Final project abstract

Three machine learning algorithms – decision trees, random forests, and K-nearest neighbors – were tested for their ability to predict the binary categorical variable “diagnosis” of the “Breast Cancer Wisconsin (Diagnosis)” dataset from UC Irvine, which consisted of 10 continuous measures of size and quality of cancerous tissue samples, each with a column for mean, maximum (“worst”) and standard error (“se”) values, for 30 columns in total. Models were iterated based on accuracy score. Decision trees seemed to perform best as a tree of one variable, which was used as an accuracy benchmark. Random forests were based on a set of 8 somewhat-arbitrarily chosen variables, with 200 trees (mtry = 5) yielding the best accuracy. For K-nearest neighbors, using the same set of 8 variables, a plot of training predictions vs. test predictions suggested k=16 as optimal, though it didn’t appear significantly better than other nearby k-values. The most accurate models of each algorithm were compared via a more comprehensive set of metrics derived from confusion matrices: accuracy and F1 scores appeared strong across the board, and comparable from model to model. K-nearest neighbors distinguished itself by avoiding false negatives better than the other two (i.e., noticeably higher recall score), which seems advantageous over avoiding false positives in the context of determining whether a tumor is malignant or not: a false positive can be mitigated by a second opinion, whereas a false negative may very well mean an un- or late-treated malignant tumor.