

Bellevue University

DSC 680

Winter 2021

**Project One:**

**Trends in Cardiovascular Health Related to Heart Disease**

January 8, 2021

Jennifer Ruiz

**Abstract**

Heart attack and stroke events are a leading cause of death across the United States. Though these events have been studied extensively during the last half century, questions remain regarding the factors that influence these events. Work in this domain has largely focused on heart defects and certain underlying risk factors, such as cholesterol levels and high blood pressure.

In this project, the author chose to focus on other possible factors influencing heart disease and heart attack events. Using machine learning methods, the author seeks to determine underlying risk factors that predict heart disease and heart attack events. Data visualization methods will be used to explore the relationships between factors and identify potential areas for future research.

*Keywords:* Random Forest, Heart Disease, Machine Learning, Python

**Introduction**

The prevalence of cardiovascular disease has grown in the United States over the last century. In the interest of public health, researchers have long studied its causes. According to the CDC (2020), heart disease is the leading cause of death among both men and women, claiming one in four lives. Research has uncovered common risk factors that increase one’s risk of heart attack or prolonged cardiovascular disease. The most studied of these factors are high blood pressure, high cholesterol, and smoking (Texas Heart Institute, n.d.).

In the United States, the Cleveland Clinic has served as a nerve center for cardiovascular research around the world. The clinic collects extensive data from its patients regarding heart-related issues and the prevalence of cardiovascular disease. Using the Cleveland data, the author seeks to focus on identifying trends and underlying causes of cardiovascular disease.

**Research Question**

This project endeavors to analyze the impact of multiple factors on cardiovascular disease for understanding and to predict the disease process. To achieve this end, a comprehensive statistical and visualization analysis will be conducted.

Key Research Questions include:

1. What impact does fasting blood sugar have on the instance of heart disease?
2. Do exercise induced cardiovascular symptoms (angina, palpitations, arrythmia) predict myocardial infarction?
3. Does decreased valve activity or defect predict heart disease?
4. What factors are the greatest indication of heart attack or stroke event? What factors have the least impact in predicting heart disease?

**Background Information**

**Methods**

***Data Understanding***

When approaching this project, the author began by examining the individual attributes contained in the dataset. The source data is coded categorically with several levels for some attributes. Attributes were visualized during Exploratory Analysis using Python (see appendix one) to gain insight into relationships within the data.

***Data Preparation***

Significant data preparation took place in order to accommodate the two methods purposed for this project. Two separate versions of the dataset were prepared prior to analysis. To prepare for the logistic regression, all data was transformed from numeric to categorical with one or more levels, as needed. This helped simplify exploratory analysis and ease of understanding. The base dataset was used for the random forest classifier.

***Modeling***

Data preparation allowed for a random forest classification model to be performed on the data. In additional to the random forest classification model, a logistic regression model was also built to predict cardiovascular disease based on several features within the dataset.

The author chose to a utilize random forest model to fully explore the story within the data. The model utilized all features within the dataset with the target variable being the presence of cardiovascular disease. The model was created using the sklearn package.

**Results**

Exploratory analysis and the models created produced significant insights into cardiovascular disease. Exploratory analysis revealed that heart disease is more prevalent among women than men in the dataset. It also revealed no significant difference in blood sugar levels based on the presence of heart disease. The data also revealed that a fixed defect, angina, and non-typical chest pain were also more prevalent in those with cardiovascular disease. Two final areas of interest were down-sloping heart rate on stress test and a higher maximum heart rate.

The logistic regression model was run twice. The first test produced an accuracy rate of 0.82 with 29 + 21 correct predictions and 5 + 6 incorrect predictions. The second test produced an accuracy rate of 0.94 with 13 + 16 correct predictions and 0 + 2 incorrect predictions.

The random forest classifier did not perform as predicted during the analysis. After multiple tries, the classifier continually produced a prediction of 1.0, indicating perfect prediction of the test data.

**Discussion**

The analysis produced some surprising results and challenges during the analysis were encountered. The performance of the random forest classifier model is questionable. This portion of the analysis will be repeated in future projects and additional options explored. The logistic regression models produced increased accuracy with each run. This indicates better prediction as the model learns the data. Some surprising insights gained during analysis were that resting heart rate, fasting blood sugar, and resting echocardiograms were not significant predictors of heart disease. Blood pressure also showed no tendency to effect rates of cardiovascular disease.

**Conclusion**

As cardiovascular disease continues to be a leading cause of death in the United States and around the world, it is important to thoroughly understand all factors that contribute to the presence of the disease. This analysis has shown several under-explored factors that predict cardiovascular disease. Applying this insight to prevention and patient education can create better outcomes for those with higher risk factors.

**References**

CDC. (2020). Heart Disease Facts. Retrieved on December 5, 2020 from <https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20disease%20is%20the%20leading,1%20in%20every%204%20deaths>.

