1 Title: Heart Disease Prediction: Machine Learning for Good

2 Group Members

Group Members: Yeqian Yang (yy1420), Xiaoxuan Liu (xl2590)
Member responsible for submissions: Xiaoxuan Liu will be responsible for all submissions.

3 Backgroud

3.1 Description of Problem

Medical organizations such as hospitals and clinics generate huge amount of healthcare data during their daily operations. However, unfortunately these data has not been fully exploited to discover the hidden patterns and relationships between patient's medical profiles such as age, gender and blood pressure etc. and the risk of heart disease.

Weng et al.[1] pointed out that current approaches to predict cardiovascular risk are mostly ineffective in identifying people who are highly likely to suffer from heart disease or people who receive unnecessary intervention. Some researchers have made use of various machine learning models to aid physicians in early diagnosis of heart disease. Gonsalves, Amanda H., et al.[2] used three supervised learning techniques Naive Bayes, Support Vector Machine and Decision Trees to experiment on the South African Heart Disease dataser of 462 instances. Dinesh, Kumar G., et al.[3] compared the prediction performances of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region.

In this project, our main contributions would be 1) we are going to perform extensive experiments on various machine learning models to predict heart disease based on the patient's medical profiles and compare their prediction accuracy. 2) Since the datasets previous papers used are quite small usually with only 300 to 400 instances, we are going to solve this issue by combining all the available datasets with same feature dimensions to provide a larger dataset and also conduct parameter tuning using K-fold cross-validation.[4]

3.2 Motivation for Problem

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.[5] 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.

3.3 References

[1] Weng, Stephen F., et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?." PloS one 12.4 (2017): e0174944.

[2] Gonsalves, Amanda H., et al. "Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis." Proceedings of the 2019 3rd International Conference on Deep Learning Technologies. ACM, 2019.

[3] Dinesh, Kumar G., et al. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms." 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[5] https://www.cdc.gov/heartdisease/facts.htm

[6] https://archive.ics.uci.edu/ml/index.php

4 Plan

4.1 Description of Methodology

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).[6] 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 statistical analysis of the datasets to better understand the distribution of the data and select relevant features. 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.

4.2 Proposed Experiments

We'll firstly preprocess the datasets. We'll combine the four datasets collected from different regions. Since most missing values indicate that the patient did not take that specific test, we'll drop these records. If the proportion of missing values represents greater than 50% of that feature, we'll drop the feature.

Prior work usually use classification models to conduct disease prediction. In our project, we 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 going to build the Classifiers using the combined datasets and conduct parameter tuning using K-fold cross-validation. We are interested in finding the best parameter setting for different models, comparing the performance of these different models to find out the best prediction model for heart disease.

4.3 Some Relevant Datasets

The raw datasets we use are from UCI Machine Learning Repository, and the heart disease datasets are separated into several sub sets from different region's hospitals(cleveland.data, hungarian.data, switzerland.data and long-beach-va.data, heart-disease.names). Here are the features relevant to our project:

```
        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: Dataset description