Our goal for the project has been evolving as we learn more about our data. Currently, using the diabetes data we found on the UCI repository, we are trying to predict the probability of readmission by a patient to the hospital within 30 days. As far as progress towards this goal, we started by cleaning the data. Initially the data had 50 variables (features). To reduce the number of variables that we will be using, we started by removing features that will have no effect on our desired outcome, including the “encounter id” and “patient number” that essentially act as indeces. As we looked at each variable, we also noticed some missing data. The variables that had missing data were categorical, so we imputed the missing values with the mode of the non-missing values. This can introduce some bias into the data, but it is still better than removing the observations with missing information. Our dataset is quite large (101,766 observations), so doing a complete-case analysis was considered. However, we thought that the best model was to preserve all of the observations, while reducing the number of features as much as possible.

Despite what we had said in our data collection report, it turns out that the cleaning of the data was more involved than we initially thought it would be. Along with the removal of unnecessary variables and imputing missing data, most of the variables needed a mapping of some sort to assign each level of the categorical information to a number. Some mappings were already provided in the folder where the dataset was found. An example of a common mapping that we chose dealt with several drug variables. The data said “Up” if the dosage was increased during the patient’s visit, “Down” if it was decreased, “Steady” if the dosage was not changed, and “No” if that particular drug was not prescribed. We ended up using the mapping “No” = 0, “Down” = 1, “Steady” = 2, and “Up” = 3. This or something similar needed to be done for most of the variables. In addition, we normalized the variables so that they were all at the same scale. Some variables like “readmitted” were binary, meaning there was only 0 or 1. However, for a variable like the No-Down-Steady-Up example, the numbers 0, 1, 2 and 3 were converted to 0, 0.333, 0.667, and 1. This turned out to be a tedious process to do this for what ended up being 40 variables that we will use in our initial models (39 used to predict the “readmitted” outcome).

Once the data was cleaned, we knew that we wanted to find some model to form a baseline for the accuracy of predicting the probability of readmission by a patient. To achieve an initial accuracy percentage, we first split the data into a training set (90000 observations chosen randomly) and a testing set (the remaining observations). Because the response variable in question is binary (they were readmitted within 30 days, or they were not), we chose to initially perform a logistic regression to predict the probability of getting a “yes”. However, logistic regression is not necessarily our main technique that we will use. We wanted to experiment with some sort of model to start seeing which features would be significant. Further feature selection techniques will be discussed below, along with techniques that may be used to find an accurate model once relevant features have been selected. After doing the logistic regression using all of the variables in the normalized and cleaned data, the training set was able to correctly predict 88.89% of the results in the testing set. We will try to improve upon this number as we discover other techniques to achieve our prediction goal.