Project Proposal

Rachel Roggenkemper, Jacob Perez, and Brendan Callender

I. Introduction

Your organization, the American College of Cardiology (ACC), is dedicated to improving cardiovascular health by providing healthcare professionals with cutting-edge tools and resources to enhance patient care. Recognizing that there is a crucial connection between diabetes and cardiovascular disease, we aim to support your mission by developing a predictive algorithm to help doctors assess patient risk for diabetes and cardiovascular disease. In particular, we aim to develop an algorithm which accurately predicts patient risk for both males and females to create more equitable patient outcomes. Additionally, our algorithm will address the variability in diagnoses across doctors by offering a data-driven approach that can be consistently used across practices to make patient diagnoses more consistent, enabling earlier interventions and improved outcomes.

The strong association between diabetes and cardiovascular diseases emphasizes the need for an integrated approach to risk assessment. Leveraging data from a multispecialty hospital in India, the CDC's Diabetes Health Indicators dataset, and a diabetes-focused study from Sylhet Diabetes Hospital in Bangladesh, we will analyze key predictors of patient risk within these disease states. Our algorithm will emphasize the use of clinically accessible variables, ensuring practicality and ease of use in healthcare settings. This will empower doctors to make informed, consistent decisions, aligning with the ACC's goals of advancing medical practice through innovation.

We will provide your organization with a comprehensive predictive tool that enhances diagnostic consistency and reduces false negatives, or cases where patients are not treated when they should be, directly benefiting patients at risk. By facilitating earlier detection of diabetes and cardiovascular diseases, our project aligns with ACC's mission to transform cardiovascular care and improve heart health. Together, we can ensure that healthcare providers are better equipped to address these interconnected conditions, ultimately saving lives and enhancing the quality of care.

II. Previous Work

1. Machine learning in precision diabetes care and cardiovascular risk prediction

This article explores the application of machine learning algorithms, including random forests, gradient boosting machines, and neural networks, to personalize diabetes management and predict cardiovascular risk. The study emphasizes the potential for advanced algorithms to improve diagnostic accuracy and guide treatment decisions. The authors evaluate model performance using metrics such as accuracy, precision, recall, and F1 score to assess the effectiveness of different models. The authors also discuss the regulatory framework surrounding clinical decision support (CDS) tools and how they require rigorous testing before use which our models would also require. Similar to our project, the authors seek to improve patient outcomes using data-driven methods to better predict diabetes and cardiovascular diseases. One distinction is that our project emphasizes equal model effectiveness for both males and females which will require thorough exploratory analysis and testing.

2. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning

This article evaluates several machine learning models, including logistic regression, decision trees, random forests, and support vector machines, to predict the risk of diabetes and cardiovascular disease. The study uses metrics such as accuracy and ROC-AUC, highlighting ROC-AUC's value in maintaining consistent performance when adjusting classification thresholds. The authors demonstrate that including lab variables significantly enhances model accuracy, while models relying only on non-lab variables remain viable for broader application scenarios. Similar to our project, this study emphasizes the importance of metrics like ROC-AUC and uses interpretable methods such as decision trees. Our project will also measure model success using ROC-AUC however, we will also consider the specificity of models which will measure a model's ability to prevent false negatives. Additionally, while this article developed separate models for scenarios with and without lab variables, our project seeks to integrate both types of variables into a single model to ensure broader usability.

3. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques

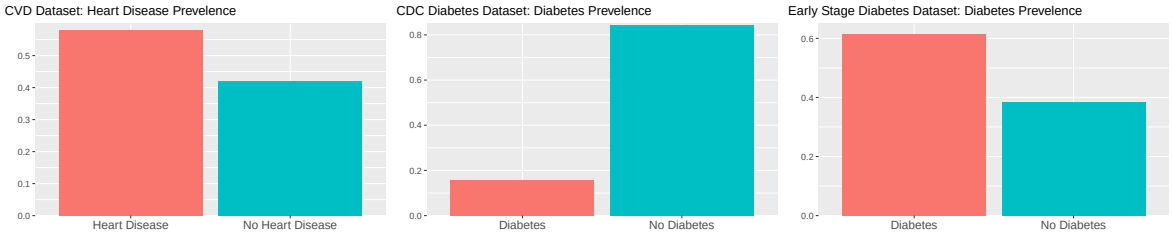
This article outlines a study that uses data from the 2014 Behavioral Risk Factor Surveillance System to build predictive models for Type II diabetes using various machine learning techniques such as support vector machines, decision trees, random forests, neural networks, Gaussian Naive Bayes, and logistic regression. The models were evaluated using metrics such as accuracy, sensitivity, specificity, and ROC-AUC. While the neural network achieved the best overall performance, the decision tree was preferred for its higher sensitivity, making it suitable for initial screenings. In addition to identifying well-known risk factors like age and BMI, the study highlighted new potential risk factors such as oversleeping and frequent health checkups. Unlike this study, which focused exclusively on Type II diabetes, our project aims to develop model predictions for both diabetes and cardiovascular disease for clinical use

by doctors. Additionally, we emphasize ensuring equitable outcomes for males and females, addressing biases which were not a focus of this study.

III. Exploratory Analysis

Distribution of Diagnoses

The grid of plots below shows the distribution of diagnoses across our three datasets. We notice that the cardiovascular disease (CVD) dataset and the early stage diabetes dataset both have somewhat balanced distributions of the response variables. We see that in the CVD dataset has about a 60/40 split with most cases truly having heart disease. The same follows for the early stage diabetes dataset with a 60/40 split with diabetes the majority. However, in the CDC diabetes dataset we notice an imbalance with over 80% of the observations being cases where the patients did not have diabetes. We will have to take into consideration these imbalances when fitting our predictive models.

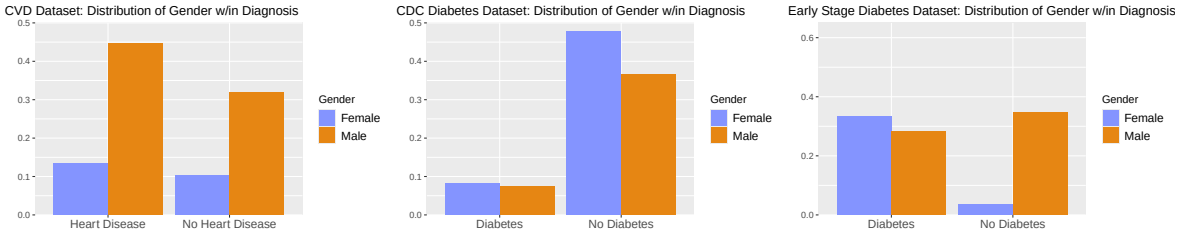


Distribution of Gender within Diagnoses

Because we are concerned with having equality of outcomes across gender, we looked into the distribution of the diagnoses within each gender. The plots below show the proportion within each disease state broken up by gender. For the CVD dataset, we notice the distribution of diagnoses is roughly the same for both males and females both having more cases of heart disease. However, we do see there are much more male cases than females cases for both heart disease and no heart disease cases. For the CDC diabetes dataset, we see that the distribution of diabetes diagnoses is roughly the same with there being more female cases for diabetes and no diabetes diagnoses. Lastly, for the early stage diabetes dataset, we see a very large difference in the number of female cases with no diabetes. This imbalance will likely need to be adjusted for in our model fitting process.

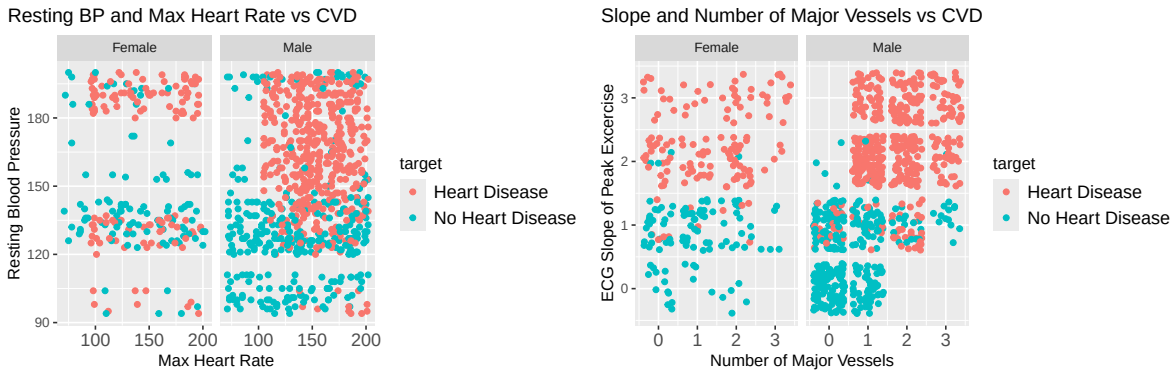
Relationship between Predictors and Diagnoses

In this section we explore some of the relationships between predictors in the dataset to the different diagnoses and also observe if any of the relationships differ across males and females.



