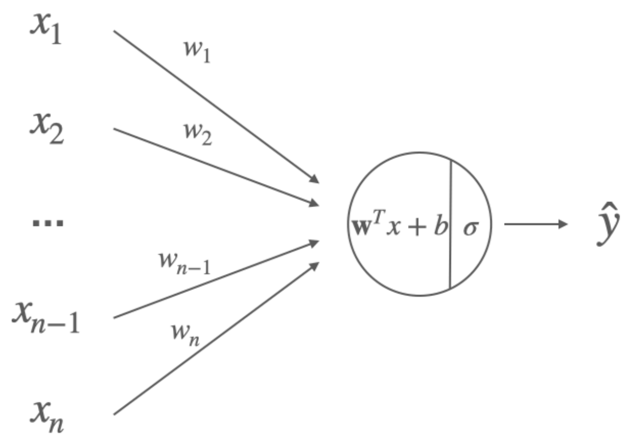
Introduction



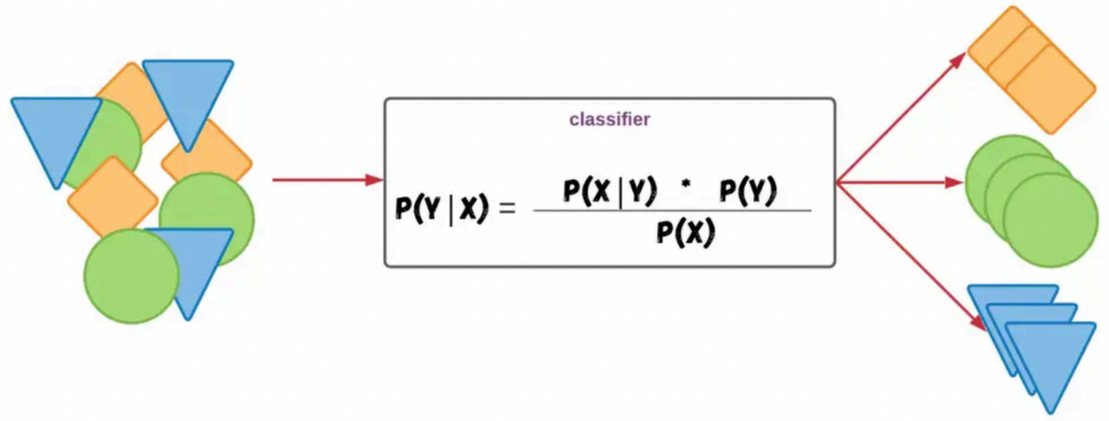
Linear regression



Logistic regression

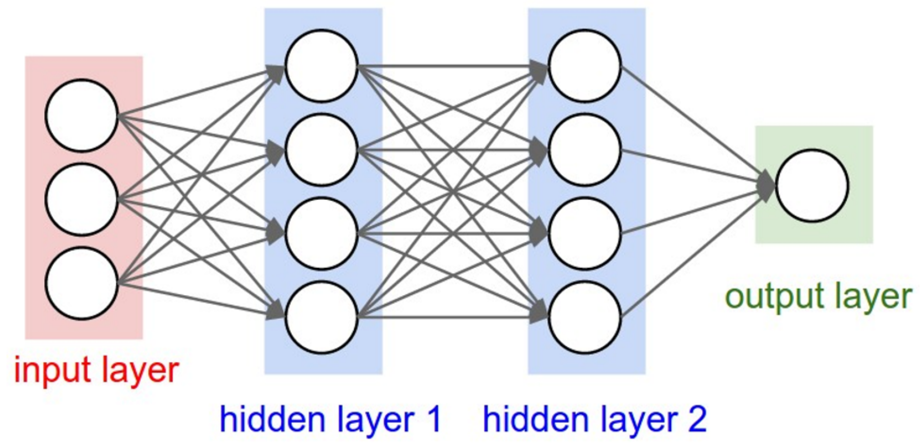


Naive Bayes

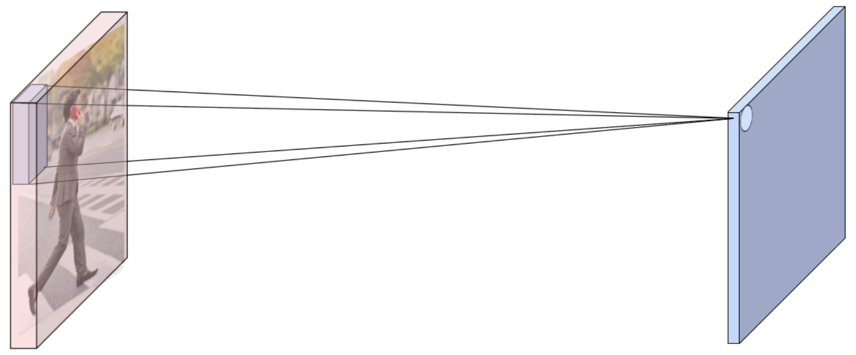


KNN

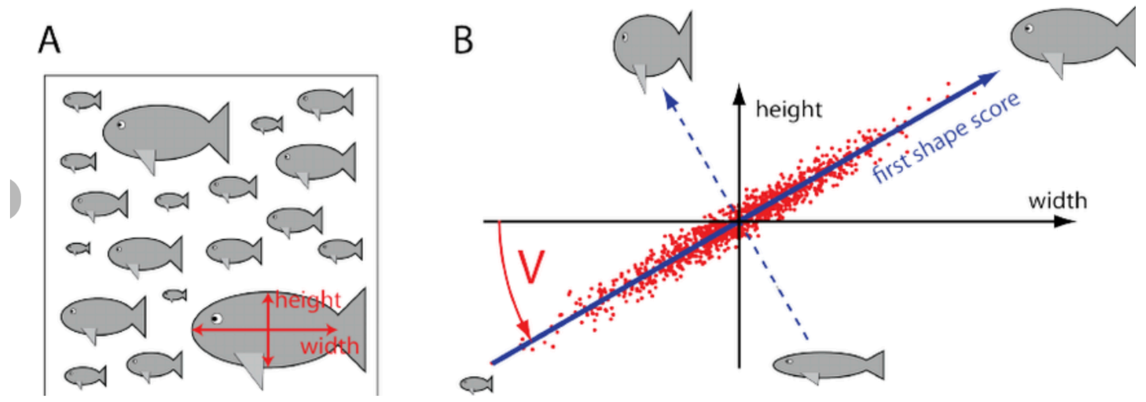
Neural networks



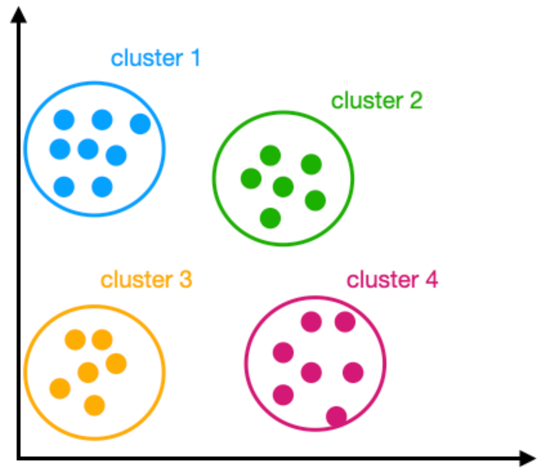
Convolutional neural networks



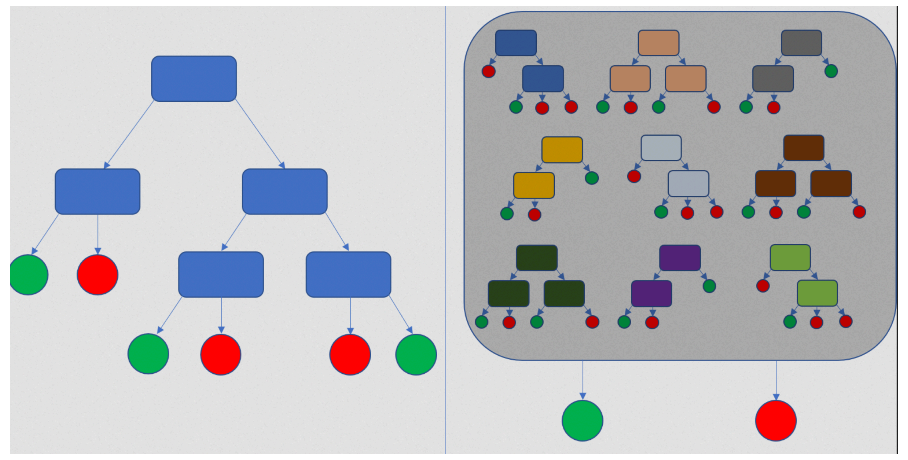
Dimensionality reduction



Clustering

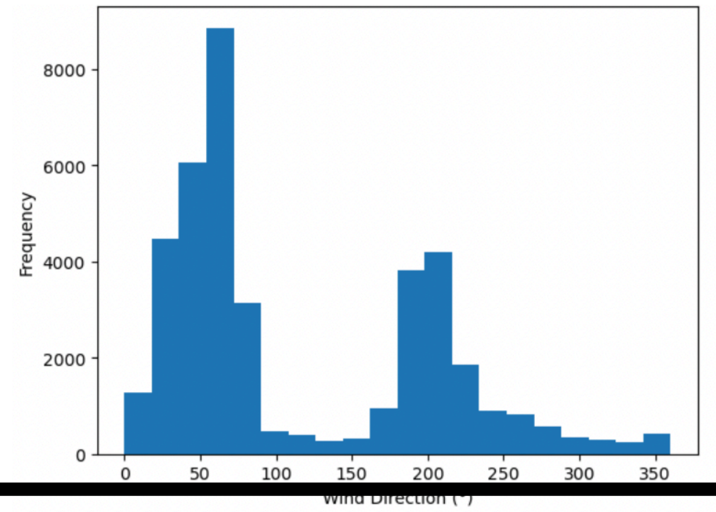


Decision-trees



AI ethics

Reinforcement learning



AI and its societal impacts

As engineers in this era we might use or develop AI-based tools

We need to understand the impact of our choices in designing a tool, using a dataset, and deploying a tool, examples:

- Why should I use a ML-based approach?
- Who will benefit from my product?
- Are there biases in the data, or the methods I have used?
- Can I explain the decisions of the approach I have used?
- What are some harmful consequences?
- What are the environmental impacts of my choice?

https://www.nature.com/articles/d41586-024-01184-4

https://keefeandkeefe.com/

Critical thinking

Even broader than our engineering applications…

- **Who is deciding that the AI narrative should dominate our discussions, courses, approaches?**

