1 Title: Heart Disease Prediction: Machine Learning for Good

2 Summary of the Plan

2.1 Motivation and Background

According to the Centers for Disease Control and Prevention, heart disease is the leading cause of death for people of most ethnicities in the United States. About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths. Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack. In this project, we plan to use data mining techniques to understand the relevant causes of heart disease and exploit machine learning algorithms to predict it. Better prediction hopefully leads to better prevention of heart disease.

2.2 General approach

In this project, we plan to use the clinical data obtained from the Cleveland, Hungarian, Switzerland,Long Beach VA heart disease database (UCI machine Learning Repository). These datasets segregate the patterns possibly related to the heart disease. We'll first perform data processing such as data cleaning, data integration and deal with missing values. Then we'll conduct statistically analysis of the datasets to better understand the distribution of the data and select relevant features. Some example features are as shown in the following table. Finally we'll implement several machine learning models to compare the prediction accuracy of these models so that we can find the most suitable model with the best performance.

```
age − age in years
sex − (1 = male; 0 = female)
cp − chest pain type
trestbps − resting blood pressure (in mm Hg on admission to the hospital)
chol − serum cholestoral in mg/dl
fbs − (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg − resting electrocardiographic results
thalach − maximum heart rate achieved
exang − exercise induced angina (1 = yes; 0 = no)
oldpeak − ST depression induced by exercise relative to rest
slope − the slope of the peak exercise ST segment
ca − number of major vessels (0−3) colored by flourosopy
thal − 3 = normal; 6 = fixed defect; 7 = reversable defect
target − have disease or not (1=yes, 0=no)
```

Function 1: Some example features

2.3 Suggested experiments

Prior work usually use classification models to conduct disease prediction. In our project, we'll start with experimenting on logistic regression and then we plan to implement several classification models such as KNN, SVM, naive bayes, decision tree and random forest. We are interested in comparing the performance of these different models to find out the best prediction model for heart disease.

3 Group Information

Group Members: Yeqian Yang (yy1420), Xiaoxuan Liu (xl2590)
Member responsible for submissions: For the project, which will contain data cleaning and machine learning model implement. We will equally separate the both parts. So we will have same workload during the project implementation. And Xiaoxuan Liu will upload this project proposal.