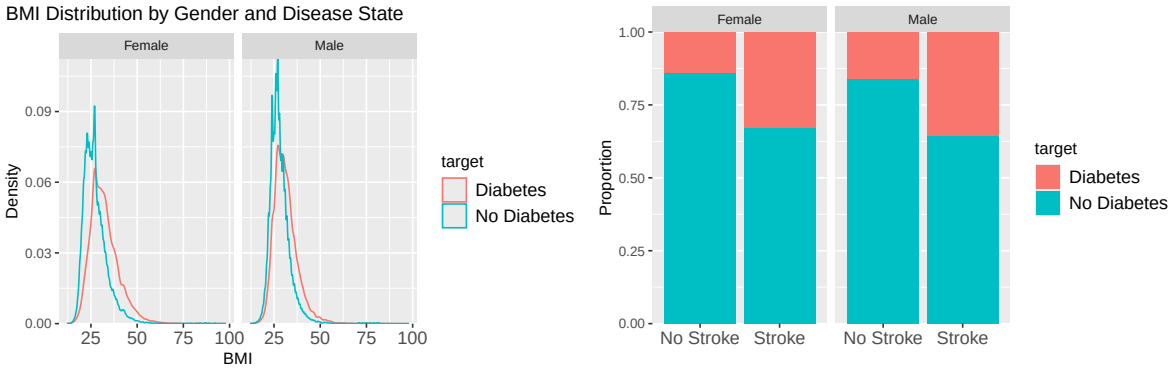
Cardiovascular Disease Dataset

Below are two plots constructed from the CVD dataset showing the relationship of 2 different predictors to heart disease prevalence. These plots are also broken up by gender to observe any potential interaction effects across males and females. For the plot on the left, we observe that higher resting blood pressure is associated with higher prevalence of heart disease. We notice that heart disease becomes more prevalent around a blood pressure of 180 while it becomes more prevalent for men around 150. For the plot on the right we notice that a higher ECG slope of peak exercise value of 2 or higher is associated with higher rates of diabetes for males and females.



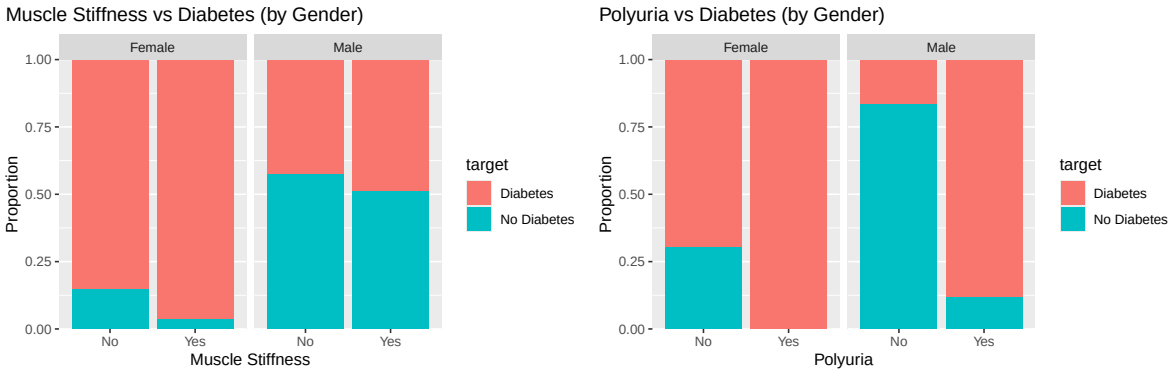
CDC Diabetes Dataset

Below are two plots constructed from the CDC diabetes dataset. These plots are also broken up by gender to observe any potential interaction effects across males and females. For the first plot on the left, we notice that individuals with diabetes tend to have larger BMIs than those without diabetes. This holds true for both males and females. In the second plot, we notice that individuals who had a stroke are more likely to have diabetes than individuals who did not have a stroke. This trend is the same for both males and females. Within this dataset, we noticed most relationships between predictors and the response were the same for males and females. Because the sample size of the CDC dataset is so large, the trends found in our exploratory data analysis are much more reliable than our other datasets which could happen by chance with smaller sample size.