- **Who is deciding how I should be spending my moments, what news I should read, how I should interact with my friends, what movies/clips to watch, where to shop, what food to eat, etc?**

- **What are some benefits and potential dangers of using AI in my courses and work?**

- **How can I use/develop AI to empower myself and others, while being aware of its challenges?**

Plan for reflection and ethics in our course

1. **Discussion on conditions for AI to benefit our societies**

   Today in class with Prof. Sascha Nick

   https://forms.gle/qrWcMXppq9kZ4EK6A

2. **Videos to watch at home (link also on our Moodle page)**
   https://mediaspace.epfl.ch/channel/MOOCS+Ethique+num%C3%A9rique/56510

   Prepared by Dr. Johan Rochel

3. **Bonus assignment (2 points)**

   To be posted on 04.12.2024 and to return on Moodle by 20.12.2024

EPFL

Decision Trees for classification

$\{x^i, y^i\}_{i=1}^N$

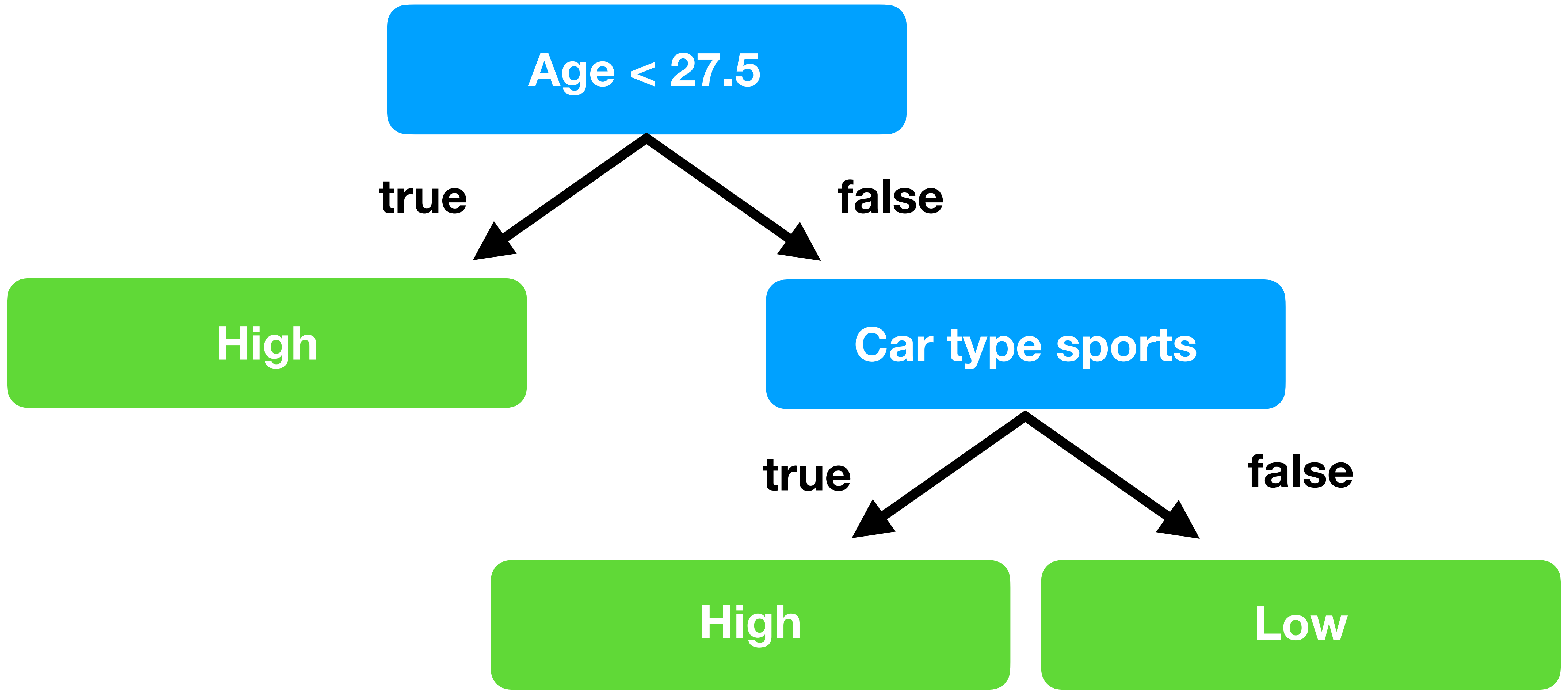
$x^i \in \mathbb{R}^d$, $y^i \in \{0, 1\}$

Which feature to use at each depth to do a split?

For the continuous feature, at what value to do a split?

For the categorical feature, which category to use for the split?

Continuous Feature	Categorical Feature	Class label
Age	**Car type**	**Risk**
23	family	high
17	sports	high
43	sports	high
68	family	low
32	family	low
20	family	high



Classification trees

Greedy approach: choose a feature and the split sequentially based on minimising a performance metric, for example, the **Gini impurity** of a node

Gini impurity of a leaf node: based on empirical probability of class

$$\sum_{l=1}^{K} p_l \sum_{l' \neq l} p_{l'} = \sum_{l=1}^{K} p_l (1 - p_l)$$

ex: $K = 2$ (binary classification): $p_1 p_2 + p_1 p_2 = 2 p_1 p_2$

Gini impurity of a node

Weighted sum of the gini impurity of the two leaf nodes associated with the node

* decide the feature & the feature value for split, based on minimizing gini impurity.

Note: Criteria other than gini index (such as entropy) are also used for node split

Classification Trees

Example - constructing the tree

Gini impurity to construct the classification tree

1. compute Gini impurity for different features and at different split values
2. choose feature and split with lowest Gini impurity

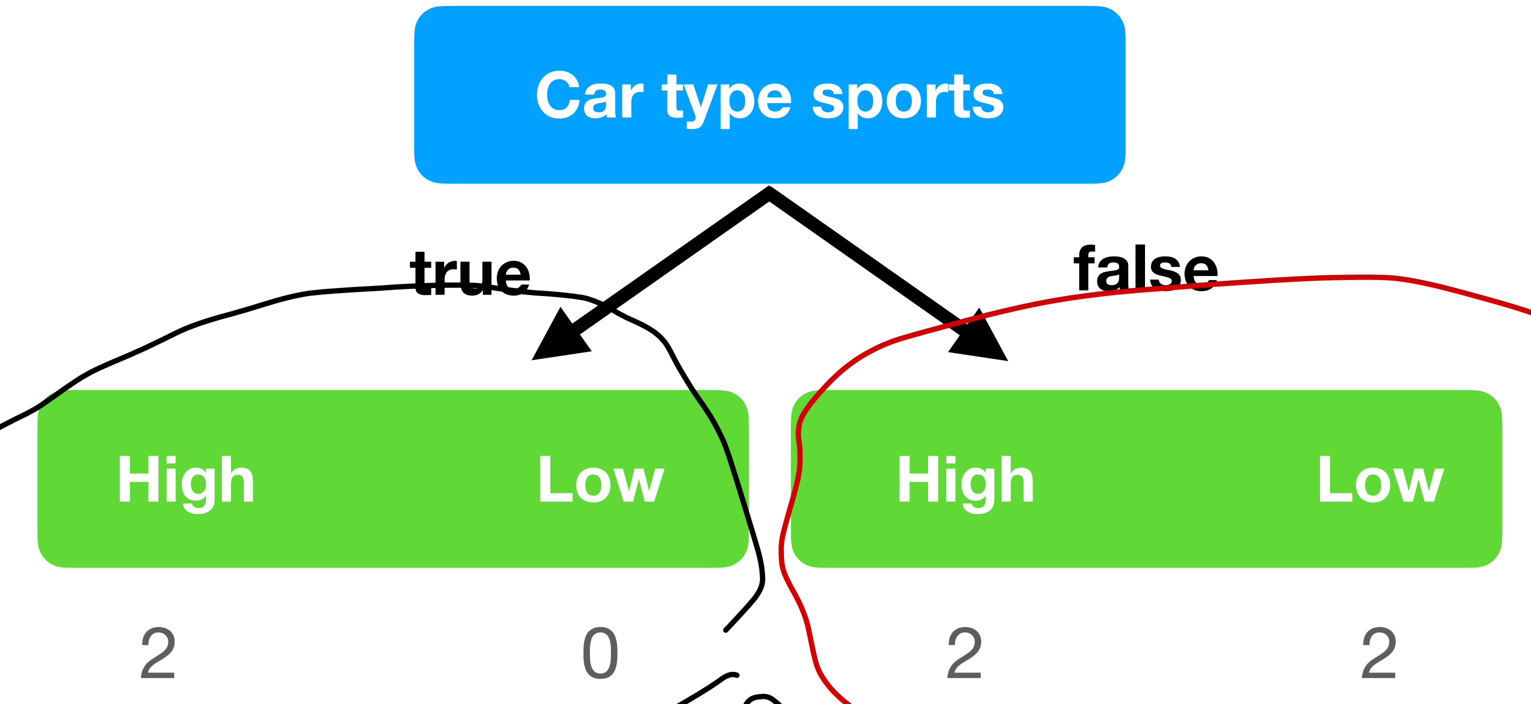
Continuous Feature	Categorical Feature	Class label
Age	**Car type**	**Risk**
23	family	high
17	sports	high
43	sports	high
68	family	low
32	family	low
20	family	high

Classification trees

Criteria for choosing a feature and a split

gini impurity of node :

$$\frac{2}{6} \times 0 + \frac{4}{6} \times \frac{1}{2} = \frac{1}{3}$$



gini impurity for left leaf node

class 1 : high risk

$$\frac{2}{2} \times 0 + 0 \times \frac{2}{2} = 0$$

gini impurity for right leaf node

$$\frac{2}{4} \times \frac{2}{4} + \frac{2}{4} \times \frac{2}{4} = \frac{1}{2}$$

Continuous Feature	Categorical Feature	Class label
Age	**Car type**	**Risk**
23	family	high
17	sports	high
43	sports	high
68	family	low
32	family	low
20	family	high

