Heart Disease Prediction with Machine Learning

Aritzi Piedras Silva
DSC 680
Milestone 3 – Project Draft
April 5,2021

\

Acknowledgements

In the growing field of machine/deep learning, I would like to acknowledge the Bellevue Professors that have provided many examples and rationale towards building, evaluating, and examining different models. Something that stood out to me while being progressing in this course is how there is not a consistent framework for why a model is selected – However, it is important to follow the many different disciplines that would need to be used for research plans that enhance its credibility, validity, and replicability.

Abstract

Heart disease continues to be one of the leading causes of death for Americans each year despite of race and gender. According to the Center of Disease Control and Prevention, forty percent of premature deaths related to heart disease can be prevented by modifying different risk factors such as smoking, physical inactivity, unhealthy diets, and stress levels. Unfortunately, medical errors and undesirable results are the reason for the need for computer-based diagnosis to achieve high quality medical procedure results. Machine learning in Medical health care is an emerging field of extreme importance that may provide diagnosis and prognosis. The main goal regarding the use of machine learning in the medical field, specifically in cardiovascular disease is to assess and summarize the overall performance in prediction ability of different algorithms.

Keywords: Predictive Analytics, Machine Learning, Medical Health Field, Logistic Classification Regression.

Problem Statement/Hypothesis

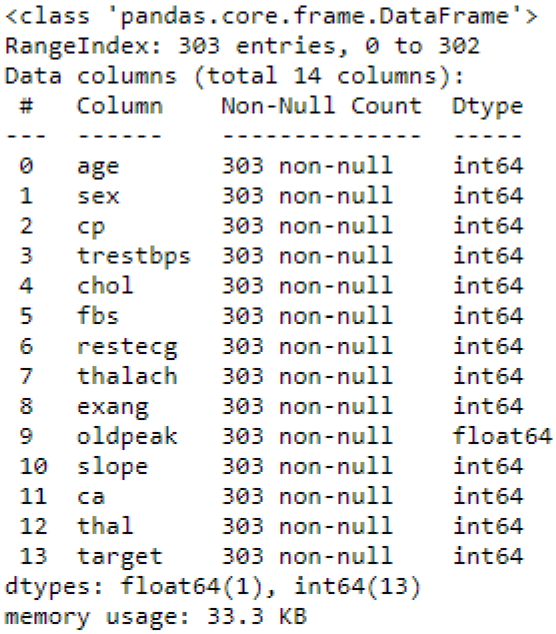
Heart disease, also known as cardiovascular disease, is a condition that affects the heart and has been the United States number one killer. According to Robin Donovan's article in *Healthline*, about 610,000 American die from cardiovascular disease that includes several different races such as Hispanics, Black people, and White people [6]. As previously mentioned, heart disease is deadly but its also extremely preventable in most people by adopting healthier lifestyle options that can allow the average people to potentially live longer. However, there are some risk factors that are impossible to control such as family history, ethnicity, sex, and age. Currently, diagnosis for heart disease consists of various test such as personal and family medical history, previous and past symptoms, blood tests, and an electrocardiogram.

This project will be focused as a classification problem in predicting heart disease on different individuals with unique attributes. The machine learning algorithms that are going to be compared in performance of accuracy in this project will be Logistic Regression and Decision tree. These two algorithms will allow us to gain better possible outcomes that may help predict early signs of heart disease and prevent future deaths based on the 13 possible attributes.

Data source

The data source is a CSV file that was located on Kaggle. The data source consists of 14 columns and 303 entries from the Cleveland database related to heart disease (See figure 1). This project will offer validity and replicability with the coding that has taken place to identify early signs of heart disease.

Figure 1:



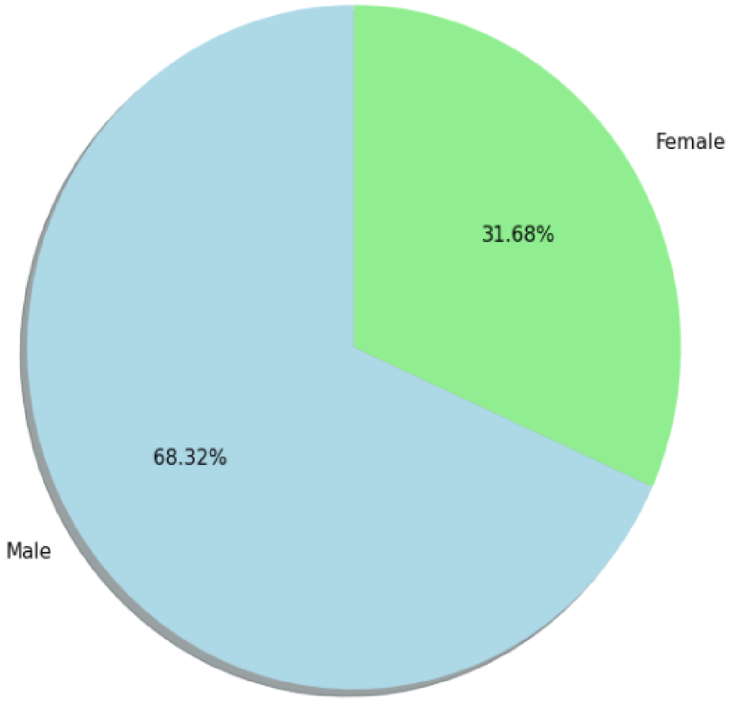
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Technical approach

