Final-Report

12/13/2021

I. Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVDs are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 32% of all global deaths. Over 85% of deaths from CVD were due to heart attack and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Cardiovascular diseases (CVDs) are the number one cause of death globally. Currently, there are several different ways for physicians to diagnose patients that they believe to be at risk for Cardiovascular Diseases (CVDs). The practices vary by country, but often include the physician checking the patient's blood pressure, cholesterol level, and conducting further tests such as exercise stress tests, X-rays, etc. Currently, there are many issues with the current diagnostic methods. A [study] (https://www.sciencedaily.com/releases/2009/11/091116103435.htm) of 500 patients found a false positive reading between 77 and 82 percent in patients in patients at risk of CVD screened by ECG, and a false negative reading between 6 to 7 percent in the same patient population . People with CVDs or who are at high risk of CVDs need early detection and management wherein a machine learning model can be of great help. Using our Cardiovascular Heart Disease data, we have two main goals. Our first goal is to create models for the purpose of prediction. These can be used to assess the likelihood of a heart disease diagnostic for potential at-risk patients based on a number of factors such as age of the and sex of the patient, blood pressure, cholesterol, heart rate, and the presence of chest pain. Our second goal is to create models for the purpose of interpretation, which can be used to provide a greater understanding of signs that at-risk patients can analyze to check their risk for CVDs. We chose to fit 3 different models to classify whether a patient has heart disease or not. The first model is a logistic regression model with variable selection performed by a lasso regression. We decided to use this model because of its interpretability so that we can examine the relationship our predictors have with the probability of a patient having heart disease. We also decided to use a random forest and a SVM because of their ability to perform classification and due to their predictive power despite their lack of interpretability. We will use a 10-fold CV to determine the best predictive model based on the classification error of each model. Due to the ability of ensemble models to reduce prediction error, we will also create an ensemble model where the prediction is the most common result of the 3 individual models (https://www.sciencedirect.com/topics/computer-science/ensemble-modeling). Finally, we compare the 10-fold CV errors for each individual model as well as the ensemble to find the one with lowest classification error and therefore highest predictive accuracy.

II. Data

We obtained our data from Kaggle (https://www.kaggle.com/fedesoriano/heart-failure-prediction). The dataset was originally provided by Dr. David Aha, a researcher at the US Naval Research Laboratory. It was created by combining five heart datasets from Cleveland (303 observations), Hungary (294 observations), Switzerland (123 observations), Long Beach, Virginia (200 observations), and Stalog (270 observations), for 918 unique observations. This makes it one of the largest available heart disease datasets. The dataset has 11 predictor variables, which are age, sex, chest pain types, resting BP, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, oldpeak, and slope of the peak exercise ST segment. Sex, chest pain types, resting ECG, exercise-induced angina, and ST slope are categorical variables, while the rest are numeric. We made sure to factor the categorical variables in order to make sure R would treat them as categorical. For our purposes, the dataset has one response variable, which is whether or not the patient has heart disease. Because in the dataset the frequency of heart disease is roughly equal to the frequency of non-heart disease, the data is likely not based on the general population, where we would expect

the frequency of heart disease to be a lot lower. As a result, the inference and predictions from our models do not apply to the general population, but instead apply to the population this dataset was drawn from, which is patients who are at risk of heart disease.

There were a total of 170 observations with missing values in Cholesterol and Resting Blood Pressure(BP). (Cholesterol value of 0 and Resting Blood Pressure value of 0 are considered missing since these are impossible to reach in real life). The visualizations of the missing values in [Table 00] shows that there are 170 observations with missing Cholesterol value and 1 observation with missing Resting BP value. For the one observation with missing Resting BP, it also had missing Cholesterol value. This observation was disregarded for analysis under Missing Completely at Random (MCAR) assumption. For the 169 observations with missing Cholesterol values, imputations were conducted based on Missing at Random (MAR) assumption. Multiple Imputation using Chained Equation (MICE) in R was used to impute the values. Five chains of imputations were conducted for Cholesterol with each chain using the default method of predictive mean matching (ppm) method since Cholesterol was a numerical variable.

Variables	Description	Type
Age	Age of the patient in years	
Sex	Sex of the patient [M/F]	
ChestPainType	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	
RestingBP	Resting blood pressure [mm Hg]	
Cholesterol	Serum cholesterol [mm/dl]	
FastingBS	Fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	
RestingECG	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]	
MaxHR	Maximum heart rate achieved [Numeric value between 60 and 202]	
ExerciseAngina	Exercise-induced angina [Y: Yes, N: No]	
Oldpeak	Oldpeak [Numeric value measured in depression]	
HeartDisease	Output class [1: heart disease, 0: Normal]	

III. Methodology

IV. Results

V. Conclusion