Decision Trees

Splitting continuous features

tid	Age	Risk
0	23	high
1	17	high
2	43	high
3	68	low
4	32	low
5	20	high

Reorder the data depending on Age

tid	Age	Risk
1	17	high
5	20	high
0	23	high
4	32	low
2	43	high
3	68	low

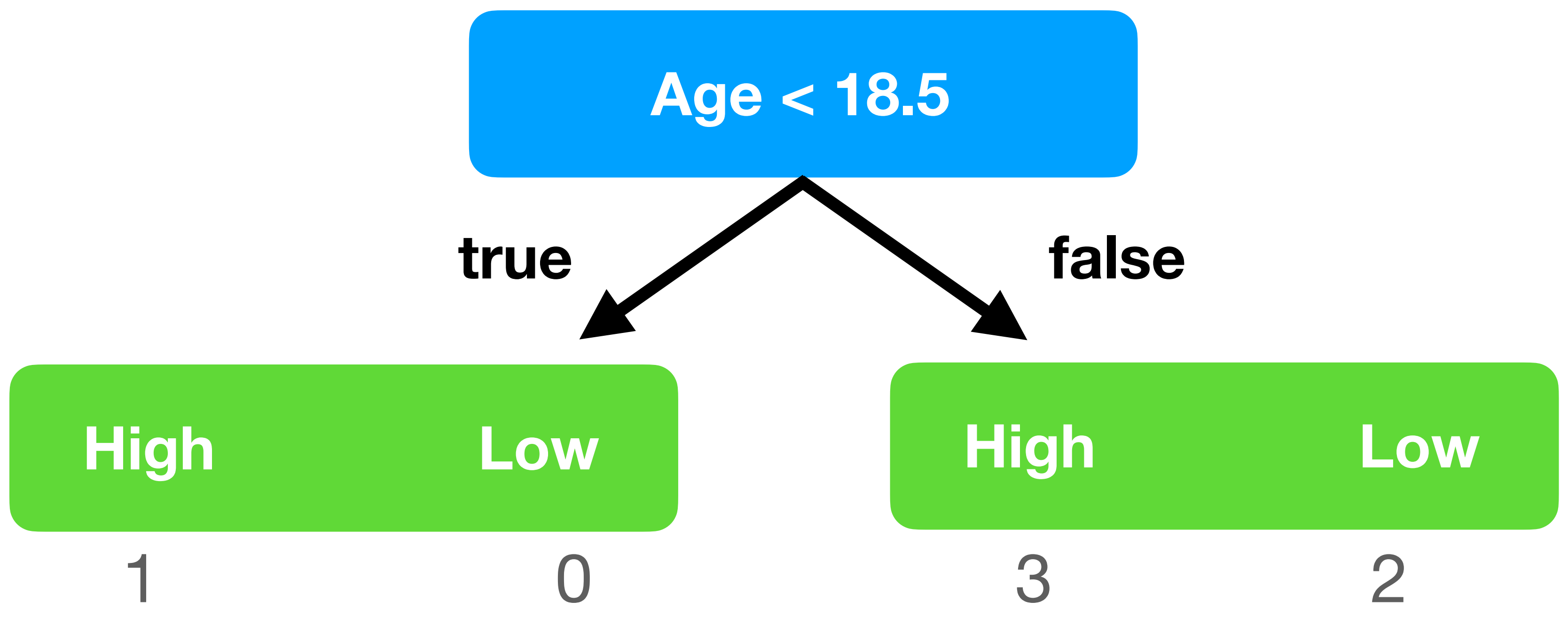
$18\frac{1}{2}$

$21\frac{1}{2}$

$27\frac{1}{2}$

$37\frac{1}{2}$

$55\frac{1}{2}$



gini impurity of left leaf node :

$$2\left(\frac{1}{1} \times 0\right) = 0$$

gini impurity of right leaf node

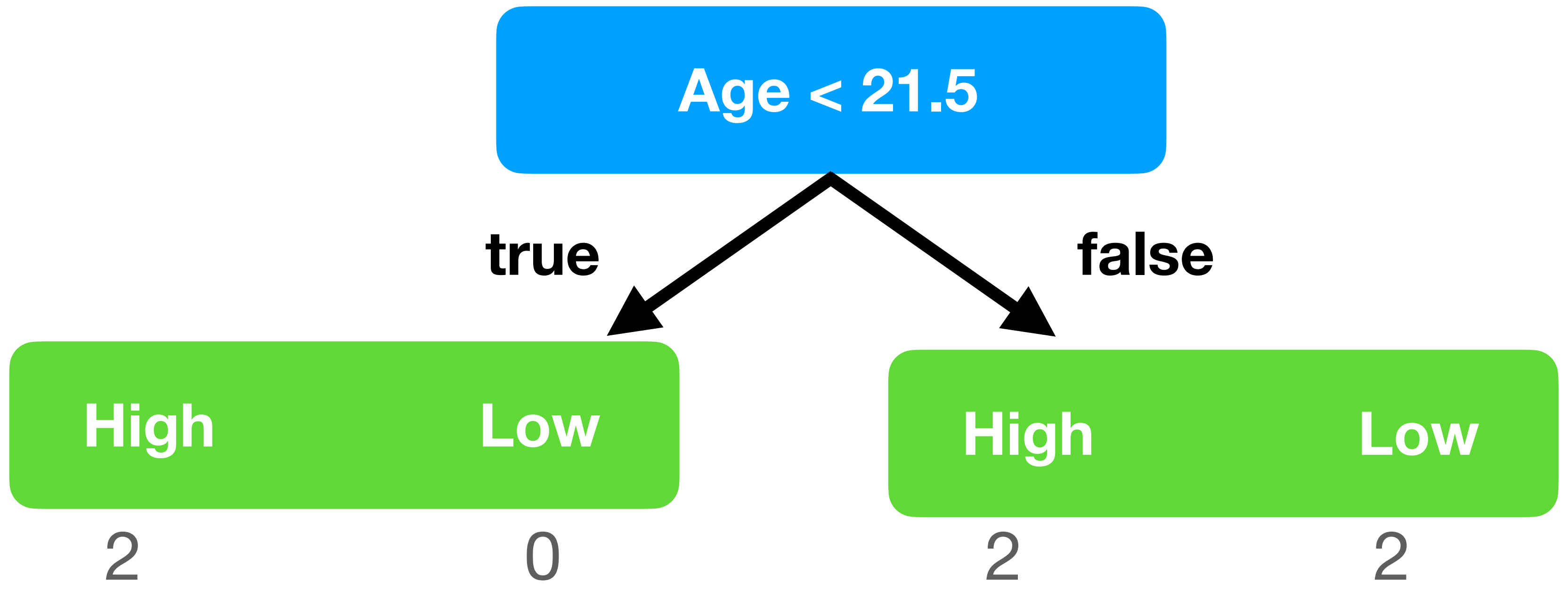
$$\frac{3}{5} \times \frac{2}{5} + \frac{2}{5} \times \frac{3}{5} = \frac{12}{25}$$

gini impurity of node " Age < 18.5 "

$$\frac{1}{6}(0) + \frac{5}{6} \times \frac{12}{25} = \frac{2}{5}$$

Decision Trees

Splitting continuous features



tid	Age	Risk
0	23	high
1	17	high
2	43	high
3	68	low
4	32	low
5	20	high

Reorder the data depending on Age

tid	Age	Risk
1	17	high
5	20	high
0	23	high
4	32	low
2	43	high
3	68	low

gini impurity of "Age < 21.5" $= \frac{1}{3}$

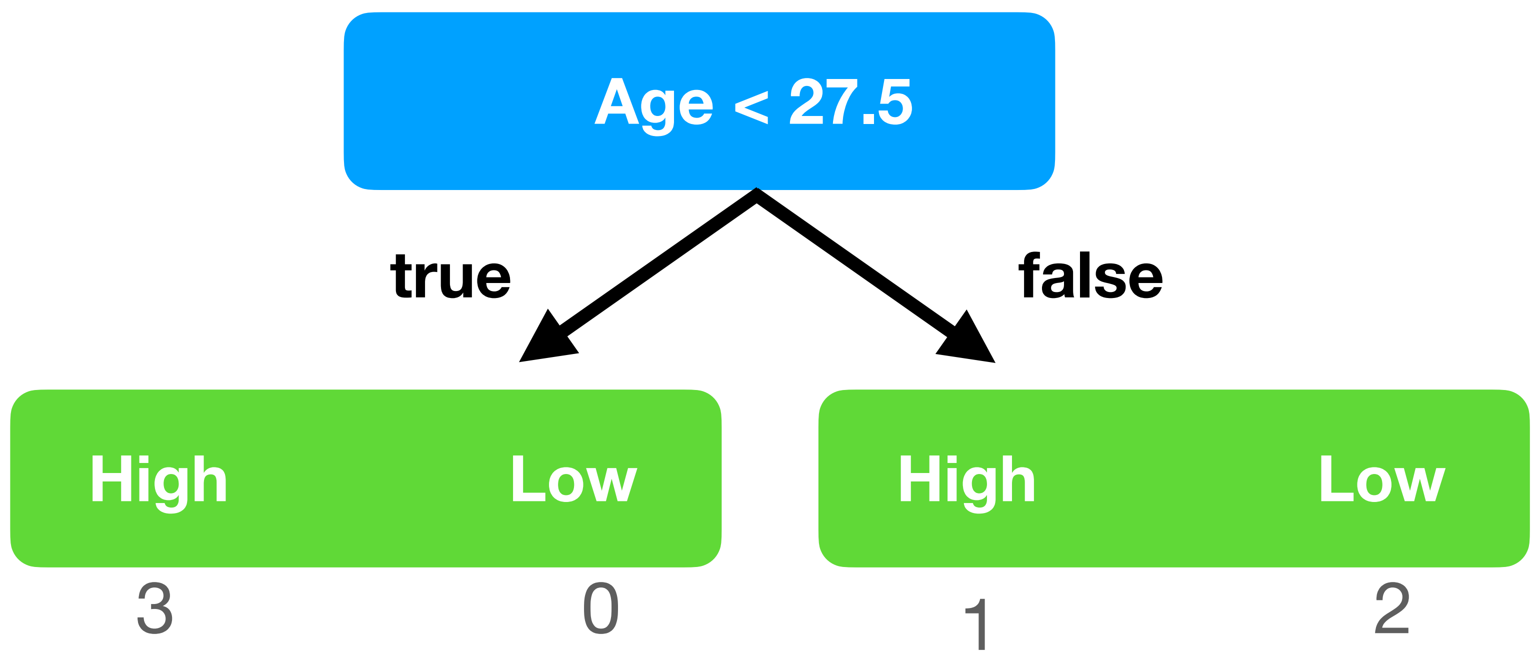
Decision Trees

Splitting continuous features

tid	Age	Risk
0	23	high
1	17	high
2	43	high
3	68	low
4	32	low
5	20	high

Reorder the data depending on Age

tid	Age	Risk
1	17	high
5	20	high
0	23	high
4	32	low
2	43	high
3	68	low



gini impurity of "Age < 27.5"

