**Classification of Heart Disease**

JIAYING YUAN, ATUL ZACHARIAS, RUIZHE ZHANG, SAM ZHANG, LING ZHONG

Cardiovascular diseases (CVDs) are the number one cause of global deaths, which leads to the importance of study about its causes and risk. Although the CVDs may have various types and causes, heart failure is one of the most common consequences and leads to millions of deaths every year all over the world.

As it is such a serious disease, we are interested in building machine learning models based on the Heart Failure Prediction Dataset to classify the condition of heart disease. We will use a dataset containing 11 features that can be used to predict a possible heart disease, and one more attribute which is the true heart disease condition. The dataset was created by combining different datasets already available independently but not combined before, which can be found under the Index of heart disease datasets from UCI Machine Learning Repository. It contains the ‘Age’ , ‘Sex’, ‘ChestPainType’, ‘RestingBP’, ‘Cholesterol’, ‘FastingBS’, ‘RestingECG’, ‘MaxHR’, ‘ExerciseAngina’, ‘Oldpeak’, ‘ST\_Slope’, as the potential predictors, and ‘HeartDisease’ as the response variable.

We will split the dataset into a training and testing set at first. Then our target is to train several classification models to predict whether a patient with certain conditions has heart disease or not and determine the best model for the dataset. We will perform the following steps to carry out the classification: data cleaning and transformation; variable selection; model selection(including KNN, LDA and Decision Tree, etc.); validation analysis; discussion and summary.

**Reference**

[1] Heart-Disease Dataset, UCI Machine Learning Repository.

<https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

[2] Fedsoriano, Heart Failure Prediction Dataset.

<https://www.kaggle.com/fedesoriano/heart-failure-prediction>