Introduction:

Classification and regression trees (CART) are a type of machine learning algorithm, which is designed to learn from labeled data and then make predictions on unlabeled data. In this case, the label indicates whether a patient has transitioned from mild to severe Alzheimer’s. The advantage of CART is that it represents a model as a decision tree and is thus easily interpretable. By contrast, ensemble methods utilize many different trees (or in general learners). One such ensemble method, AdaBoost, learns from a variety of trees, weighting each one depending on accuracy. Our initial belief was that for explanatory purposes, CART would be best, but for prediction purposes AdaBoost would be best. If the prediction results turned out to similar, then CART could be used for both prediction and explanation. We have found however that …. [Juan-Felipe, please fill in]

Methods:

First, patients were assigned target variables in two ways. For the *t* = 1.25 year trees, patients who remained in the study at 1.25 years from baseline and were still considered MCI were assigned a target of ‘1’. Patients still in the studywho had transitioned to severe Alzheimer’s before this time were assigned a target of ‘2’.. For the second test, *t* = 2.25 years and we considered only patients who remained in the study that length of time. These times (1.25, 2.25 years) were chosen in order to achieve a balance in patient outcomes.

For the CART work, first, a 10% leave-out set is selected from the overall population of MCI patients. This portion of patients is kept separate while building regression trees. The other 90% of patients are used in cross-validation to find the most important and successful predictors. Precision and recall statistics were used to measure predictor success, as well as the MATLAB function “predictor importance”, which quantifies the impact of the predictor on the mean squared error of the tree divided by the number of tree nodes associated with that predictor. Thresholds of absolute importance and importance weighted by precision and recall were set iteratively, and predictors not meeting that threshold were removed from consideration for that iteration. The threshold associated with the most accurate regression trees was ultimately chosen, and the predictors that met this threshold were saved for the final step.

From this list of predictors, 10,000 sets were randomly selected, with each predictor having a 50% chance to be included in a given set. Each such set was tested by 90% cross-validation using regression trees. The most accurate regression tree was ultimately chosen along with the corresponding predictor list. Finally, a model was built training with the entire cross-validation set and testing on the 10% leave-out set. This gives the CART accuracy model.

For AdaBoost, the same strategy was used (10% leave-out set and cross-validation based training on the other set) but without any separate step of feature selection. The result gives that AdaBoost accuracy model.

[If there is room we can also try an AdaBoost experiment where CART chooses the predictors. In that case, there would be three sets of results.]

Discussion:

The results of AdaBoost show a clear trend towards increased predictive capability as MRI data and 6-month data were added. [Juan-Felipe should adjust the numbers here] Precision increased by 6%, 4% and recall increased by 13%, 5%, for *t* = 1.25, 2.25 years respectively, from the first to the third dataset. With the MRI and 6-month data, predicting for *t* = 1.25 years, AdaBoost achieved precision and recall (83.0%, 90.3%) higher than the split (82.1%). Predicting for *t* = 2.25 years, the AdaBoost model achieved precision and recall (70.0%, 70.8%) higher than the split (59.9%) by an even greater margin, suggesting a greater certainty of these results.

Part of future work is to analyze longitudinal data and consider different methods of quantifying neural atrophy. The absolute data used here will be added to longitudinal data taken over several visits (up to several years apart). The factors associated with the transition from mild to severe may become more clear when considering the changes in variables in addition to their starting points. A reassignment of the target variable to a continuous measurement could provide further resolution.

The addition of AdaBoost software was successful likely because its algorithm chooses more weakly correlated markers than strong ones. AdaBoost might thus be a more appropriate choice for Alzheimer’s disease because the disease is complex.