Trends in Cardiovascular Events related to Heart Disease

# **Jennifer Ruiz Winter 2020 GitHub Portfolio:** <https://github.com/jfruiz15/DataSciencePortfolio>

# Which Domain?

This project comes from the healthcare domain. Information in this domain is used to directly impact healthcare research and treatment for patients. Please see reference list below for annotated information.

# Which Data?

I will be analyzing the University of Cleveland’s Heart Health data set. The data is available to the public through Kaggle through the following link: <https://www.kaggle.com/ronitf/heart-disease-uci>.

# Research Questions? Benefits? Why analyze these data?

For the purpose of this project, I will be predicting cardiovascular events by the presence of heart disease. I will use prediction and classification analyses to better understand the data. My research questions are:

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?

I chose to approach the data with these research questions as they have not been investigated in past analysis. The primary factors looked at in previous analyses were age, sex, and family history as well as cholesterol and blood pressure levels. I believe analyzing these factors could provide new insight which would be beneficial to the overall understanding and treatment of these events.

# What Method?

For this project my plan is to use a random forest analysis to better understand the factors that most influence the presence of heart disease. The dataset is coded categorically so using the random forest analysis will best serve my research questions. I will use multivariate regression models to predict myocardial infarction and stroke events. The goal here is understand which under-utilized factors are the best predictors of heart disease and cardiovascular events.

# Potential Issues?

This dataset is complex with a large amount of data. Though it is already coded with dummy variables and has been cleaned, there may be missing information that needs to be addressed. Also, with the size of the dataset, training the regression model may take considerable time. I will need to prepare for this to ensure delivery on time.

# Concluding Remarks

Head disease and cardiovascular events are a leading cause of death in the United States. Understanding the factors that influence these conditions is key to treatment and prevention. Most research in the area focuses solely four factors: age, sex, family history, and physiological symptoms such as blood pressure and cholesterol. These factors have proven links to both heart disease and cardiovascular events. Yet, there are possible contributing factors that have not yet been explored. In this project, the author seeks to examine factors such as exercise-induced cardiovascular events, resting blood sugar rates, and valve defects in predicting heart disease and cardiovascular events. The author will use both random forest analysis and multivariate regression analysis to achieve this goal.

References:

Li, L. (2019). Classification and Regression Analysis with Decision Trees. Retrieved on December 4, 2020, from <https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054>

This article discusses using Decision tree analysis for both numeric and categorical data. Also discusses basics of random forest implementation for ensemble learning.

Chakure, A. (2019). Random Forest Regression. Retrieved on December 4, 2020 from <https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f>

This article discusses the nuances of ensemble learning. It also provides a more in-depth explanation of random forest algorithms for classification and regression. A sample Python implementation is also provided.

Kansal, H. (2020). Multivariate Linear Regression. Retrieved on December 5, 2020 from <https://medium.com/dev-genius/multivariate-linear-regression-dfd18a26431d>

This article discusses the benefits of using a multivariate linear regression model vs. a multiple regression model with health-related data. The author highlights the complexity of this data domain and the unique challenges to predicting in this category.

Menon, P. (2017). Data Science Simplified Part 5: Multivariate Regression Models. Retrieved on December 5, 2020 from <https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015>

This article outlines how to interpret multivariate regression models as well as the challenges to working with such models. The author highlights common interpretation mistakes and methods to check model accuracy.

Revert, F. (2018). Interpreting Random Forest and other black box models like XGBoost. Retrieved on December 5, 2020 from <https://towardsdatascience.com/interpreting-random-forest-and-other-black-box-models-like-xgboost-80f9cc4a3c38>

This article discusses overall methods for interpreting random forest models. The author makes the important distinction between performance and interpretation. The differences between overall and local interpretation is also outlined.

Grover, P. (2017). Intuitive Interpretation of Random Forest. Retrieved on December 5, 2020 from <https://medium.com/usf-msds/intuitive-interpretation-of-random-forest-2238687cae45>

This article outlines 4 major methods of interpretation for random forest models. The author explains each method as well as its benefit to creating an overall interpretation for models.

Mene, S. (2020). Predicting Heart Disease Using Regression Analysis. Retrieved on December 6, 2020 from <https://medium.com/swlh/predicting-heart-disease-using-regression-analysis-486401cd0a47>

This article outlines regression analysis for predicting heart disease. This article outlines challenges to working with this data type as well as model interpretation. This will be used as a reference in my analysis.

Krisnamoorthy, V. (2018). Your Score Predicts the Chance of a Heart Attack. Retrieved on December 6, 2020 from <https://medium.com/trineomics/your-score-predicts-the-chance-of-a-heart-attack-7a3f4803eb44>

This article outlines background of heart disease and cardiovascular event risk factors. The author outlines the data behind common risk factors such as age, sex, and overall health. It also discusses methods that could be used to predict heart disease that warrant further investigation.

Rawat, S. (2019). Heart Disease Prediction. Retrieved on December 6, 2020 from <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>

This article predicts heart disease using common risk factors. It utilizes the Cleveland dataset and outlines drawbacks within the dataset.

Zaidi, J. (2020). Project: Predicting Heart Disease with Classification Machine Learning Algorithms. Retrieved on December 6, 2020 from <https://towardsdatascience.com/project-predicting-heart-disease-with-classification-machine-learning-algorithms-fd69e6fdc9d6>

This article discusses classical algorithms used to predict cardiovascular events. It provides sample implementation of regression and decision tree analyses. The article also highlights the challenges and limitations of predicting health outcomes and disease instance.