The technical approach that will be used will be part of the CRISP-DM process. With the CRISP-DM, I will focus on business understanding, data preparation, modeling evaluation and development. Each stage of this process will focus on framing up a research plan and help structure and organize the final project.

The data will be preprocessed for cleaning, I will be changing the column names so they are easier to read for an audience that may not be familiar with the abbreviations. The data has been cleaned by dropping null values (in which this dataset did not), convert types of data from integers to categorical format when necessary, check for unique values, and the basic statistical details like percentile, mean, standard deviation etc. After the data clean up, I began to play around with exploratory data analysis to see what visualizations I can create to identify potentially hidden patterns and/or outliers. The packages used to create visualization were Matplotlib and Seaborn – however, I decided to play around with Pandas-Profiling to see what the automated output could potentially be. The first thing I wanted to check out with visualization was how gender is affected by heart disease, what percentage of women compared to what percentage of males have been diagnosed with heart disease (See figure 2).

Figure 2:



At this point the data is ready for modeling in which I created a pipeline for machine learning algorithms to train and test the data using sklearn library. The X variable became the ‘Target’ which dictates whether a data point entry was diagnosed with heart disease, while the Y variable became the thirteen other features. I played around with two models so far being Logistic Regression and Decision Trees, I plan on adding a third model to see how it may differ such as K-Nearest neighbor or Random Forest. For Logistic Regression I achieved an 85% in accuracy performance while Decision trees performed at a 78% accuracy. I have hit a roadblock with Grid Search as I cannot get passed an error, but I am hoping that Grid Search and Cross validation will provide me with better parameters to improve the accuracy percentage for all models used.

Conclusion

Heart disease is the leading cause of death for both men and women of different ethnic groups in the United States. However, it is important to note that society strive to find better ways of preventing this stoppable killer. With the use of machine learning and predictive analytics, we could find computer-based options that will allow society to decrease the complication and/or death rate caused by heart disease. This project will allow different researchers to find better ways to prevent heart disease complications based on the attributes provided.

References

[1] Bhanot, K. February 12, 2019. Predicting presence of Heart Diseases using Machine Learning. Retrieved: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c

[2] CDC. Heart Disease Facts. Retrieved: https://www.cdc.gov/heartdisease/facts.htm

[3] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon [PDF-494K]. Retrieved: https://www.cdc.gov/nchs/data/databriefs/db103.pdf

[4] Lloyd-Jones, D. M., Larson, M. G., Beiser, A., & Levy, D. (1999). Lifetime risk of developing coronary heart disease. The Lancet, 353(9147), 89-92.

[5] Lloyd-Jones, D. M., Larson, M. G., Beiser, A., & Levy, D. (1999). Lifetime risk of developing coronary heart disease. The Lancet, 353(9147), 89-92.

[6] Medline Plus. How To prevent heart disease. Retrieved: https://medlineplus.gov/howtopreventheartdisease.html

[7] Rawat, Shubhankar.August 10, 2019. Heart Disease Prediction. Retrieved: https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

[8] Rich-Edwards, J. W., Manson, J. E., Hennekens, C. H., & Buring, J. E. (1995). The primary prevention of coronary heart disease in women. New England Journal of Medicine, 332(26), 1758-1766.

[9] Sanchis-Gomar, F., Perez-Quilis, C., Leischik, R., & Lucia, A. (2016). Epidemiology of coronary heart disease and acute coronary syndrome. Annals of translational medicine, 4(13).

Questions

What are the common signs of heart disease in both genders?

Which gender is most likely to have suffer from heart disease?

What lifestyle changes could people follow in order to prevent heart disease at an earlier age?

How does chest pain play a role in diagnosing the earlier signs of heart disease (while active/resting)?

At what age does blood sugar level become more relevant in prevention of heart disease?

Does high cholesterol become a sign of heart disease?

Do unhealthy eating habits play a role in heart disease?

What is the average age in which both genders begin to experience heart disease?