Model Analysis report.

The supervised learning models selected were K Nearest Neighbors and Decision Trees. The unsupervised learning model selected was K Means Clustering. Our output variable, patient readmission was split into three different outcomes, readmission within 30 days, readmission after 30 days, and no readmission. Ultimately the choice was made to not combine the readmission types. It is reasonable to conclude that readmission within 30 days and readmission after 30 days represent different contexts for patient health. Grouping them together would possibly result in the model making conclusions that do not fit the reality of the scenario for a given patient.

Decision Trees were chosen due to their capability of handling a multi classification scenario. This is done through splitting the data into different branches, and being able to handle imbalanced class distributions. K Means Clustering and K Nearest Neighbors were similarly chosen in order to handle multi classifications. K Nearest Neighbors could assign data to different classifications based on surrounding data points, and K Means Clustering could cluster data based on their classifications.

Ultimately however, all models performed fairly poorly, even with the use of hyperparameter tuning. Our Decision Tree model came out to an accuracy of 48.9%. Hyperparameters for max depth and impurity were tweaked to improve the accuracy to 56.7 from the mid 40s. K Nearest Neighbors achieved an accuracy of 52.7% from the mid 40s after the hyperparameter K was set to 10. For our K Means Clustering model, the Silhouette Score was equivalent to 0.11. The number of no readmission values was 47371 compared to 54864 in reality, the number of readmissions after 30 days values was 30862 compared to 35545, and the number of readmissions within 30 days was 23533 compared to 11357.

In light of this, further accuracy testing was not done for more specific data points based on race and gender. The decision to not perform dimensionality reduction was made in order to preserve the features of the data. Since the goal was to project outcomes for patients based on race and gender, dimensionality reduction had the possibility of removing the context of these features when translating them into different dimensional space.

Given the data, the conclusion can be made that the provided features do not strongly relate to patient readmission. As stated in our initial part 1 report, we are assuming our data will be able to show any trends or relation between demographics and type of care received. Being unable to find any trends even between patients in general and outcomes could be related to a few factors. The first of these is that the data is not

drawn from one particular hospital. Hospitals in lower income areas may treat patients more predisposed to health conditions due to socioeconomic factors than their counterparts in higher income areas for example. Another is that the standards and quality of care provided is consistent across each hospital. Many different variables could go into the treatment of the patient that is not reflected by the features in our dataset. The use of one hot encoding for our demographical data could possibly have also altered the accuracy of our models, due to the curse of dimensionality.

In light of these conclusions, Part 4 will go ahead with the decision tree model, and use the ensemble version, random forest model and test more extensively. For instance, dimensionality reduction will take place. Given the context of the project, the decision was made to not use methods like data replacement for missing values in a record, in order to stay as close to real world data as possible for a patient, as giving replacement data would possibly not be accurate to what a real patient record may look like (for instance, a patient with many previous visits being given an average number of medications, when in reality their medication number would be much higher). We will use data replacement, as well as reducing the use of one hot encoding on categorical data such as age and weight. Age and weight are currently split into categorical ranges. We will instead replace these ranges by picking numerical values within them.

All results and findings are attached on the next pages, which contain the pdf of the jupyter notebook used for this project.