Report
Heart Disease Prediction using Logistic Regression

A. Introduction

Heart disease is a major health concern worldwide. If we are able to identify high-risk patients based on risk factors, we will be able to recommend lifestyle changes to these patients in order to lower their risks.

B. Data Overview

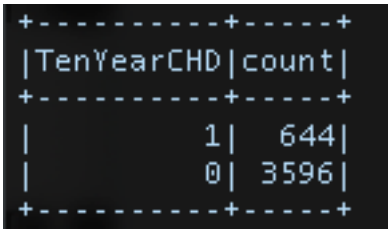
For this problem, the dataset Framingham Heart dataset was used. This dataset contains fifteen features and one target label. The feature columns include: male, age, education, currentSmoker, cigsPerDay, BPMeds,prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, and glucose. The target label is TenYearCHD (predicting whether a patient has a 10-year risk of future risk of coronary heart disease). This dataset contains 4240 instances.The supervised Machine Learning algorithm Logistic Regression within Spark MLlib is used to build the predictive model.

C. Data Pre-processing and Algorithm

Seven out of the fifteen features in this dataset have ‘string’ data type (education, cigsPerDay, BPMeds, totChol, BMI, heartRate, glucose). In order to create a data frame of vectors, these string type columns are converted to ‘double’ data type. After the conversion, 582 rows of data were removed due to missing values. Then, all the features are assembled/vectorized into a feature column. To ensure all the data falls within the same scale, standard scalar normalization was used on the feature. The entire dataset was split into 80% for training and 20% for testing. The training data was then used for training the Logistic Regression model.

D. Results

We obtained a training accuracy of 85.1%. The testing data was then used on the same model to obtain an accuracy of 86.5%. Based on the results, the algorithm is successful in predicting whether someone will get coronary heart disease in 10 years. However, we did observe an imbalance in the class label distribution: there is five times more data on label ‘0’ than label ‘1’. Improvement to this algorithm can be made by collecting more class ‘1’ data or by oversampling the class ‘1’ data.



TenYearCHD	count
1	644
0	3596