$2/9$

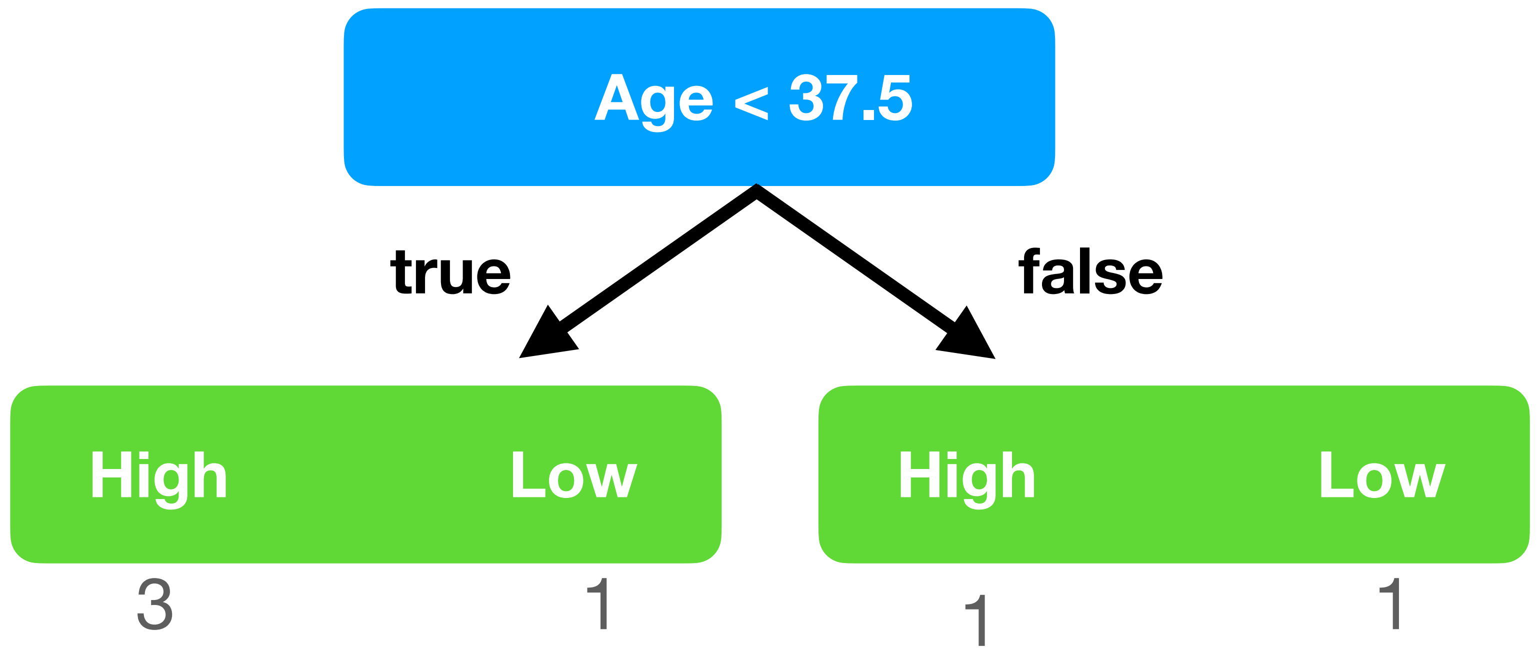
Decision Trees

Splitting continuous features

tid	Age	Risk
0	23	high
1	17	high
2	43	high
3	68	low
4	32	low
5	20	high

Reorder the data depending on Age

tid	Age	Risk
1	17	high
5	20	high
0	23	high
4	32	low
2	43	high
3	68	low



gini impurity of left leaf node

$$\left(\frac{3}{4} \times \frac{1}{4}\right) 2 = \frac{3}{8}$$

gini impurity of right leaf node

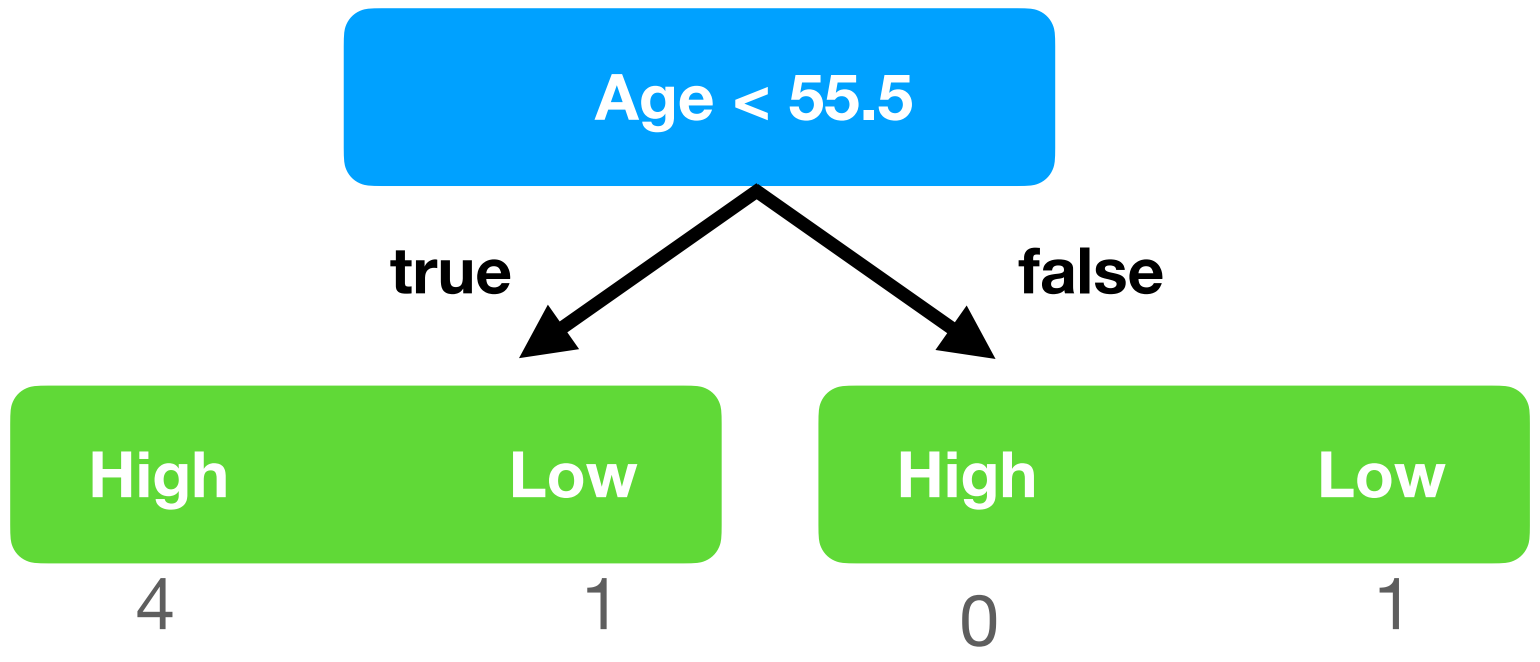
$$\left(\frac{1}{2} \times \frac{1}{2}\right) 2 = \frac{1}{2}$$

gini impurity of node "Age<37.5"

$5/12$

Decision Trees

Splitting continuous features



tid	Age	Risk
0	23	high
1	17	high
2	43	high
3	68	low
4	32	low
5	20	high

Reorder the data depending on Age

tid	Age	Risk
1	17	high
5	20	high
0	23	high
4	32	low
2	43	high
3	68	low

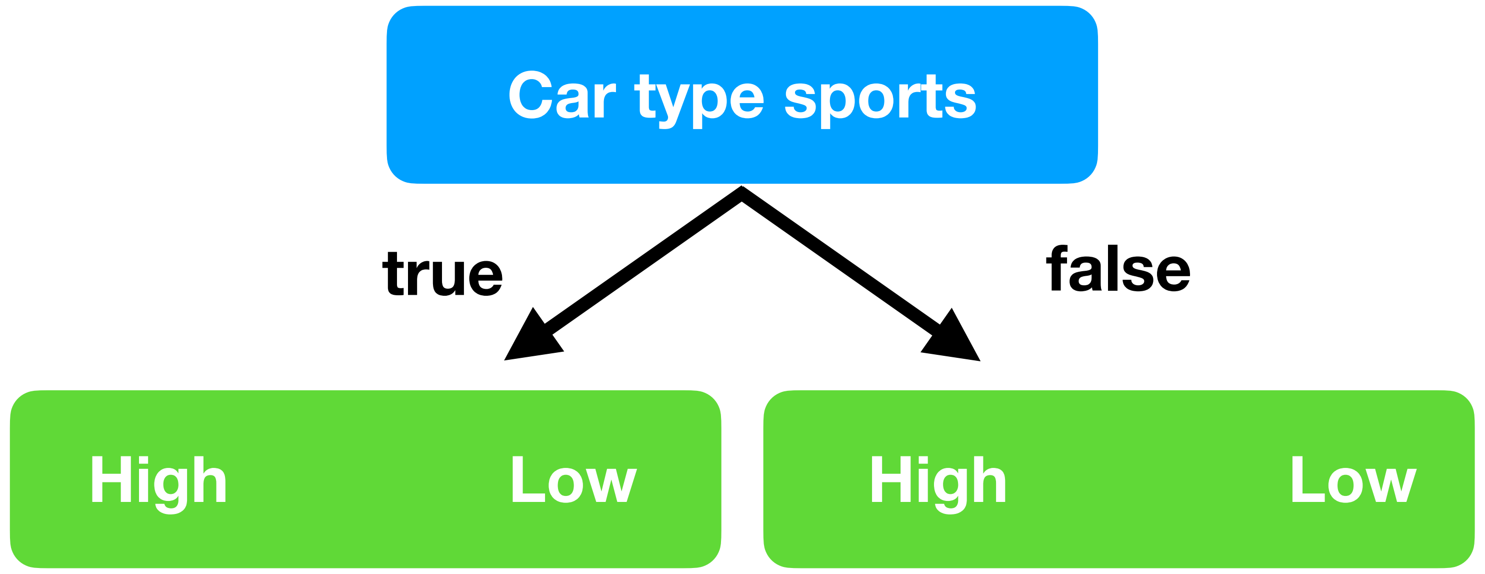
gini impurity of node "Age < 55.5"

