Lesson 4: Decision trees

Experience AI

Raspberry Pi Foundation

Starter activity

What is classification?

Discuss with the person next to you.

Lesson 4: Decision trees

In this lesson, you will:

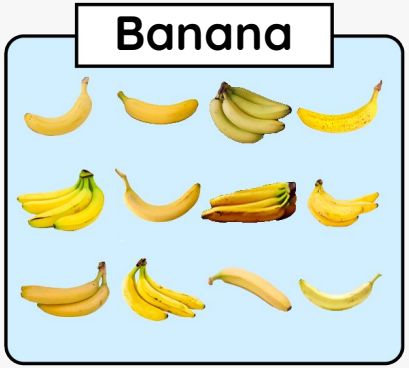
- Describe how decision trees are used to build a classification ML model
- Describe how training data changes an ML model
- Explain why ML is used to create decision trees

What is classification?

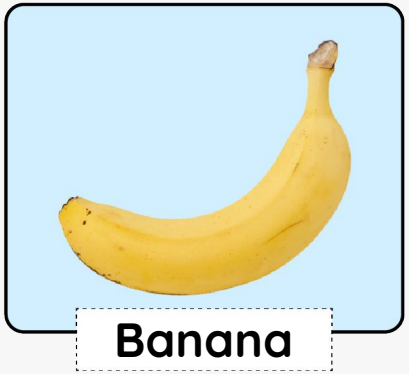
Classification is the process of assigning data to a **class** by applying **labels**.

A classification model is trained with **pre-labelled** data.

The model can then be used to **predict** a label for any new data.



Class



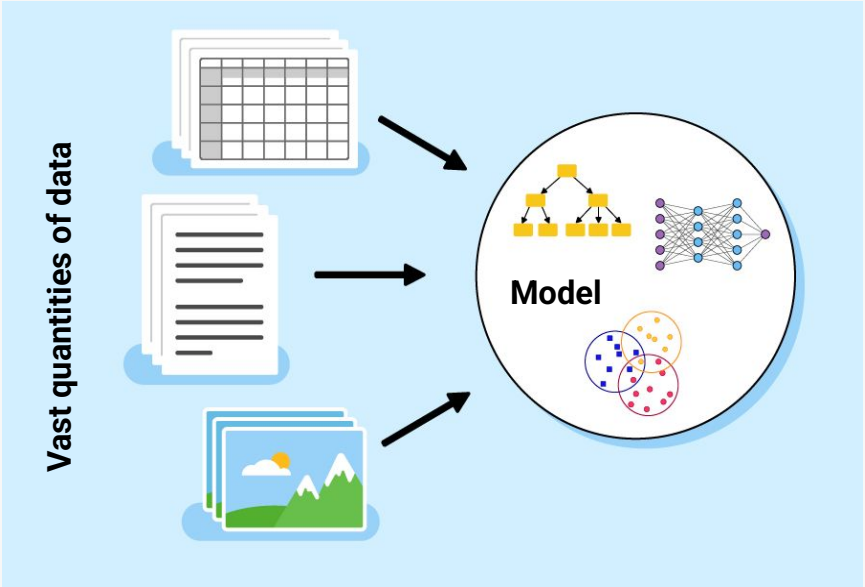
Label

What does a model look like?

There are different **types** of model that can be created using machine learning.

When choosing a type of model, you should consider:

- The type of data
- The problem you are trying to solve
- The need for **explainability**



What does a model look like?

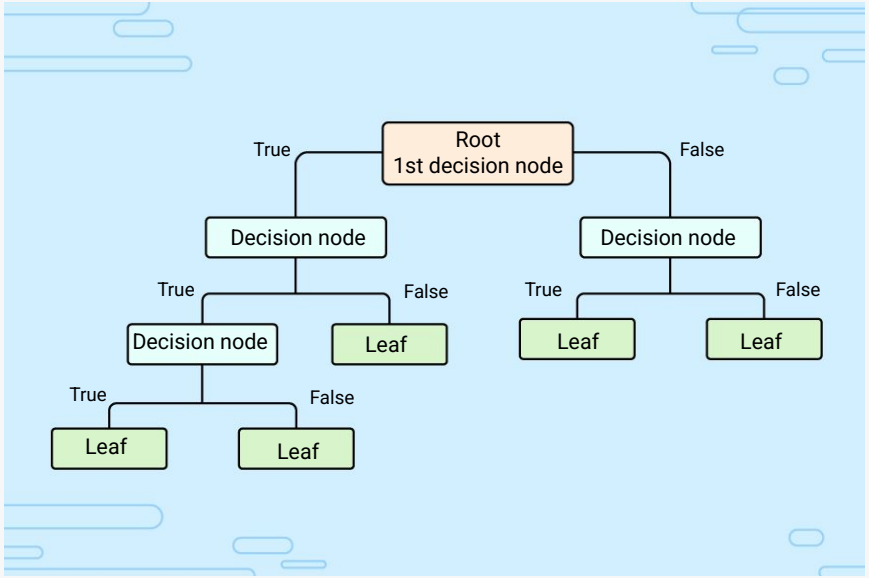
Watch the video on YouTube

A decision tree

Decision trees are a type of model that are created using supervised learning and can be used to classify data.

Decision trees are made up of **nodes**.

The top node of a decision tree is called the **root**.



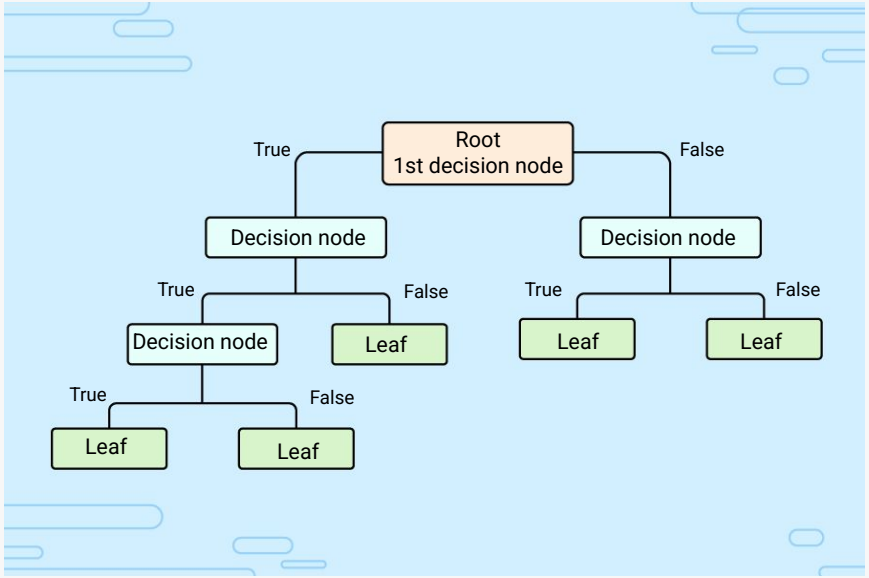
A decision tree

The nodes in a decision tree are either...

1. A decision node
2. A leaf

Decision nodes contain **conditions** that will split the data, commonly in two directions.

