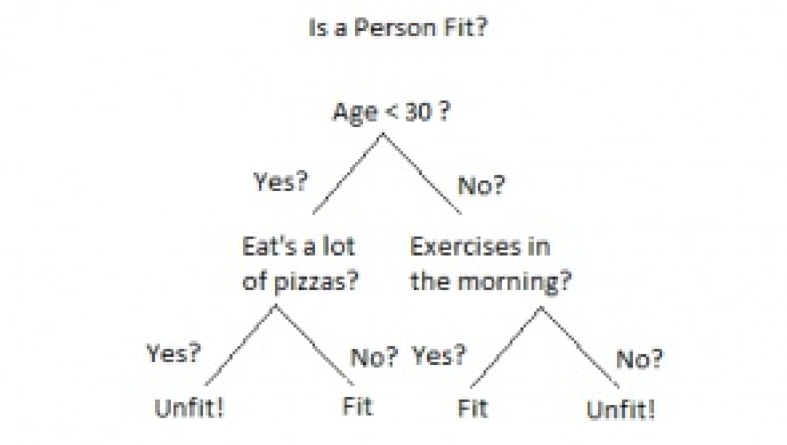
# Decision Trees and Random Forests

Miles McCain — CSC630, Machine Learning

A **decision tree** is a machine learning model that performs predictions by asking a series of *questions/tests* about the data. Decision trees can perform both regressions and classifications.

For example, given information about a person's age, exercise habits, and pizza consumption, a decision tree might predict whether a human is *fit* or *unfit* (out of shape) according to the following diagram:



*A basic decision tree diagram, by* ***Xoriant/Mayur Kulkami****. Forgive the typo on* `eat's`*!*

While helpful, decision trees on their own have a tendency to overfit to their training data. To account for this excessive variance, decision trees can be combined with one another to create an *ensemble*. Each decision tree is trained on a slightly different subset of the data so that its fit 'questions' vary slightly. The combined 'ensemble' model then performs its classifications by aggregating the classifications of its decision trees and returning the most common decision.

In effect, the decision trees each 'vote' on the proper classification. By pooling their decisions, the tendency of decision trees to overfit data is largely neutralized.

A **random forest** is a machine learning model that builds an optimized ensemble of decision trees. It randomizes the training data for each of its sub-trees, thereby minimizing overfitting (if tuned properly). The random forest's primary hyperparameter is the *number of estimators*—that is, how many sub-trees should the random forest train and consider when performing classifications.

This YouTube video may be a useful resource to understand random forests.