$$\frac{4}{5}$$

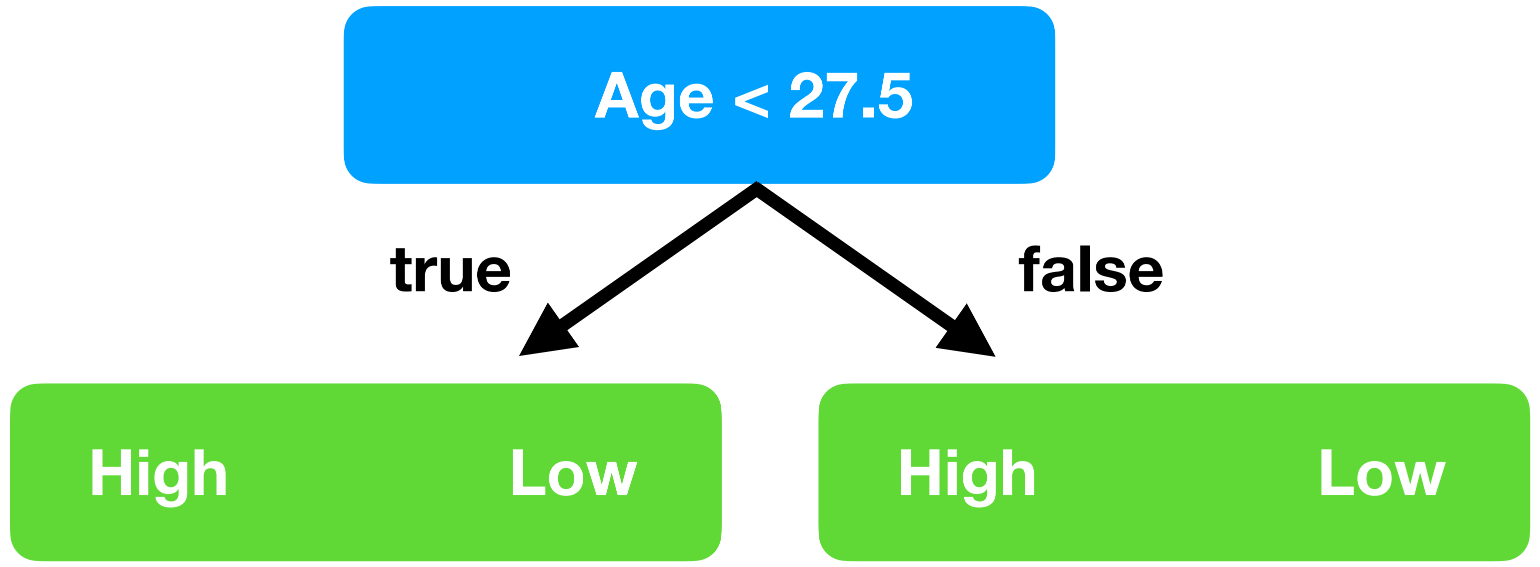
Decision Trees for classification

Age	Car type	Risk
23	family	high
17	sports	high
43	sports	high
68	family	low
32	family	low
20	family	high

Continuous Feature: Age; Categorical Feature: Car type; Class label: Risk



gini impurity: $\frac{1}{3}$



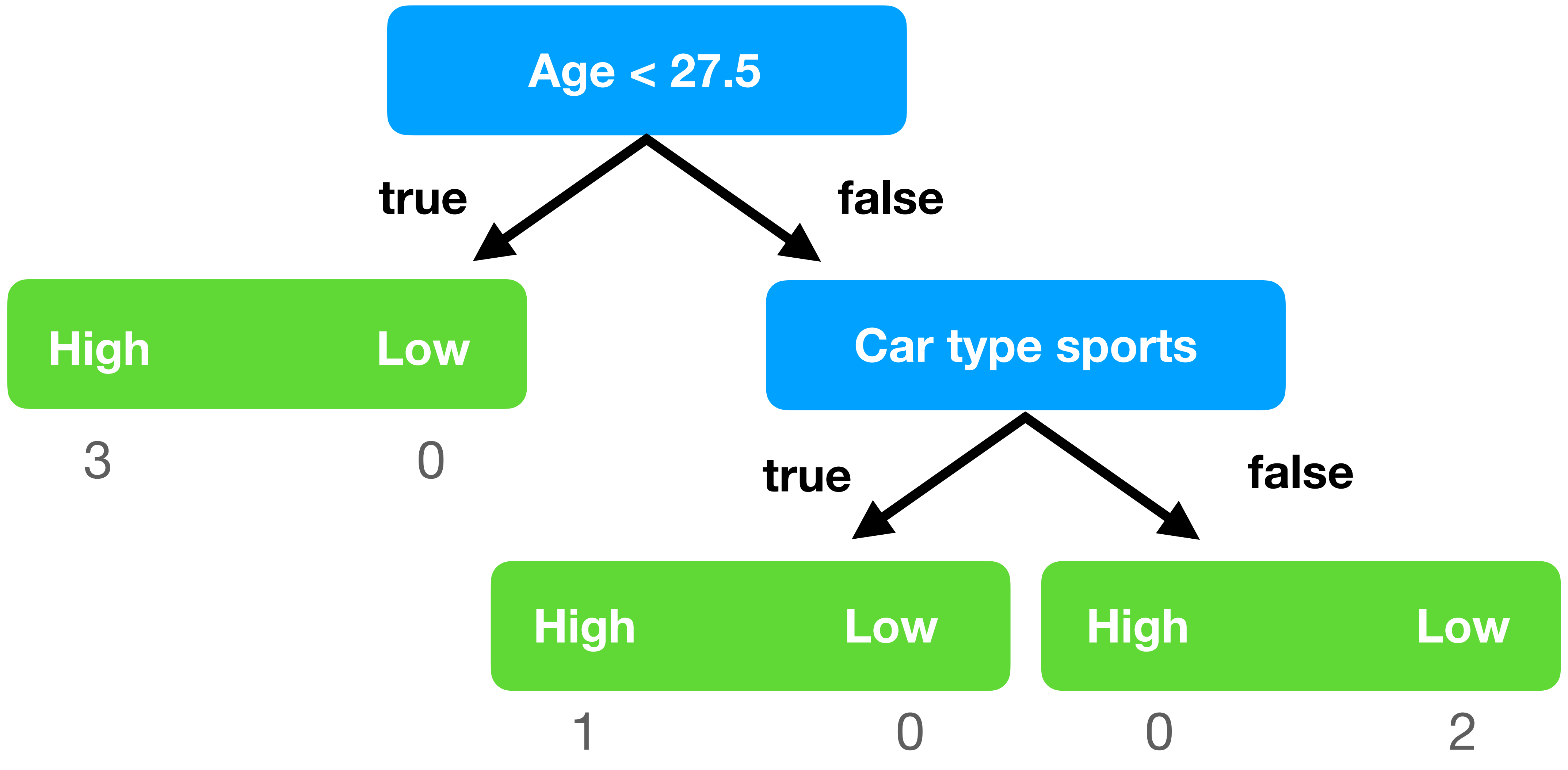
gini impurity: $\frac{2}{9}$

Classification tree

Example - Final tree

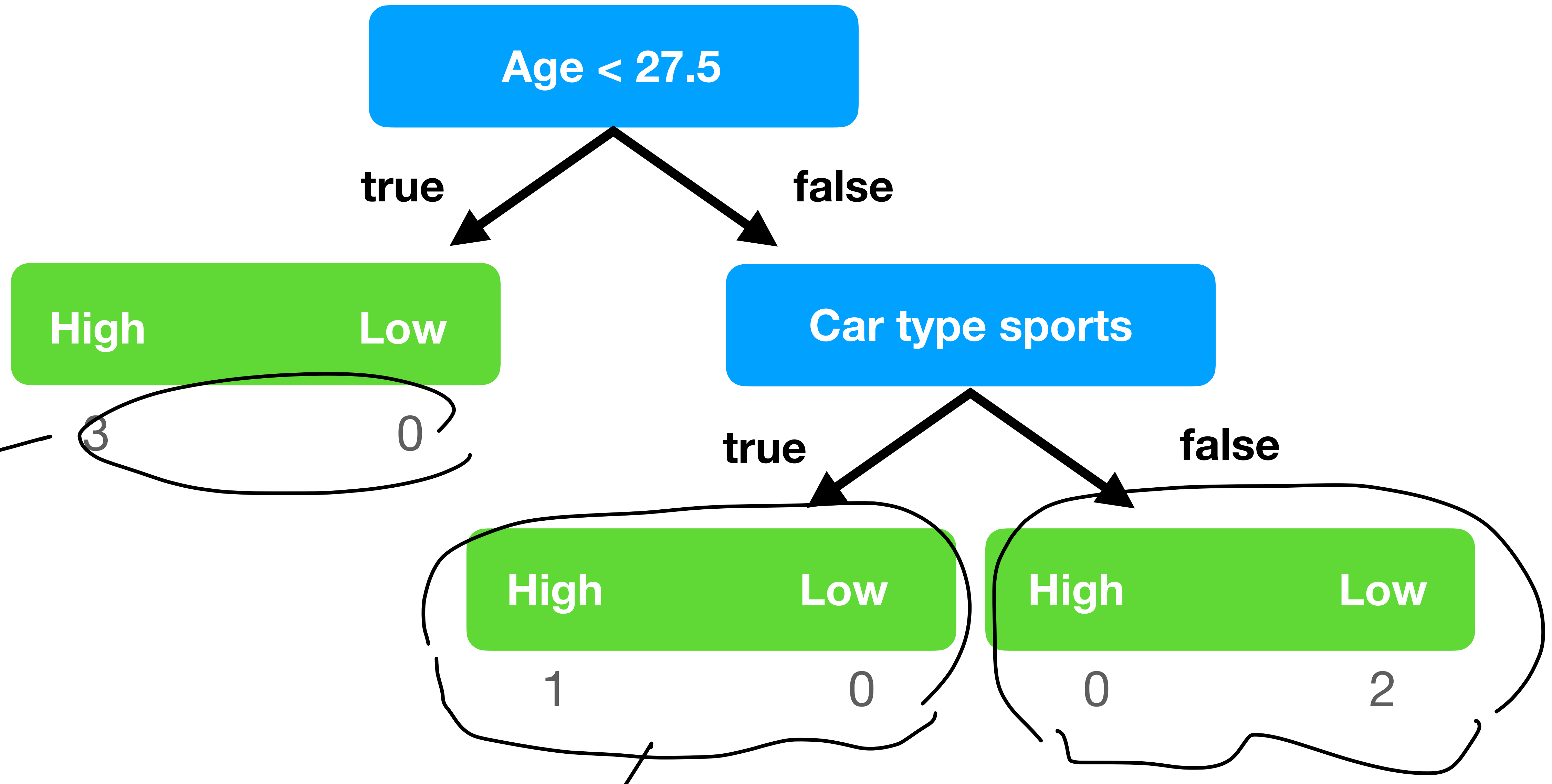
Age	Car type	Risk
23	family	high
17	sports	high
43	sports	high
68	family	low
32	family	low
20	family	high

Continuous Feature: Age; Categorical Feature: Car type; Class label: Risk



Classification tree

How do we use this for prediction?