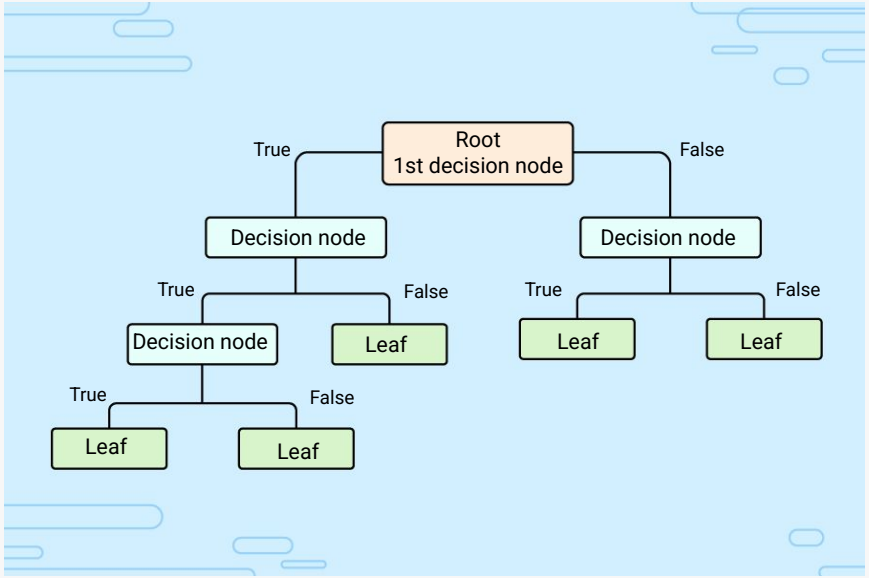
The root is the first decision node.



A decision tree

Leaf nodes will usually represent a single class.

When data is evaluated using a decision tree, the leaf you end on provides the predicted **label** for that data.



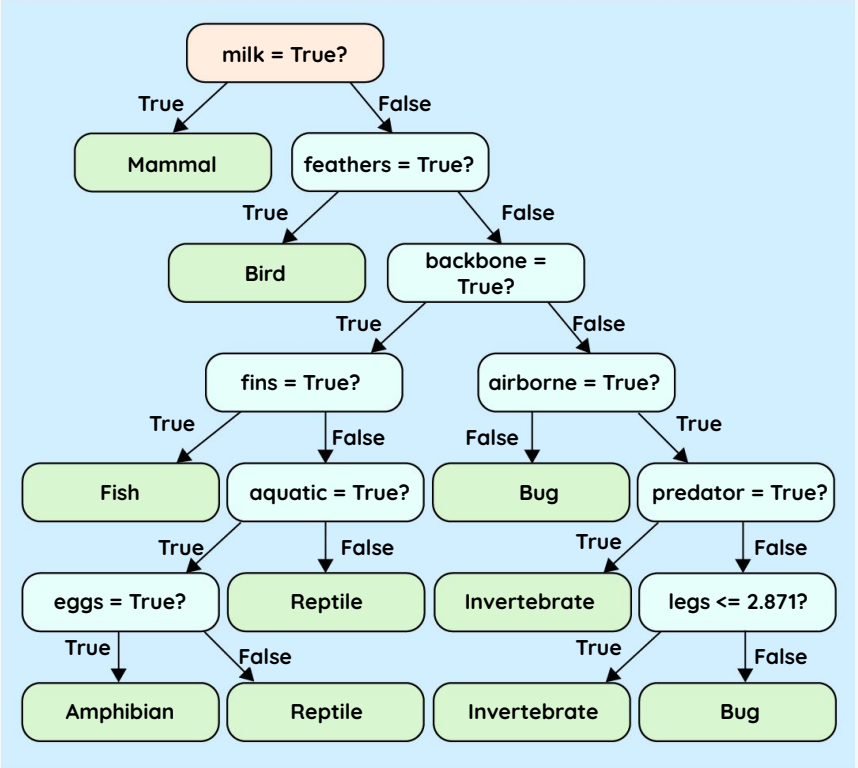
Classifying animals

This decision tree classifies animals.

The data set used to create a decision tree contains **features**.

Some features are either **True or False**, such as ‘feathers’, ‘airborne’, or ‘aquatic’.

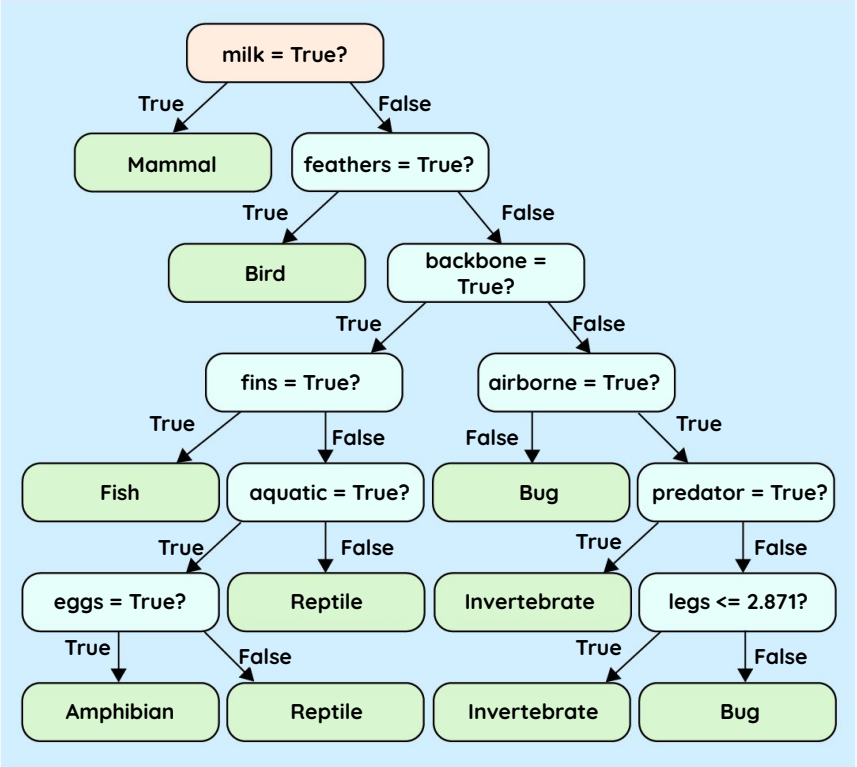
Other features are **numeric**, such as ‘number of legs’.



What would the prediction be for this animal?

hair	FALSE
feathers	FALSE
eggs	TRUE
milk	FALSE
airborne	FALSE
aquatic	TRUE
predator	TRUE
toothed	TRUE

backbone	TRUE
breathes	FALSE
venomous	FALSE
fins	TRUE
legs	0
tail	TRUE
domestic	FALSE
cat-sized	TRUE

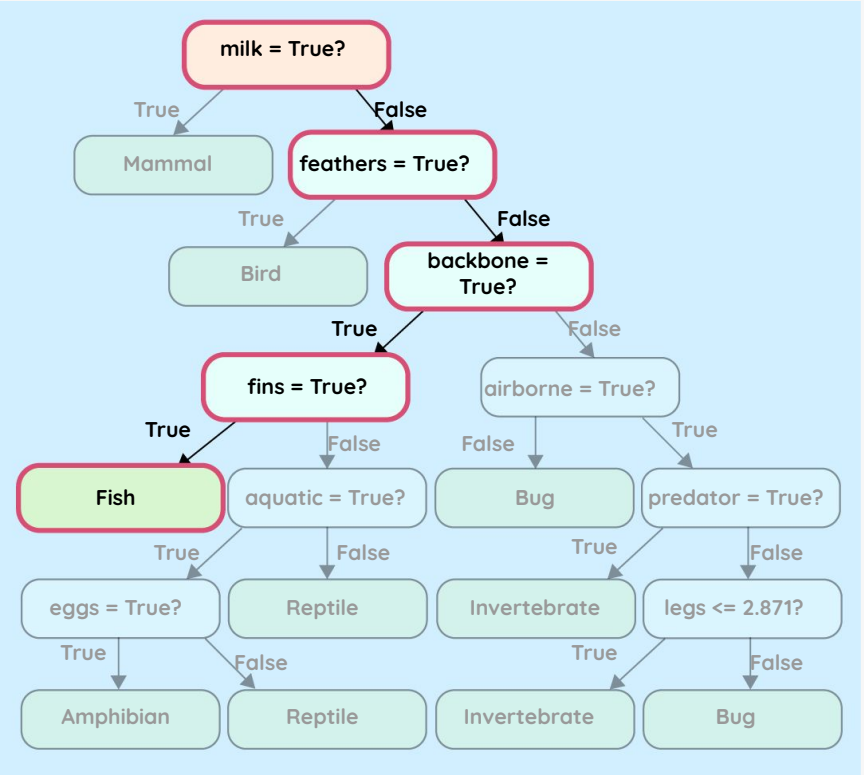


Think, pair, share

What would the prediction be for this animal?

