The model chosen for refinement in part 4 of the project was the Decision Tree model. Rather than just using one Decision Tree however, a Random Forest model was selected. A Random Forest model is an ensemble model of Decision Trees, where the results of each Decision Tree are represented in our final output, which aims to be more accurate than a single Decision Tree. As there was poor accuracy with just a single Decision Tree, it was believed that an Ensemble Model would be able to capture the different types of errors the Decision Tree was facing, and result in a more accurate output.

Ultimately, this was not the case. There was only a marginal improvement to .502 - .503 in accuracy. The hyperparameters for the Random Forest model were n_estimators (The number of decision trees to be used in the random forest), criterion (The function used to measure the quality of a split), max_depth (The maximum depth of each decision tree), min_samples_split (The minimum number of samples required to split an internal node), min_samples_leaf (The minimum number of samples required to be at a leaf node), and bootstrap (To use sampling with replacement when building decision trees).

Two approaches were taken to the data in Part 4. To use data replacement on missing values, or not. For data replacement, categorical data was replaced with the mode for the field, and quantitative data was replaced with the mean for the field. Despite the fact that giving replacement data would possibly not be accurate to what a real patient record may look like, the decision was made to at least try it and see if there would be anything gained from it. In particular, it would allow for greater values of n_estimators, as there would be more data to work with.

Ultimately the non replacement data approach proved marginally more accurate at .503. In light of this, just to test the original hypothesis that demographics of a patient would lead to different trends in quality of care, the non replacement data dataset was split into male and female records. As medical care is often centered around men (for instance, pharmaceutical drugs are frequently tested only on men), in addition to women facing different issues in their standard of care, it was expected that there would be some sort of increase in accuracy for male records, and/or a decrease in accuracy for female records. But once again only marginal differences were detected.

Once again, we are forced to restate conclusions from part 3 as to why this is. 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.

Ultimately, for future work on the project, I would simply choose a different dataset. I don't believe that the hypothesis that patient demographic data relates to medical care is inaccurate. Here there was no clear relation to readmission. But there are other outcomes that could have more of a relation, such as proper diagnosis of diabetes, that other datasets would be centered around. Or even for readmission, finding other datasets rather than the one used in the project could prove the hypothesis right.