Progress Report 2

Tyler Fusco, Gabrielle Berasi, Mari Leonard

Current Status:

Completed Analysis: We finalized the Exploratory Data Analysis (EDA) phase, which involved generating visualizations to gain insights into the data. This included a boxplot of the 'thalach' variable, a histogram of cholesterol distribution, a correlation matrix, and a scatter plot of 'old peak' versus 'slope'.

Modeling Progress: We have since executed initial runs (two iterations each) of three distinct models: logistic regression, kNN, and random forest. Analysis of the results indicates that the logistic regression and kNN models provided the most reliable and accurate performance. Conversely, one random forest model was identified as clearly overfitted, evidenced by perfect 100% scores across accuracy, recall, and F1.

Future Plans:

In the near future we are planning on running our models again using PCA to see how it affects our model accuracy. The models we have decided to run are knn, random forest and logistic regression. We will also want to compare our results using PCA to our results without using PCA to show how PCA can affect results.

We will need to complete this to prepare for our next client meeting. We will also need to work on the draft of our poster and the draft of our final presentation once we have our models completed. We also have some class activities such as the Visualization activity and our Data Ethics Debate.

Potential Issues:

While our project has progressed smoothly so far, there are several potential issues that we have identified moving forward. One concern is that our initial model runs were limited to only two iterations per model, which may not be sufficient to produce stable or generalizable results. Additionally, one of our random forest models appeared to be overfitted, achieving perfect accuracy, recall, and F1 scores. This suggests that our dataset may need more careful tuning, parameter adjustment, or additional validation to prevent overfitting. Another issue is that we have not yet clearly defined or documented our data-splitting and validation methods, such as

whether we are using cross validation or a fixed train test split, which could impact the reliability of our model comparisons.

We also recognize that the implementation of PCA in the next phase may present challenges. It will be important to decide how many components to retain and to ensure that PCA is integrated correctly into the modeling pipeline for fair comparison across models. Lastly, as we move closer to preparing our poster and final presentation, time management may become a challenge as we balance this project with other course activities such as the Visualization activity and Data Ethics Debate. Addressing these potential issues early will help ensure that our final results are accurate, interpretable, and ready for presentation.