hair	FALSE
feathers	FALSE
eggs	TRUE
milk	FALSE
airborne	FALSE
aquatic	TRUE
predator	TRUE
toothed	TRUE

backbone	TRUE
breathes	FALSE
venomous	FALSE
fins	TRUE
legs	0
tail	TRUE
domestic	FALSE
cat-sized	TRUE

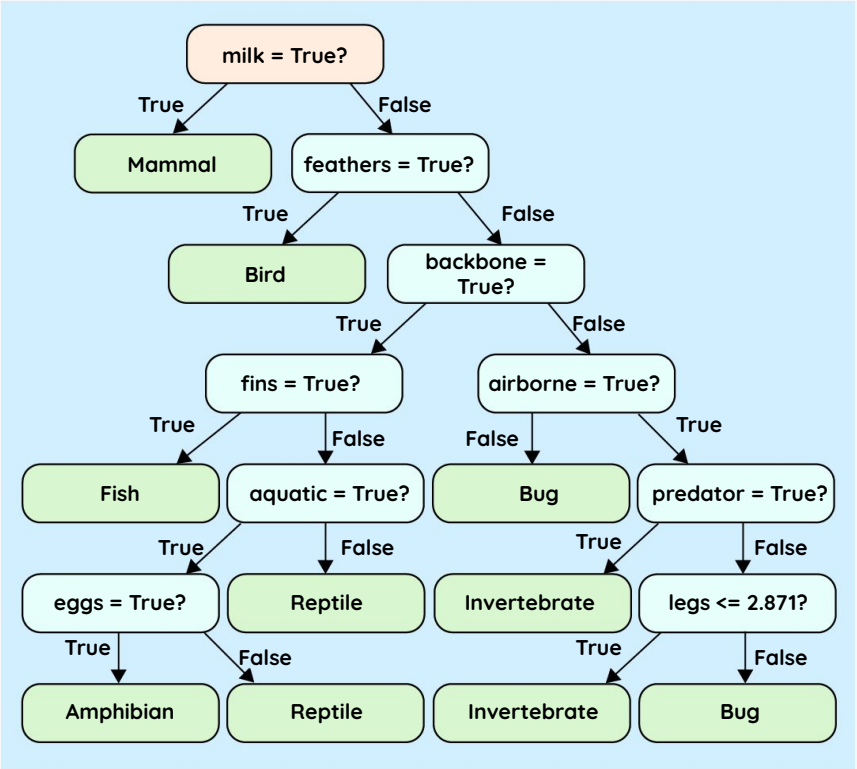


When to use a decision tree

Decision trees will only work with numeric or categorical data (like True/False).

They do not work with complex data like images, audio, video, or long text data.

Decision trees have a very high level of **explainability**, because you can follow the conditions to check the cause of a prediction.



How decision trees are made

We are going to develop a decision tree to help NASA classify new **stars** found by the James Webb Space Telescope.

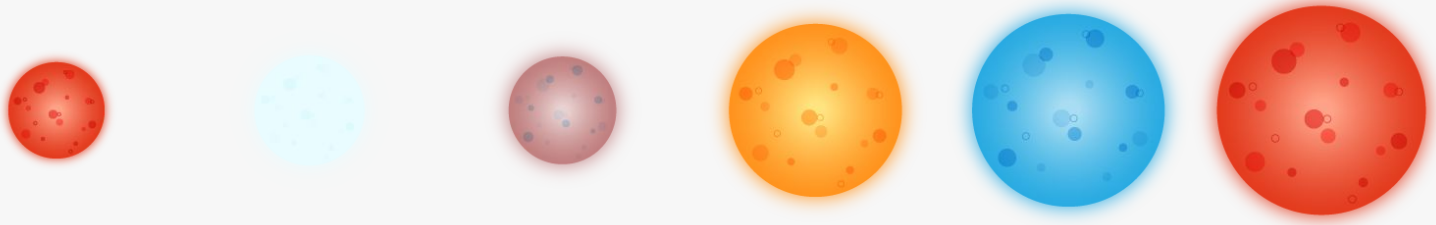
The telescope captures long-range images from space.

It captures 57GB worth of images every single day, so there is a lot of data to go through!

Types of stars

Stars can be in a number of different stages of their lives. Identifying what type of star we have found can help scientists understand our own sun by studying similar stars.

The types are:



Red Dwarf **White Dwarf** **Brown Dwarf** **Main Sequence** **Supergiant** **Hypergiant**

These will be our **labels**.

Features

We cannot use the images directly, so the data we are using is observations that can be made using the images:

- The temperature of the star
- The radius of the star
- The brightness of the star
- The colour of the star

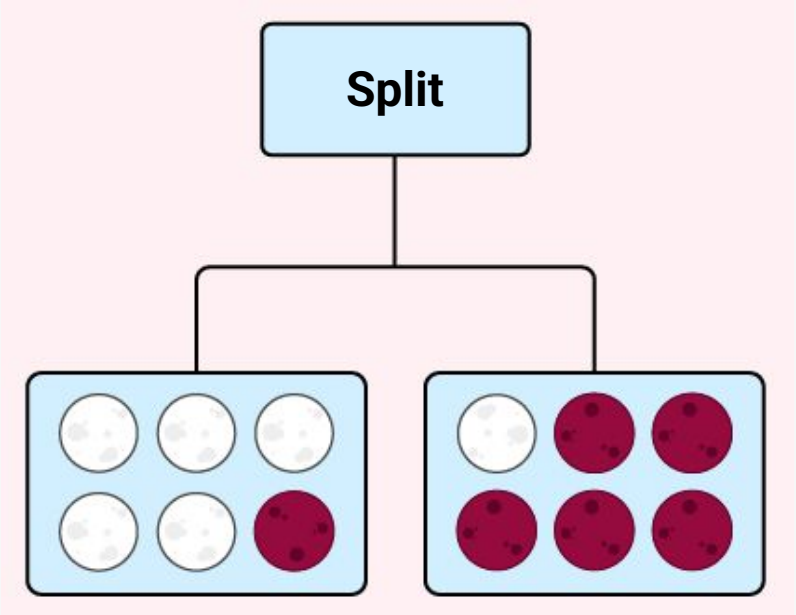
These are the **features**.

How to split the training data

When you start creating a decision tree, you start with all the training data.

As you create your decision tree, you will use **decision nodes** to split and separate the training data.

As the tree develops, the number of stars being considered at each decision node will get smaller.