Early Stage Diabetes Dataset

Below are two plots broken up by gender constructed from the early stage diabetes dataset. From the first plot on the left, we notice that individuals with diabetes are associated with having slightly more muscle stiffness than individuals without diabetes. This is true for both males and females. In the second plot, we see that individuals who have diabetes are more likely to be experiencing polyuria (having to pee a lot) than individuals without diabetes. This association is true for both males and females. One important observation in these plots is the majority cases for women being cases of diabetes. This majority, which was explored earlier, shows up in both plots and may cause us to believe a variable is predictive overall when it really isn't and it is just a trend unique to our dataset.



IV. Preliminary Results

To get a better sense of the problem, we fit initial models using logistic regression with LASSO penalties and calculated cross-validated metrics which are recorded in the tables below.

Cardiovascular Disease Dataset

For the cardiovascular disease dataset, our preliminary model performed extremely well with respect to accuracy, sensitivity, and ROC AUC. Our next steps for this model will be testing other models such as support vector machines, k-nearest neighbors, Linear discriminant analysis, and more. Additionally, we will also be evaluation model results with respect to gender treated as binary which may serve as a tie-breaker between two effective models.

Metric	Value	Interpretation
Accuracy	0.958	Our preliminary model predicted roughly 96% of cases correctly as either heart disease or no heart disease.
Sensitivity	0.961	Our preliminary model predicted 96% of cases where individuals truly had heard disease correctly.
ROC AUC	0.993	ROC AUC is hard to interpret. However we would expect a value of 0.5 for random guessing and 1 as perfect so 0.99 is extremely good.

CDC Diabetes Dataset

For the CDC diabetes dataset, our preliminary model performed very poorly. We see the sensitivity with respect to diabetes diagnoses is very poor while the accuracy and ROC AUC values are somewhat good. This is due to the high prevelance of non-diabetes diagnoses in the data which dominate the predictions. Our next steps will involve testing more models to improve the sensitivity of our final model to better catch diabetes cases. We will also be evaluate these models with respect to equal outcomes for males and females to approach a final model.

Metric	Value	Interpretation
Accuracy	0.849	Our preliminary model predicted roughly 85% of cases correctly as either diabetes or no diabetes.
Sensitivity	0.186	Our preliminary model predicted 18.6% of cases where individuals truly had diabetes correctly.
ROC AUC	0.819	ROC AUC is hard to interpret. However we would expect a value of 0.5 for random guessing so 0.819 is an improvement.

Early Stage Diabetes Dataset

For the early stage diabetes dateset, our preliminary model performed very well with respect to accuracy, sensitivity, and ROC AUC. We see the sensitivity of the preliminary model was very good which aligns with our initial project goals. Our next steps for this dataset will include testing more models and evalutating these models overall and with respect to performance for males and females.

Metric	Value	Interpretation
Accuracy	0.856	Our preliminary model predicted roughly 86% of cases correctly as either diabetes or no diabetes.
Sensitivity	0.946	Our preliminary model predicted 94.6% of cases where individuals truly had diabetes correctly.
ROC AUC	0.955	ROC AUC is hard to interpret. However we would expect a value of 0.5 for random guessing and 1 as perfect so 0.95 is extremely good.

V. Project Timeline and Goals

Please note: This document is the project proposal

Date	Deliverable	Description
11/22/24	Project Proposal	Document describing initial project goals, discussion of initial dataset exploration, and preliminary model fitting results.
12/6/24	Final Report	Final report including discussion of final results, model use instructions, limitations of project, and next steps.
12/6/24	Final Models	Final predictive models
12/6/24	Presentation	Presentation of final results (as poster). Discussion of key findings, metrics, and limitations of project.