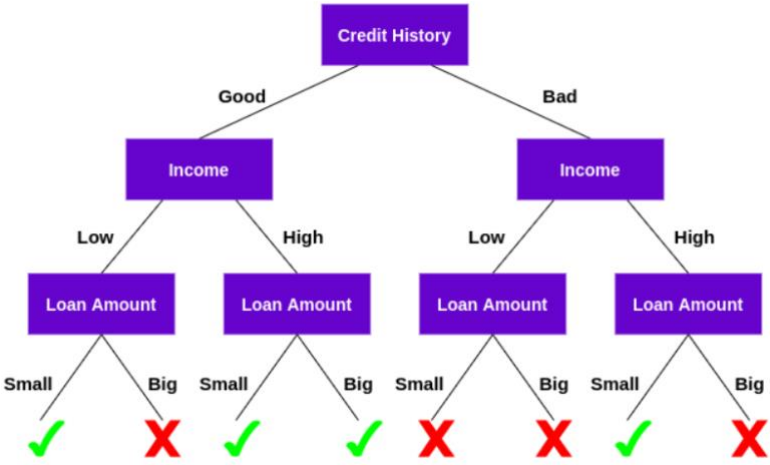
Classification and Clustering Methods Homework
By Stuart McColl

Decision Trees and Random Forests

Decision tree

In general, a **decision tree** asks a question and then provides a classification based on the answer given. A decision tree can be a series of sequential questions, which lead to a specific result.

For example, a mortgage provider will approve your application based upon certain criteria, or *questions* and the result, is whether or not the applicant is approved for the mortgage. Please see following example:



Source: [Click here for website]

Decision tree: Strengths and weaknesses

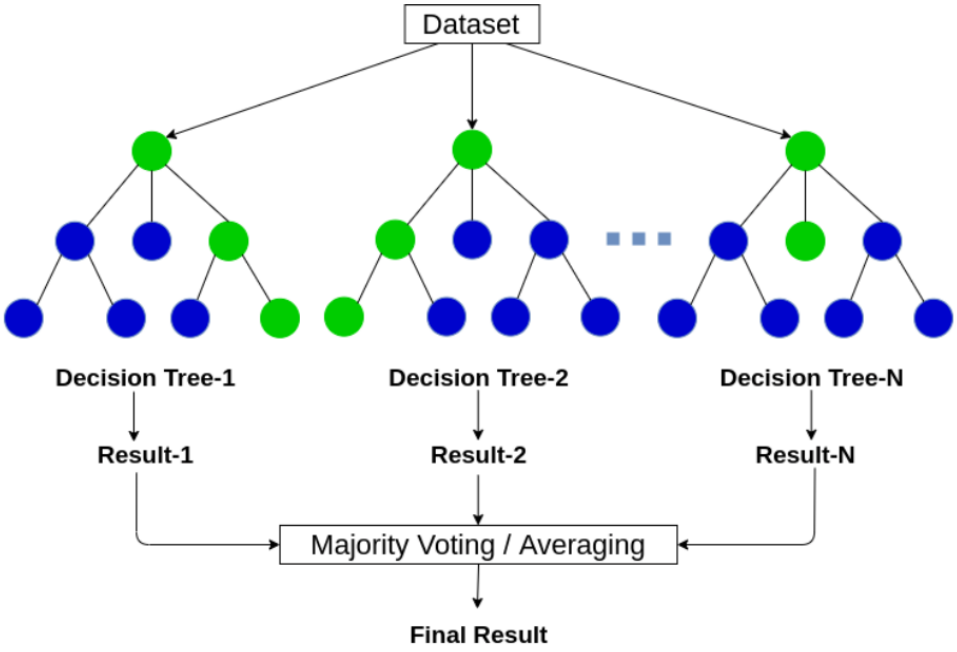
The **strength** of a decision tree is ultimately in its simplicity. They are very easy to follow and to interpret.

The main **weakness** of a decision tree is that they are vulnerable to over-fitting, which can ultimately lead to wrong predictions. This is more likely to occur when the data set is large as the tree will add more and more 'nodes', which makes the model difficult to interpret.

A decision tree, therefore, would be best suited to a smaller data set with few predictors.

Random forest

A random forest is a collection of decision trees and ultimately combines the output of multiple trees in order to generate the final output. Please see following example:



Source: Click here for website

Random forest: Strengths and weaknesses

The main **strength** of a random forest is in its ability to handle large data sets with a large number of variables.

The main **weakness** of a random forest is similar to that of the decision tree in that it is also vulnerable to over-fitting. Over-fitting is more likely to occur when the data set is considered 'noisy'.