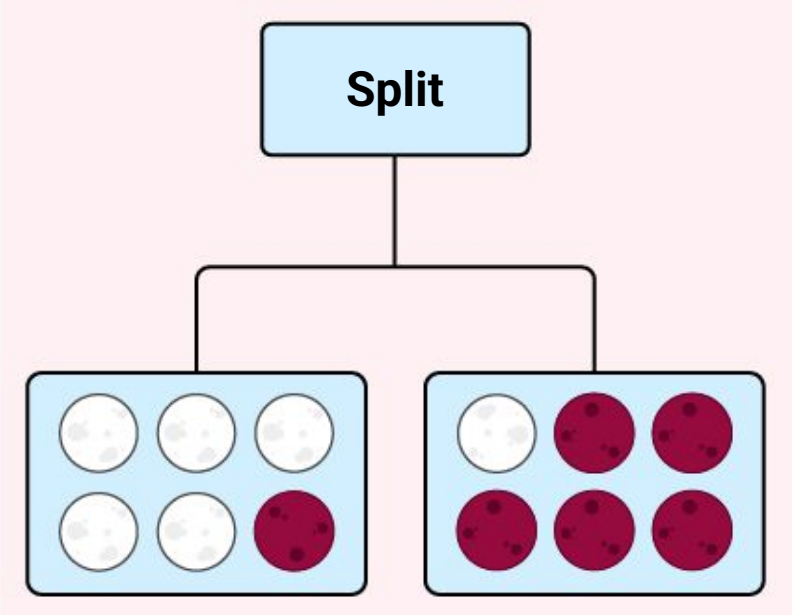


How to split the training data

The splits are made using **conditions**.

You need to decide which **feature** you will use to split the data at each decision node.

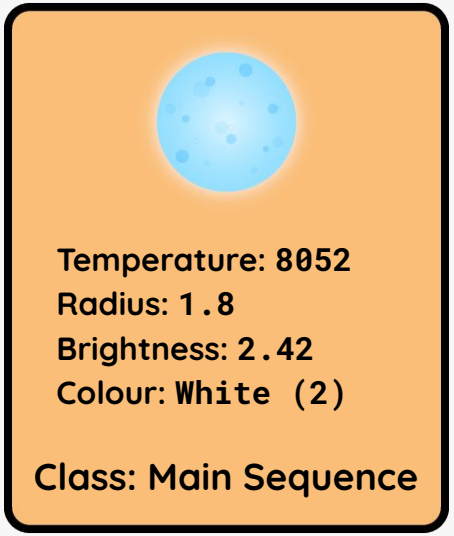
Let’s start with just two stars.



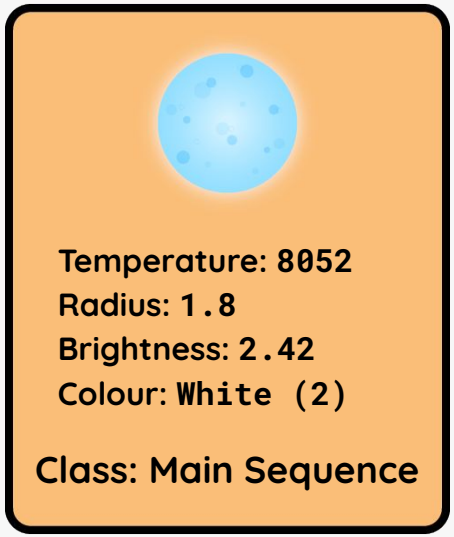
Star cards

For this activity, we are going to use **star cards**.

These cards have the data about a **single star** on them.

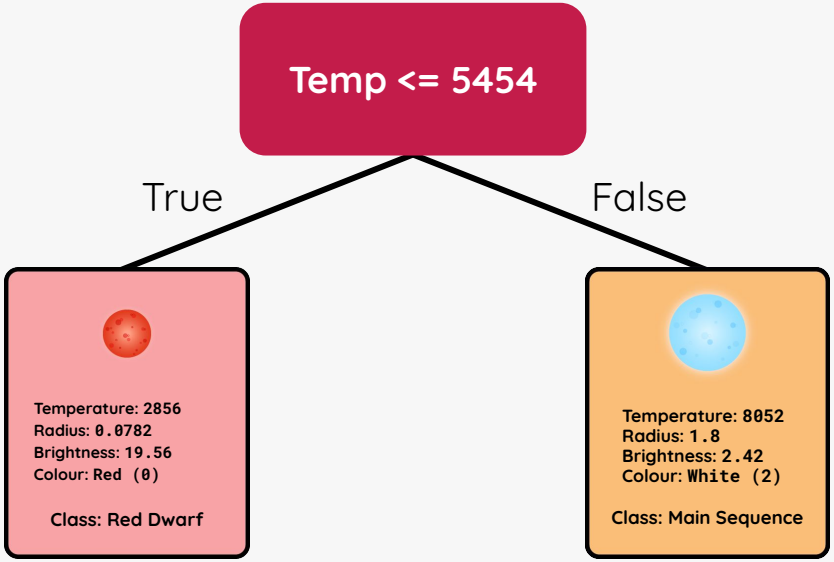


The first split



For this split, we are going to use temperature.

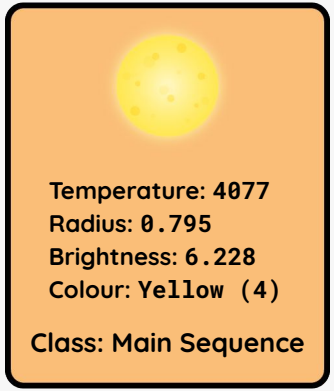
The first split

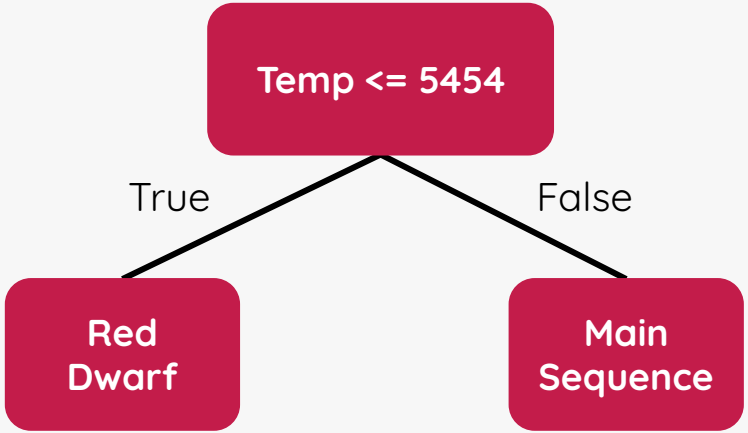


The **midpoint** between the temperatures is the value in the condition.

We have made our first split. Let's test it with another piece of training data.

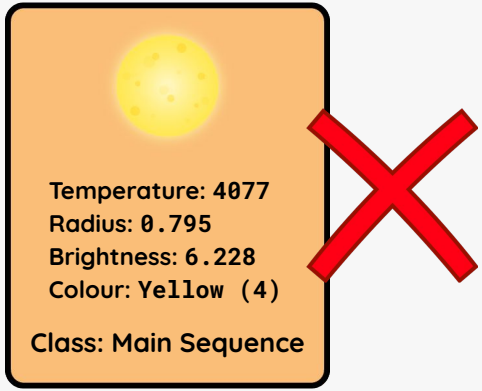
The first split

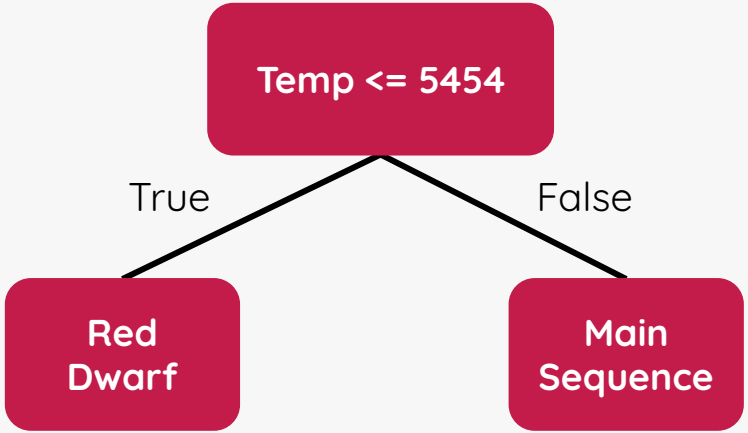




Does the decision tree correctly classify this new star?