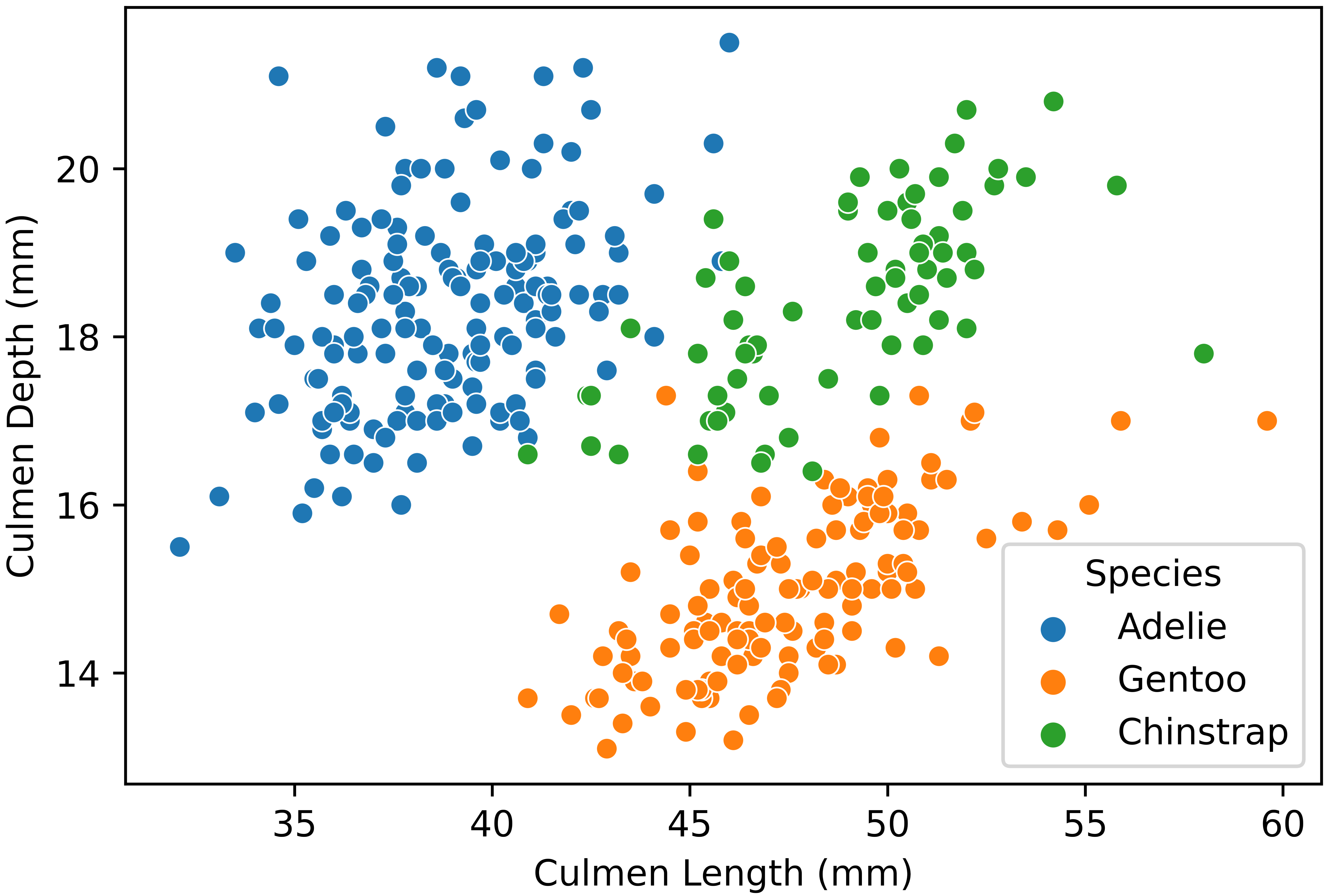
"High" risk

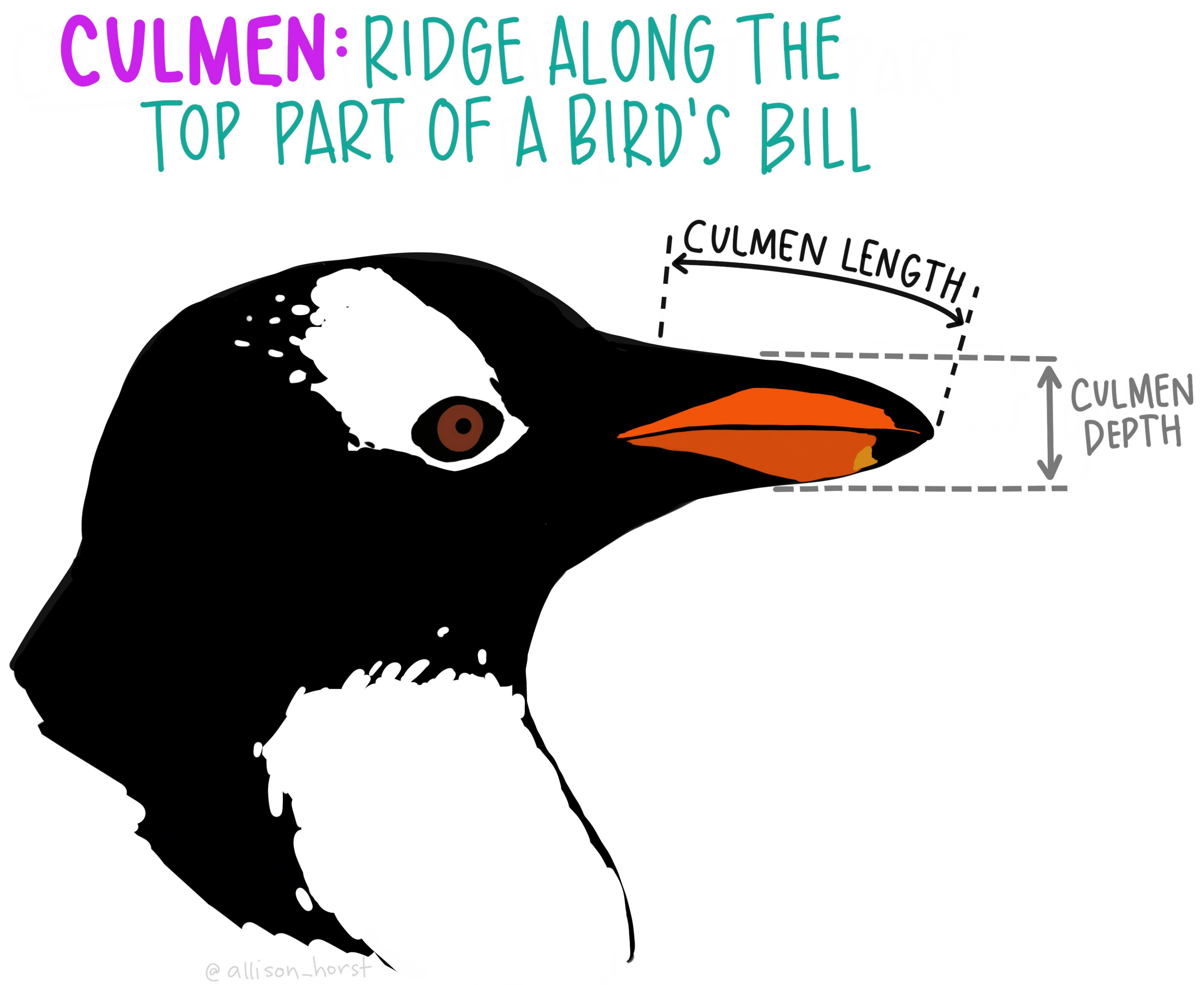
"High" risk

"low" risk

Classification tree example

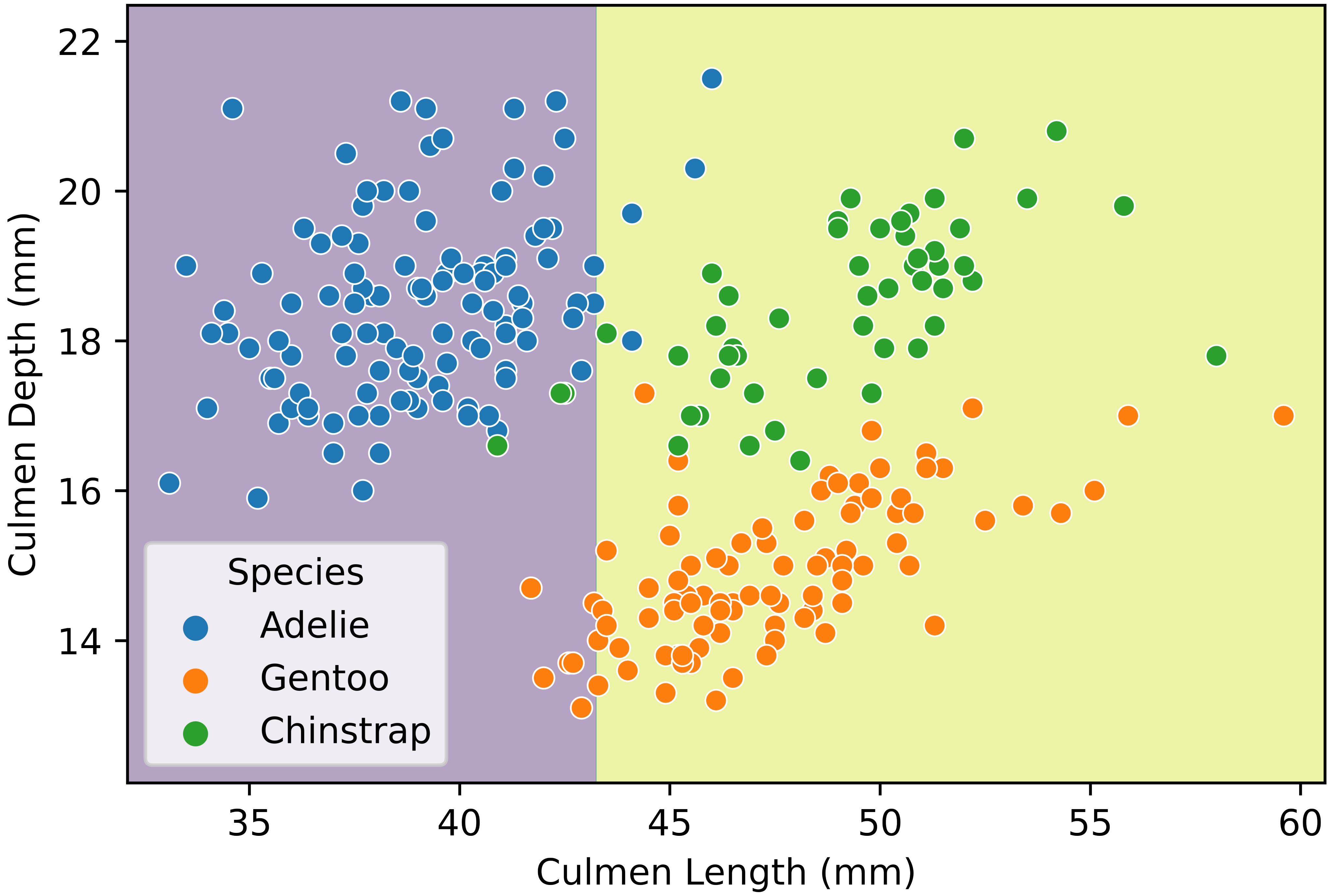
penguin data set

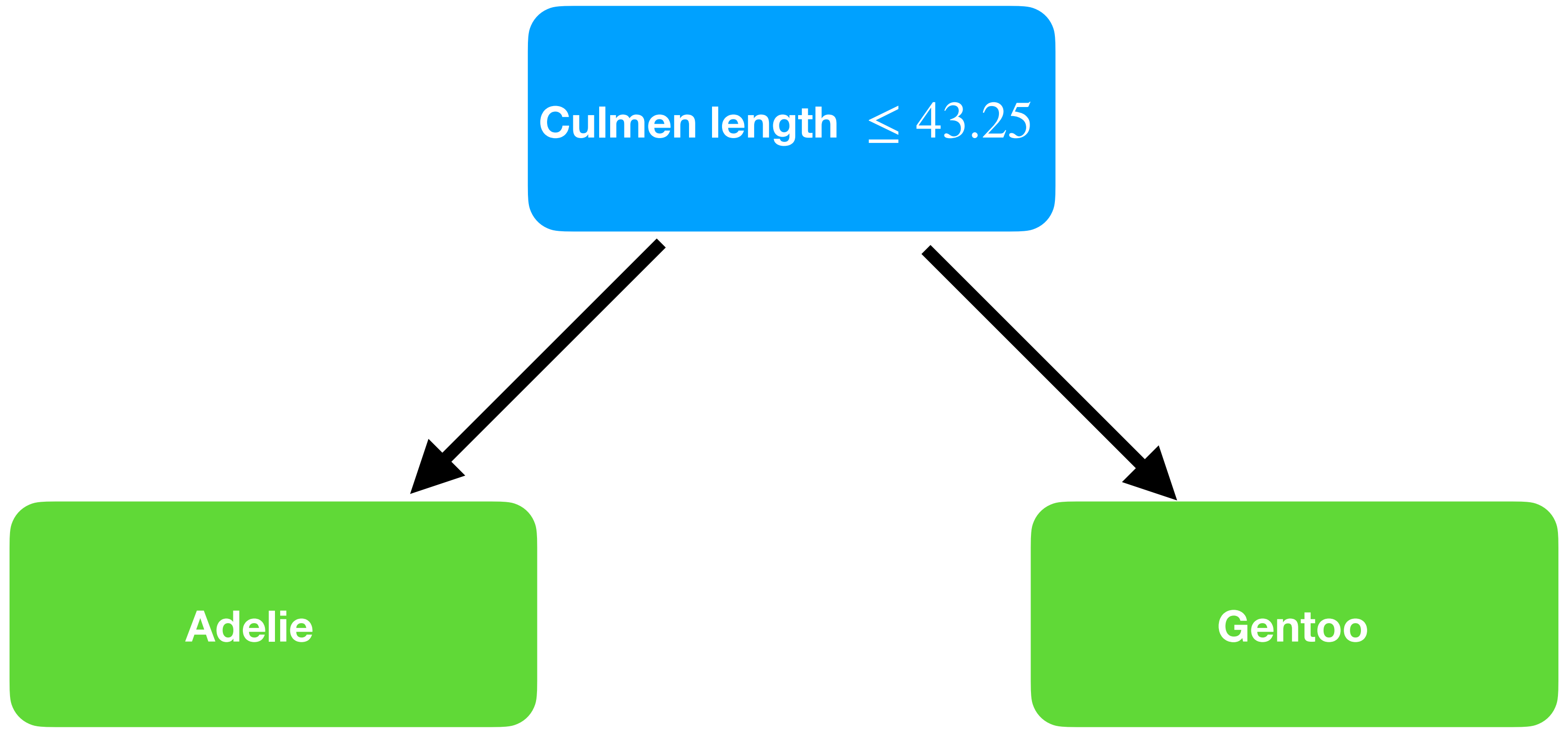




Decision Trees

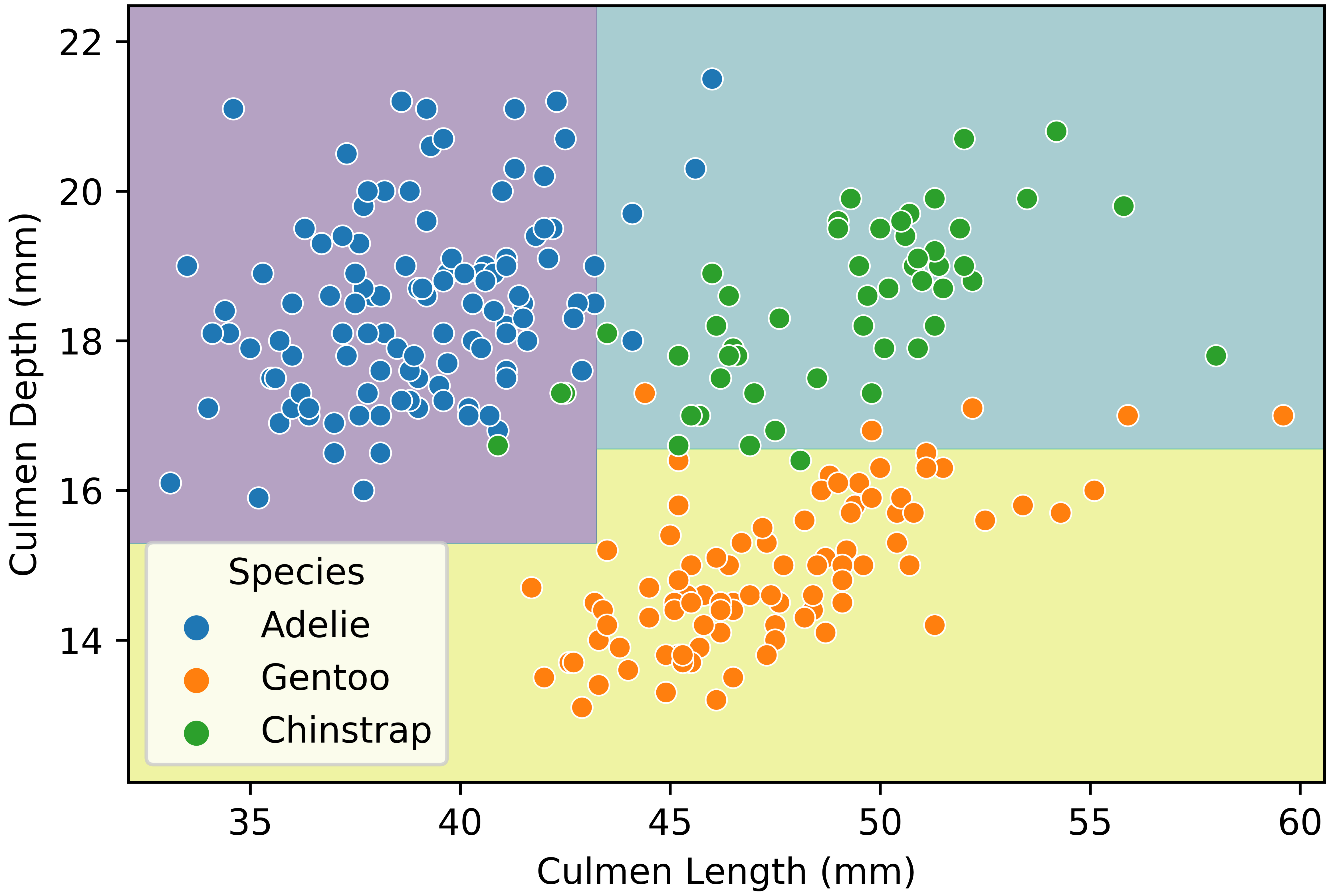
Penguins example

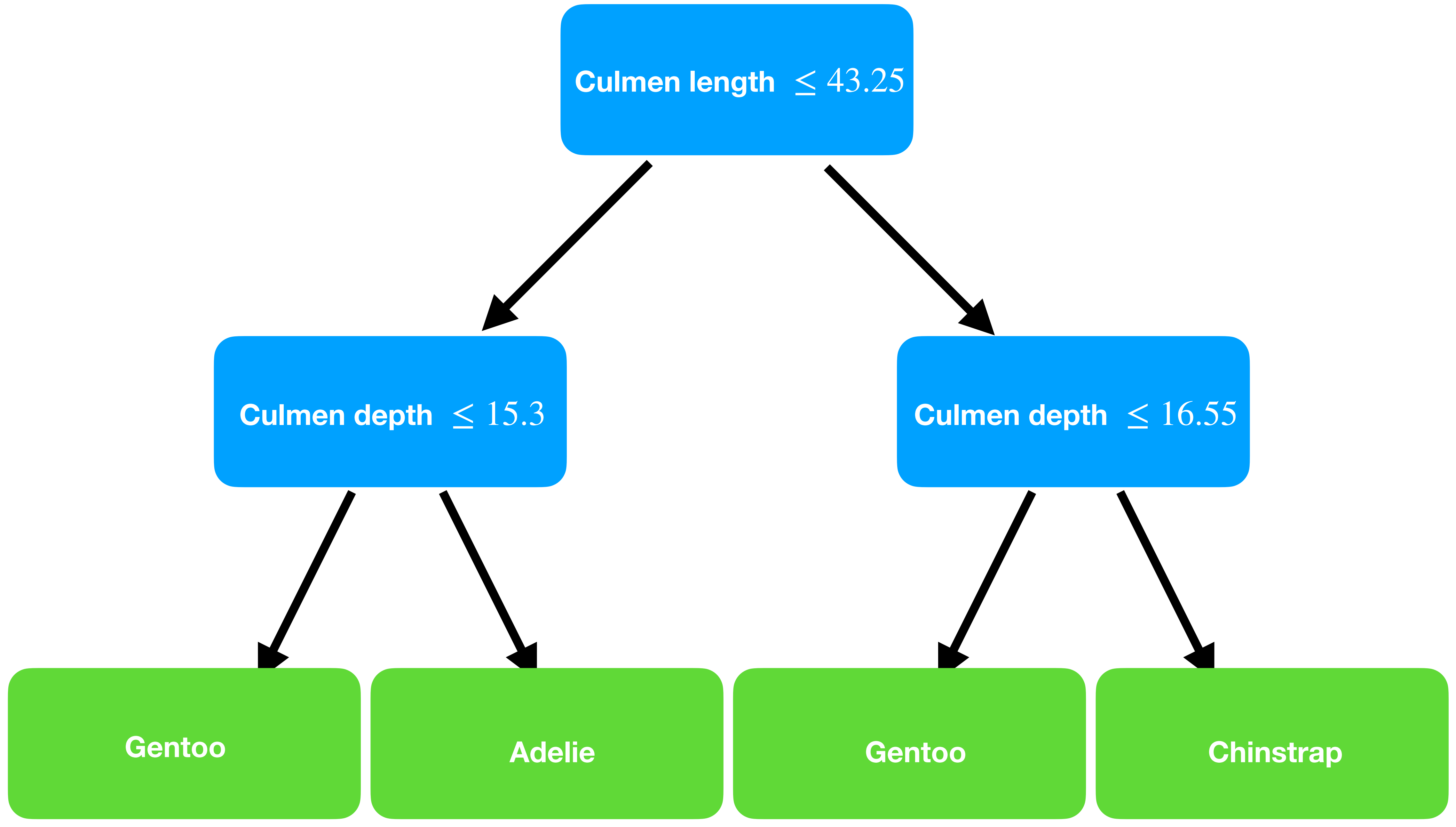




