INT305 2020 Coursework 3

Interim Report

Name : yuxuan wu

Student ID : 1716309

Approval Signature : Kangshi

Date :

Problem description :

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help. With the assistance of the prediction model, we could generate a roughly possibility of getting the diseases in the future and make early treatment

Method :

1. Visualization of each feature. Have an overview of the component or distribution of each feature. Plot a heatmap to find the potential relationship between each feature, which make a preparation for future feature selection procedure.
2. Use logistic regression and svm algorithms to include all the feature and samples to make a preliminary prediction
3. Try to optimize and increase the overall accuracy of the model.
   a) Conduct the feature selection procedure to return the most significant features
   b) Add the normalization and regularization part
   c) Design a proper dropout rate
   d) Use grid search method to return the best hyperparameters
   e) Use rbf kernel method, try to incorporate more intricate svm algorithm

Preliminary Experiment :

I finished the explorative data analysis, including multiple graphs of features. For better illustration of the data, I incorporated histogram, violin plot, bar chart to make basic comparison (in percentage) and figure out the potential relationship between each other.

I also finished the preliminary model building. I split the data into two datasets, 80% as training data while 20% as testing. I exclude the time feature since the time of recording the patients have no relationship with the formation of heart disease. Then, I fed all the features into the prediction model, either logistic regression or svm algorithm was used.

Next steps :

1. Try to optimize and increase the overall accuracy of the model.
   a) Conduct the feature selection procedure to return the most significant features
   b) Add the normalization and regularization part
   c) Design a proper dropout rate
   d) Use grid search method to return the best hyperparameters
   e) Use RBF kernel method, try to incorporate more intricate svm algorithm