The first split

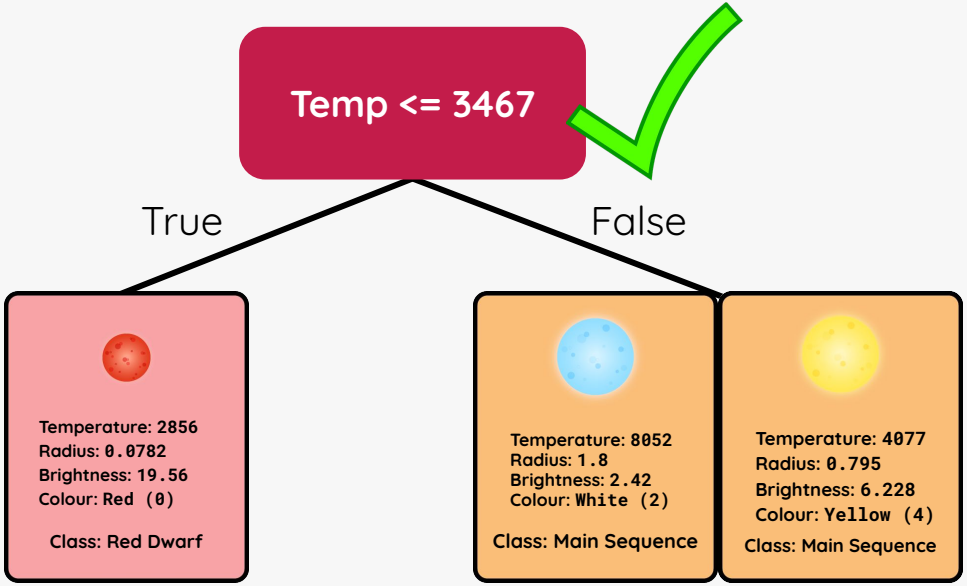




How can you fix it?

You have to adjust the split!

The first split



There you go!

As the decision tree is being trained, there will be lots of little adjustments like this.

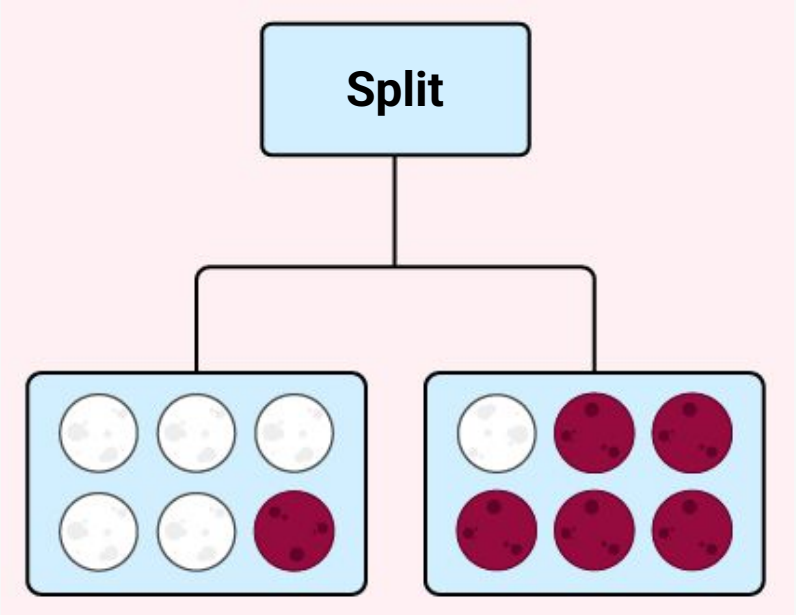
Create a decision tree

On your worksheet, you have six stars to split and classify.

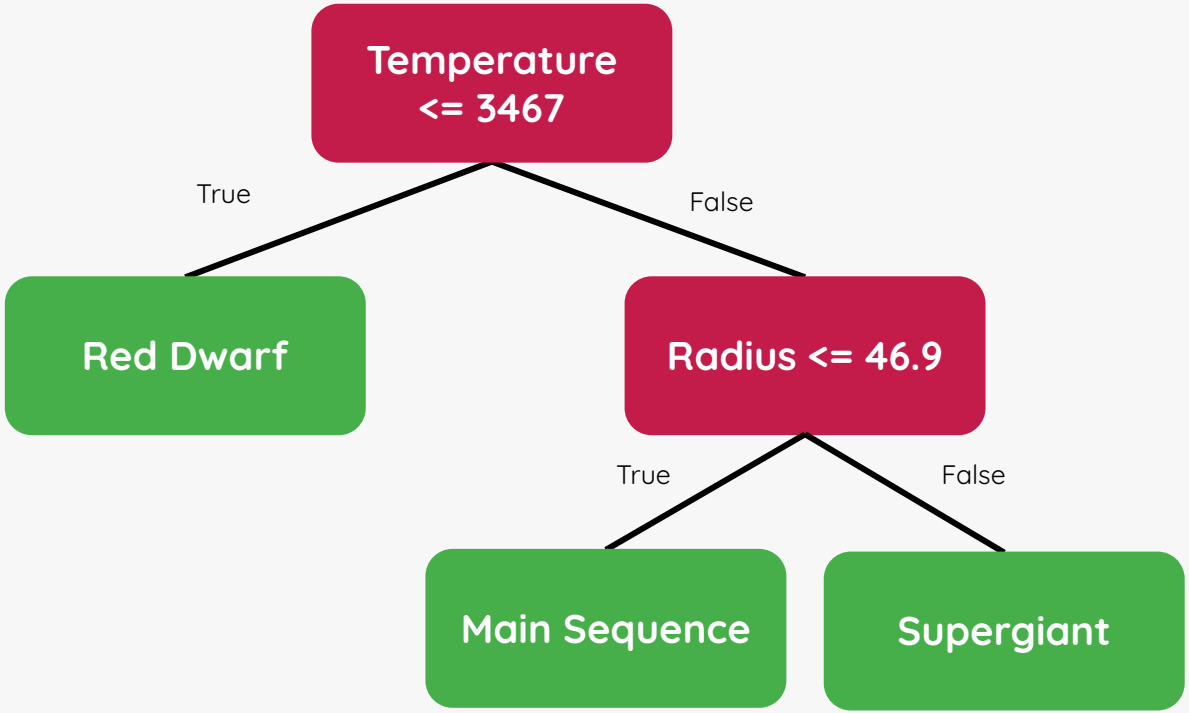
The top of the tree has already been made for you.

Try to make the last split!

The only rule is that you cannot use the same feature to split the data again, so don't use temperature.