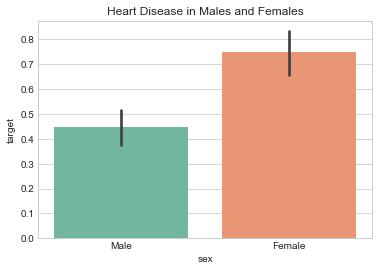
Kaggle. (n.d.). Heart Disease UCI. Retrieved on December 5, 2020 from <https://www.kaggle.com/ronitf/heart-disease-uci>

Texas Heart Institute. (n.d.). Heart Disease Risk Factors. Retrieved on December 5, 2020 from <https://www.texasheart.org/heart-health/heart-information-center/topics/heart-disease-risk-factors/>

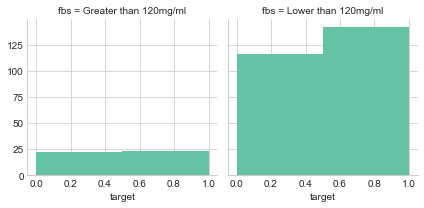
**Appendix A**

**Visuals one through thirteen are feature examination during the exploratory analysis process.**

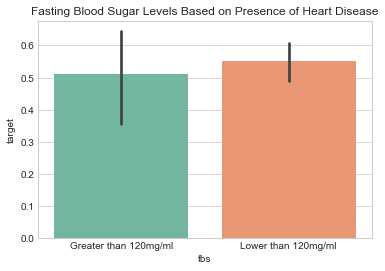
***Table One***

****

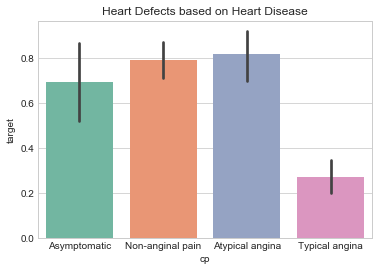
***Table Two***

****

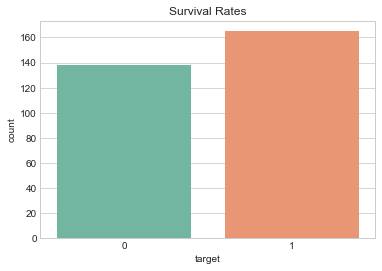
***Table Three***

****

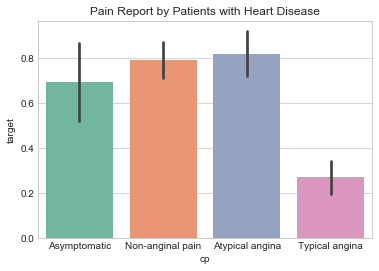
***Table Four***

****

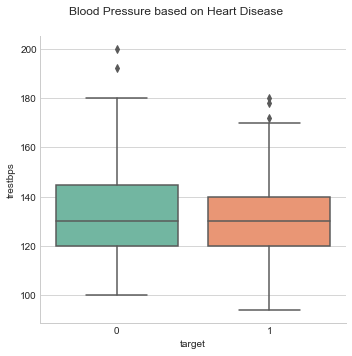
***Table Five***

****

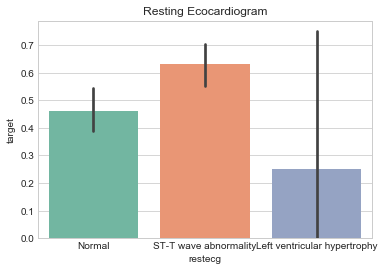
***Table Six***

****

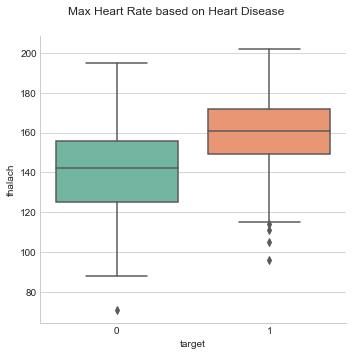
***Table Seven***

****

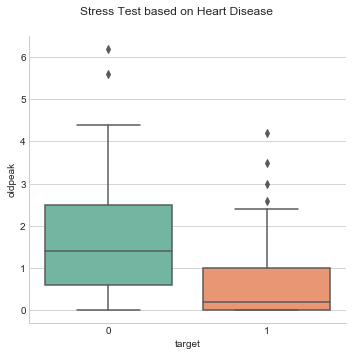
***Table Eight***

******

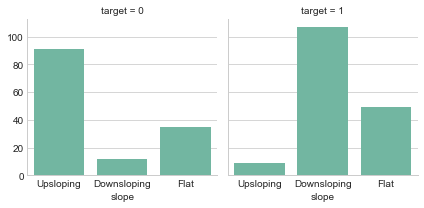
***Table Nine***

******

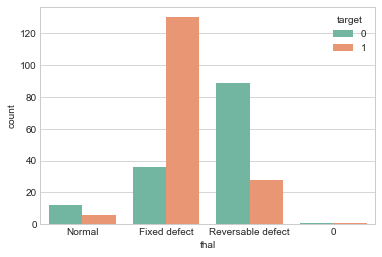
***Table Ten***

******

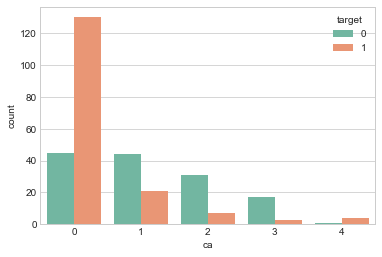
***Table Eleven***

******

***Table Twelve***

******

***Table Thirteen***

******