Decision Trees

Penguins example

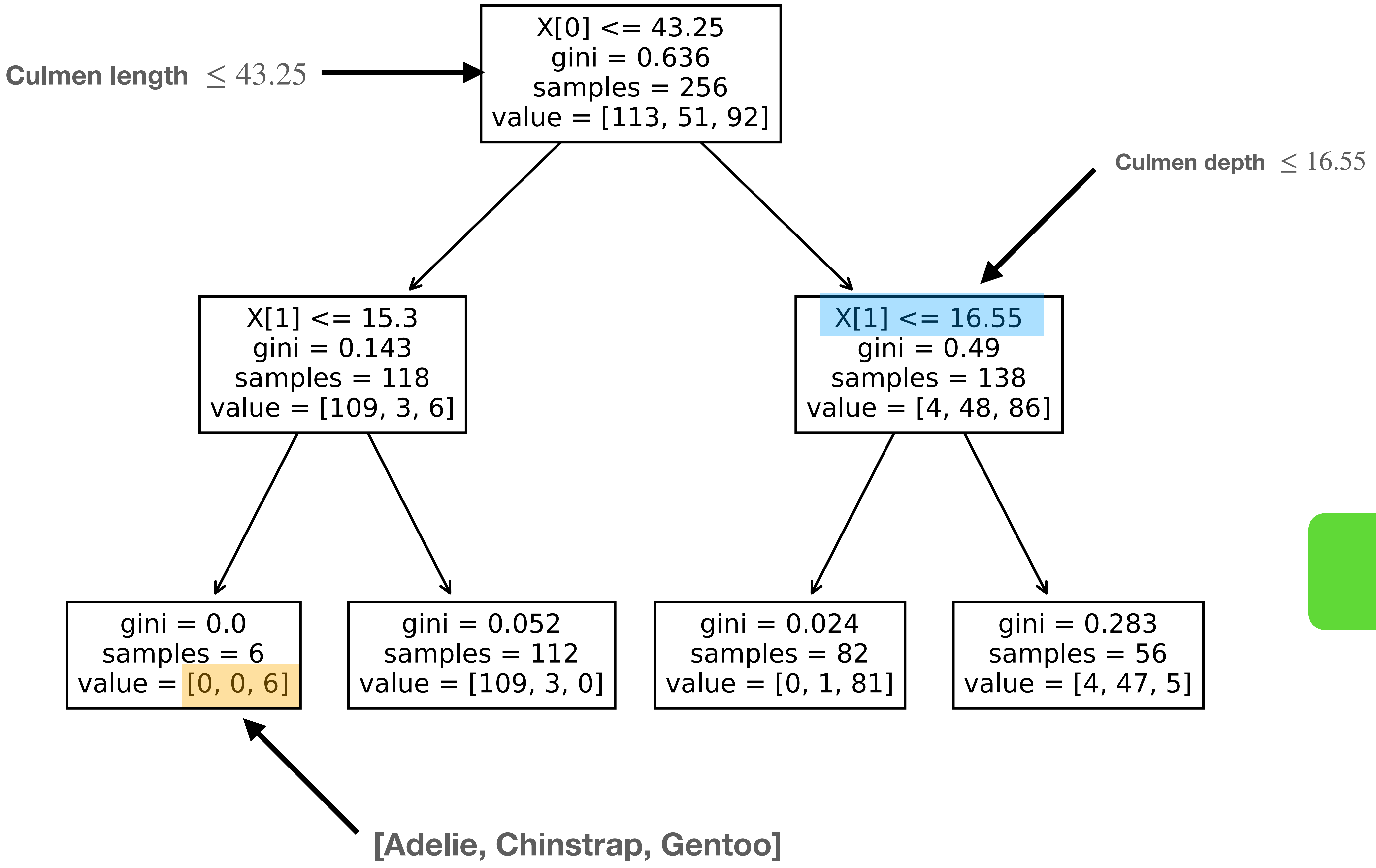


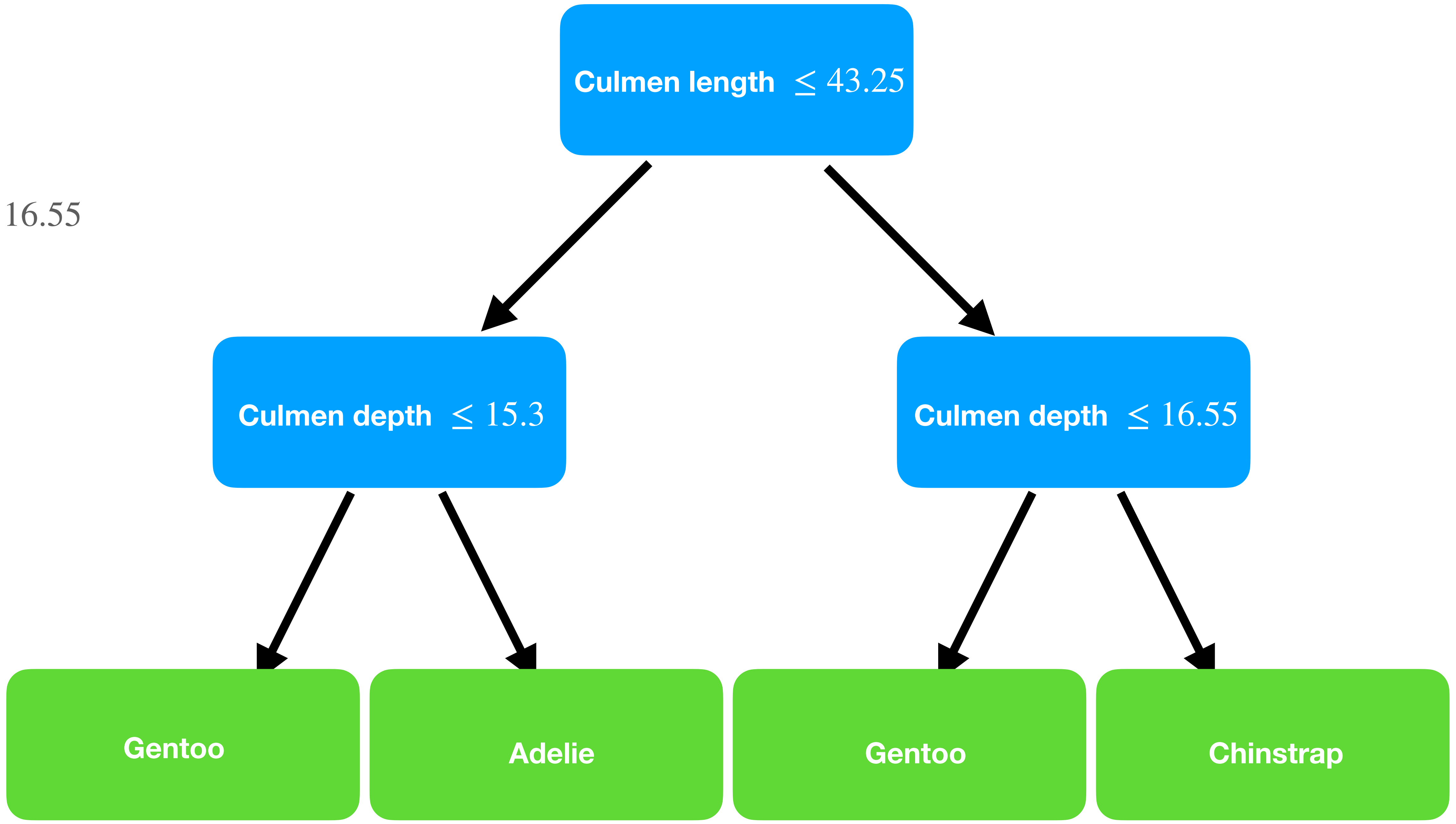


Decision Trees

Penguins example

Python output for *sklearn.tree*





Summary - decision tree construction

- We select the most discriminative **Feature and Threshold**
  - Discriminative power based on a criteria:
    - Regression tree: Average Squared Loss
    - Classification tree: Gini impurity
  - We create a node based on this feature
- We repeat for each new branch until a stopping criteria
- To ensure the tree-depth is not too high (avoid overfitting) “pruning” is done..

More examples on StatQuestion!!!: https://www.youtube.com/watch?v=_L39rN6gz7Y