Create a decision tree — Answer



Test your decision tree

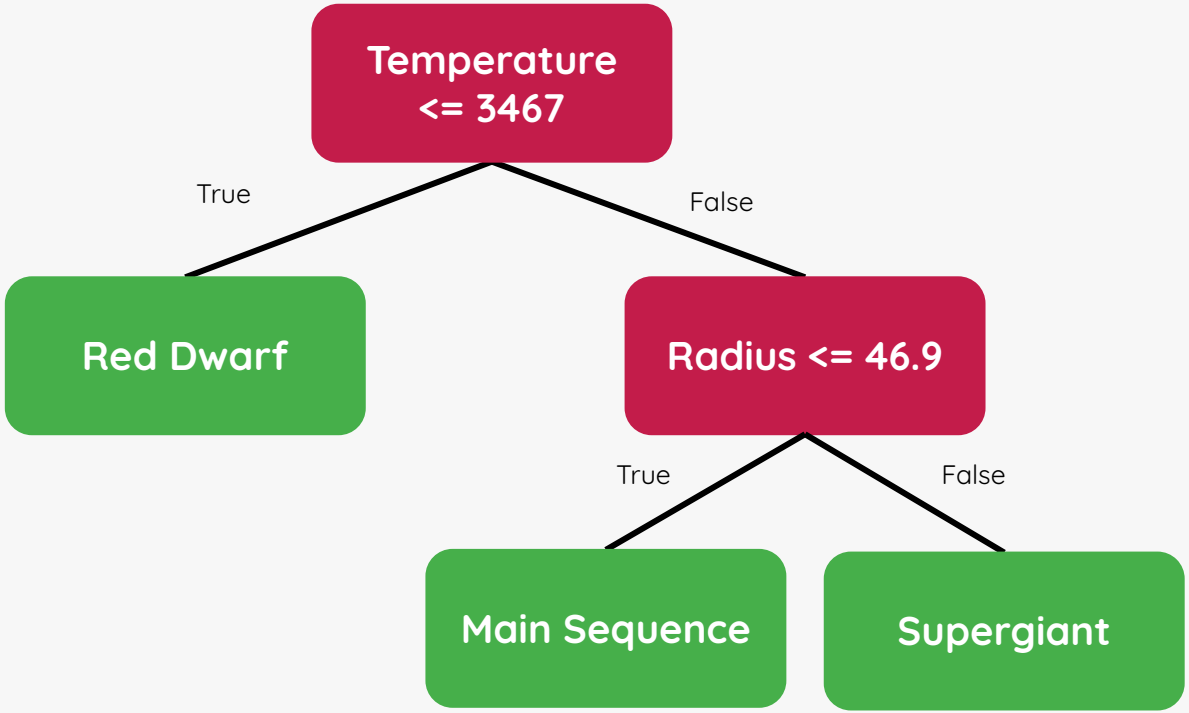
Test star 1

Temp: 36108

Radius: 10.2

Brightness: -4.4

Colour: Blue (2)



Test your decision tree

Test star 1

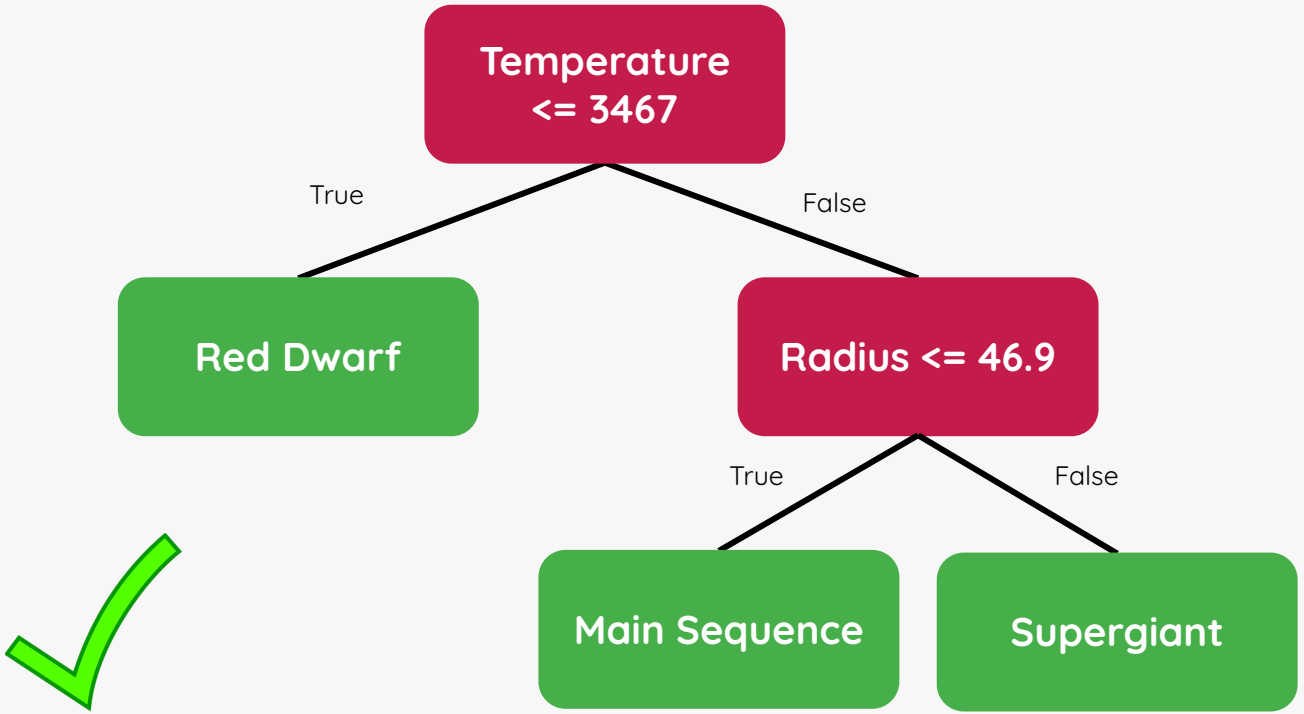
Temp: 36108

Radius: 10.2

Brightness: -4.4

Colour: Blue (2)

