Group 1 Project Proposal: Rohan Girish, Anoushka Jadhav, Shweta Kumaran, Landon Myhill

For our final project, we plan to use a heart disease dataset from Kaggle to build a model that will predict whether someone is at risk of heart disease based on different health variables. The dataset includes variables such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, EKG results, max heart rate, exercise-induced angina, and ST depression. The main goal of this model is to figure out which of these variables are most important in predicting whether someone will get heart disease and to see the effectiveness of different machine learning models at identifying high-risk patients. This will be beneficial in understanding key health metrics to help identify the risk of heart disease.

We'll start by cleaning the data by checking for missing values, outliers, and how the different features relate to one another. We might need to use SQL in order to modify the tables to a format that is suitable to the type of analysis we want to do. Next, we will encode any categorical variables and normalize numerical data. In order to effectively summarize the data, we will visualize any relationships using bar plots, histograms, and correlation matrices. For the analysis, we plan to try out models like logistic regression, decision trees, and random forests. We will use resampling techniques like k-fold cross-validation and bootstrapping to make sure the results are consistent and reproducible. We'll compare the models using metrics like accuracy, precision, recall, and AUC to figure out which one works best.

To help explain the results, we will include visualizations like ROC curves, confusion matrices, and feature importance plots. The goal is to make the findings easy to understand and helpful for something like early detection or prevention. Everything will be pulled together in a clean and organized Jupyter notebook with all our code, visuals, and commentary. We'll also upload everything to GitHub and build a Docker image to make sure the project is reproducible and easy to run. Each group member will take on a different part of the project, like one of the different machine learning techniques, results and visualization, and formatting of the presentation. Additionally, we will attempt to meet in person to work on the project and assist one another on any issues we may be facing. Overall, we want to show a full data analysis workflow from start to finish using a real-world health problem that actually matters.

https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction