Reflection

Brendan Callender, Jacob Perez, Rachel Roggenkemper

Goals

What is the main takeaway message you hope to be able to supply to the partner?

We hope to provide the American College of Cardiology with a predictive tool that improves diagnostic consistency for diabetes and cardiovascular disease while ensuring equitable outcomes for male and female patients. The main takeaway is that this tool can help clinicians identify high-risk patients more accurately and consistently, enabling earlier interventions and better patient outcomes.

What types of summaries or visualizations do you think would be impactful to communicate your results?

Key visualizations will include ROC-AUC curves to demonstrate model performance, bar charts highlighting variable importance, and confusion matrices to show the model's accuracy, precision, and recall. Additionally, comparative graphs illustrating outcomes for male and female patients will emphasize the tool's commitment to equitable predictions.

What are some conclusions that you would be excited to discover? What results might be a bit disappointing to you?

We would be excited to identify specific predictors that strongly influence risk for both diabetes and cardiovascular diseases, particularly if they are actionable or previously overlooked factors. It would also be exciting if our model achieves high sensitivity and ROC-AUC across diverse patient demographics. Additionally, we are aware that symptoms of heart disease can differ significantly between men and women, with symptoms experienced by women often being overlooked or misinterpreted. We hope our findings can bring awareness to these disparities and highlight any gender-specific risk factors that could improve diagnostic accuracy for women. On the other hand, it might be disappointing if the model fails to improve prediction accuracy over existing tools or if certain predictors show significant bias or unequal outcomes across genders.

Challenges

Are there any aspects of the data that you are worried about, or that are challenging to clean / manipulate?

A primary concern is the potential for missing data or inconsistencies across the datasets, particularly given their diverse sources. Cleaning categorical variables with inconsistent levels or significant missing values may be challenging and could affect model performance. Balancing the datasets to address class imbalances, such as fewer patients with cardiovascular disease, will also be a critical step.

Is there any information that is not in your dataset that you wish you could have?

We think it would be helpful to have longitudinal data that tracks patients over time, as this could improve the predictive power of the model by capturing trends in health changes. Additionally, biomarkers or other lab results that are not included in some datasets could enhance the precision of predictions.

Are there any other concerns you have regarding the success of your project?

One concern that we have to keep in mind is we must be careful to only use patient information that is ethically and legally allowed, as health data is highly sensitive and protected by strict regulations. Ensuring proper data handling is essential to maintain privacy and comply with legal and ethical standards.