Main Sequence



Test your decision tree

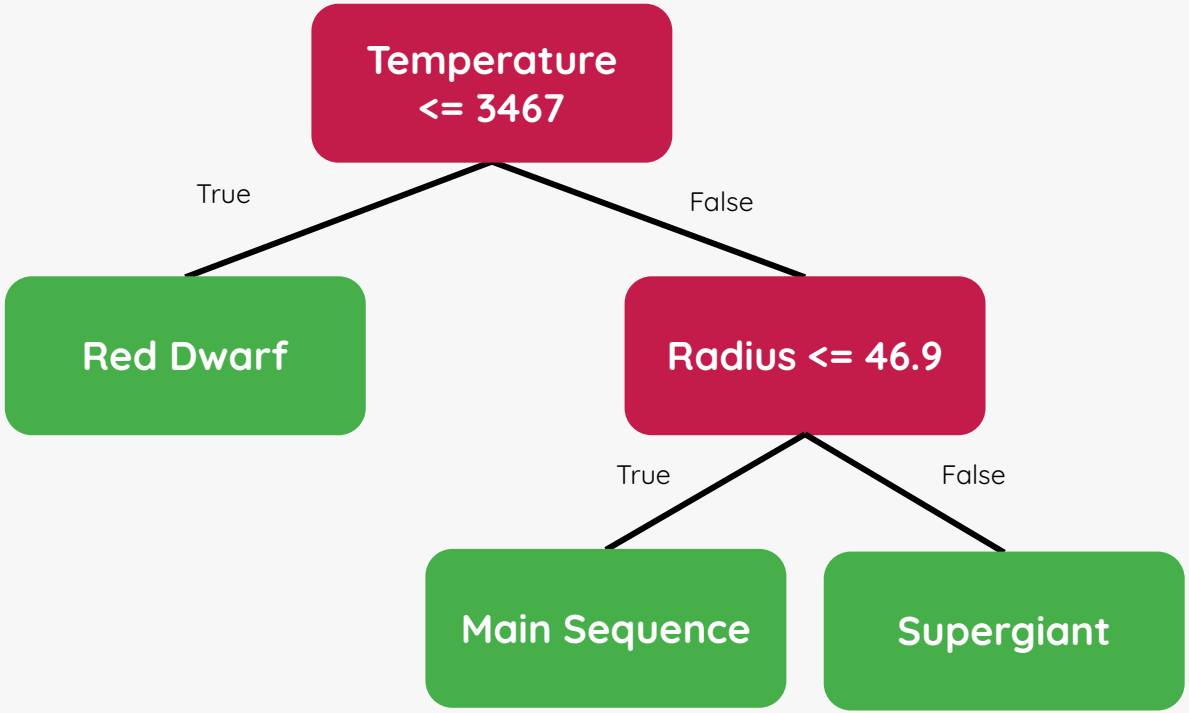
Test star 2

Temp: 11096

Radius: 12

Brightness: -5.91

Colour: Blue (2)



Test your decision tree

Test star 2

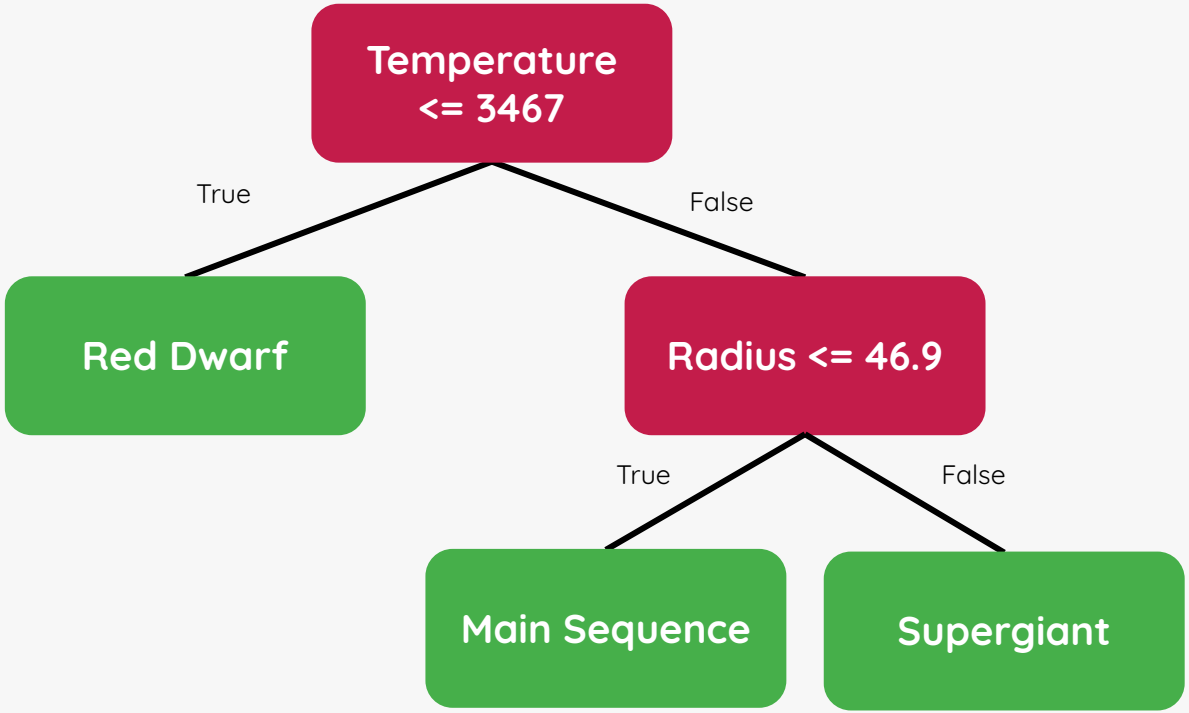
Temp: 11096

Radius: 12

Brightness: -5.91

Colour: Blue (2)

Supergiant



Data-driven models

There are some errors in this decision tree. Let's create another one.

The decision tree model that is created will change depending on the training data used to create it.

This is what it means to be data-driven.

Data-driven models

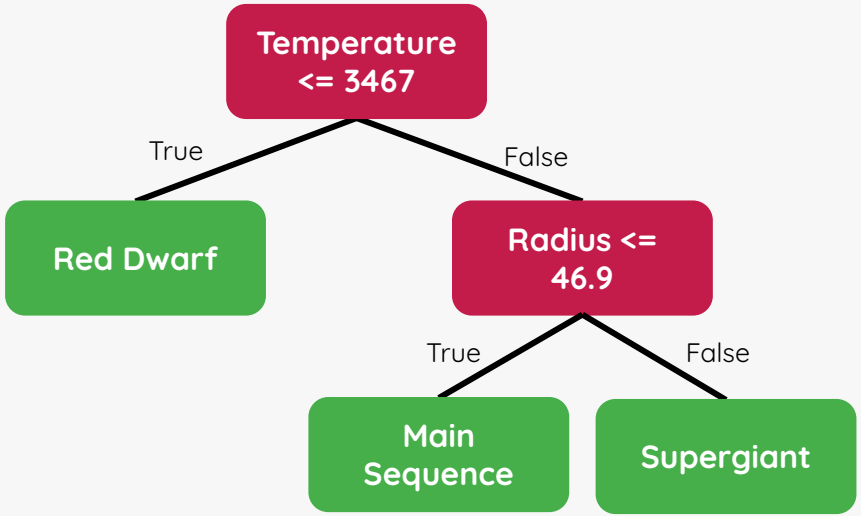
Open the second worksheet.

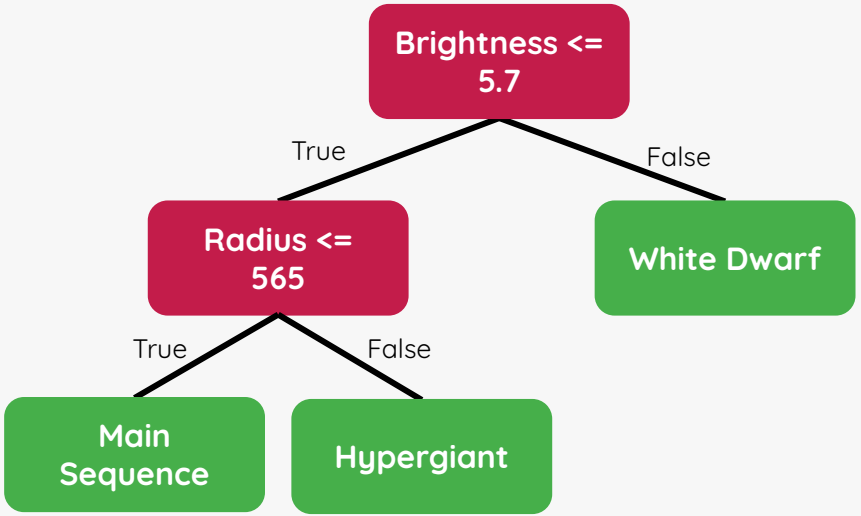
There is some different training data.

You have 4 minutes to split the data.

For this new data, the best feature for the first split is **brightness**.

