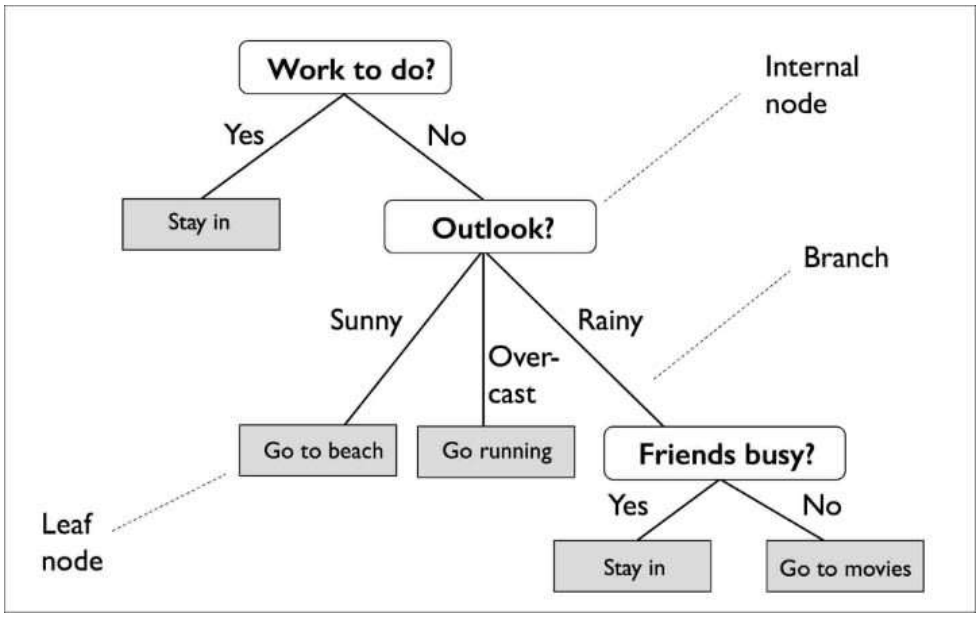
Module 3-3: Decision Trees and Random Forests

Another technique that can be used to classify data is a **decision tree**. The idea behind decision trees is that it will take a given feature from a dataset and "split" it by asking a question. Depending on the answer to this question, the decision tree will either ask more questions or make a prediction.

Let's look at an example. When creating a decision tree, we start at the top and work our way down. In this example, the goal is to predict what to currently do with free time.



"Work to do?", "Outlook?", and "Friends busy?" are all features of the data we are predicting. These are the equivalent to the "Thorns" and "Colors" in the last lesson's example.

"Stay in", "Go to beach", "Go running", and "Friends busy?" are all outputs of the data we are predicting. You can imagine them as classes, just like how "Rose" and "Daisy" were examples of classes in the last lesson.

The first thing we do is determine a possible answer to the feature we want to split on, which will eventually lead us to a concrete prediction. You can think of these as flow charts.

The order that you choose what features to split is based on how much information is gained by splitting the data. Splitting on certain features might lead you to a prediction faster than splitting on another prediction. In this example, splitting on "Work to do" first resulted in being able to make a prediction immediately if we answered yes. This restricts the options for later features and results in them not leading to unnecessary paths which could have been caught earlier.

Random Forests

Sometimes determining what order to select features might need trial and error. A tree with multiple features can be constructed in many ways with all the same content. **A random forest** is a machine learning technique that is comprised of multiple decision trees. The idea is that the data that is fed into the random forest creates multiple decision trees that split on the data in different orders.

Finally, to predict a certain outcome, the machine learning algorithm uses all of the decision trees and checks their individual responses. The random forest then returns the most popular output from the decision trees.