Random Forests

A Random Forest (RF) [Breiman, 2001] is an ensemble model composed of many individually trained DTs, and is the most popular implementation of a bagged decision tree. Each component DT in a RF model makes a classification decision where the class with the maximum number of votes is determined to be the final classification for the input data. RFs can also be used for regression where the final output is determined by averaging over the individual tree outputs. The underlying principle of the RF algorithm is that a random subset of features is selected at each node of each tree; the samples for training each component tree are selected using bagging, which resamples (with replacement) the original set of datapoints. The high performance of this algorithm is achieved by minimizing correlation between trees while reducing model variance so that a large number of different trees provides greater accuracy than individual trees. However, this improved performance comes at the cost of an increase in bias and loss of interpretability (although variable importance can still be inferred through permutation tests).