Data-driven models

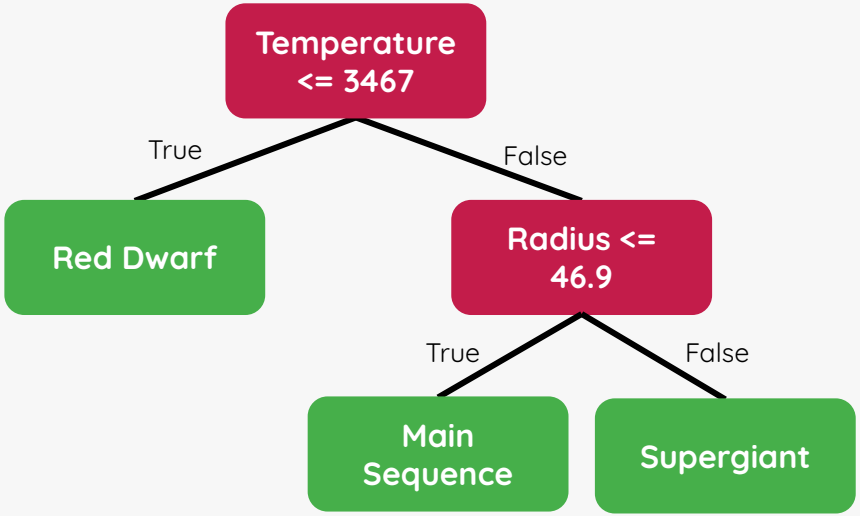




How does machine learning help?

You have been through the process of making a decision tree yourself.

How would machine learning help when creating a decision tree?

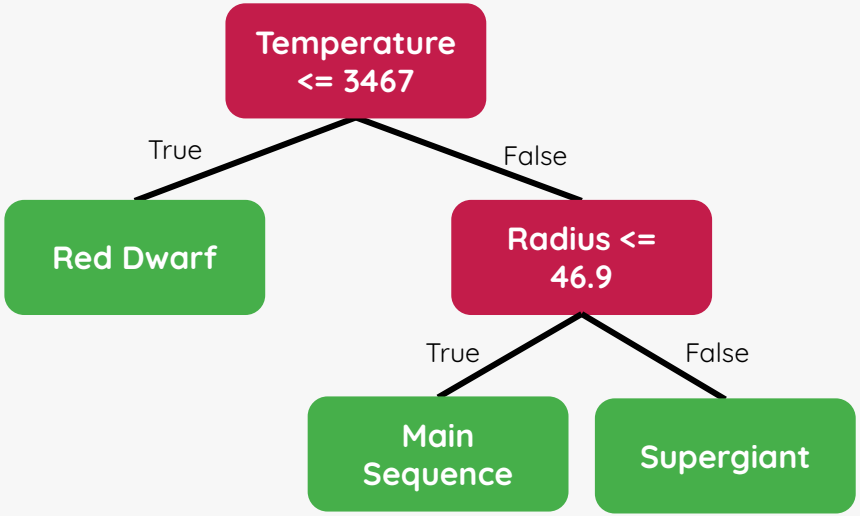


How does machine learning help?

You used 6 stars from 3 classes to make each decision tree.

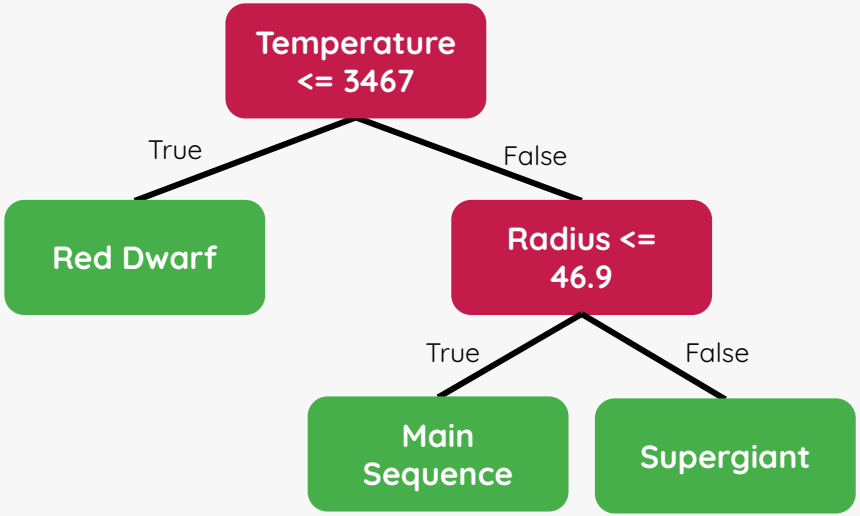
This was just a sample of **241** stars and 6 classes in the full data set.

The data set also only contained 4 features. Some data sets will contain 10 or more features.



How does machine learning help?

Let's have a look at a decision tree made using Machine Learning for Kids.



How does machine learning help?

Watch the video on YouTube

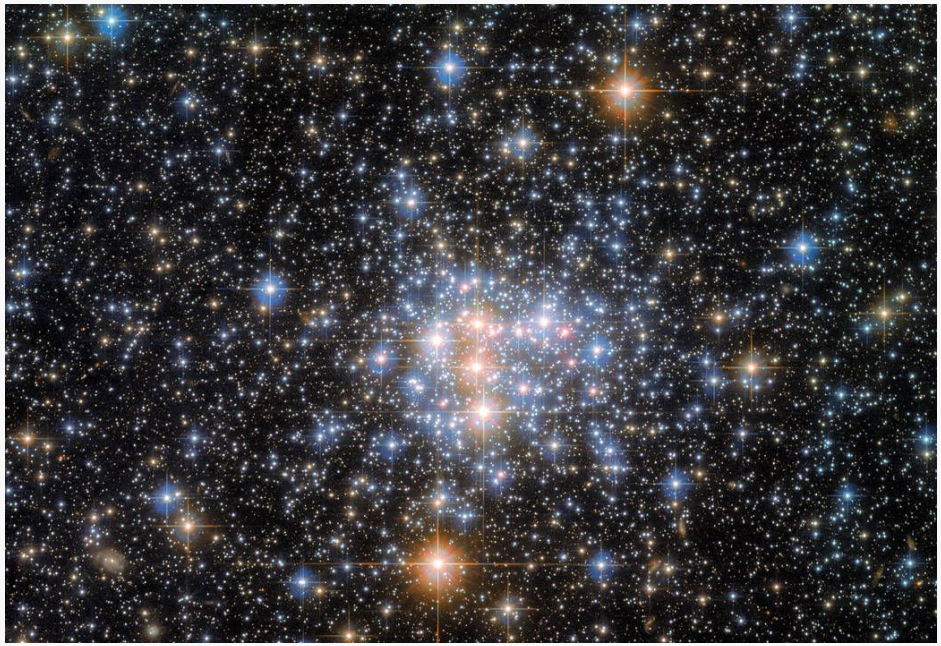
How does machine learning help?

Use your worksheet to create the star classifier.

Testing your machine learning model

Use the test data on Machine Learning for Kids to test your decision tree.

When you have finished, have a look at someone else's decision tree to see if it is the same as yours.



Decision trees in medicine

Decision trees are being considered for medical software that helps doctors diagnose patients.

Why would decision trees be a good choice of model for medical use?

Think, pair, share

Decision trees in medicine

Decision trees offer a high level of **explainability**, which is helpful when a patient or doctor would like to know why a certain diagnosis has been produced.

The data that doctors collect about patients is also quite often **numeric** or **categorical**.

Next lesson

In this lesson, you...

Learned about different types of model

Used a decision tree to predict a label for some data

Created a decision tree by hand

Used machine learning to create a larger decision tree

Next lesson, you will...

Choose a machine learning project

Create a model to solve a real-world problem