Jeremy Dondoyano

HCDE 410

A7

Machine Learning and Heart Disease

Introduction

As I've gotten older, my health and longevity have become more and more important to me. However, similar to many in my generation, I get a lot of my information, including health/exercise information, on social media sites. These sites are great for spreading messages quickly and in a visually invigorating way, however, they are often times plagued with misinformation/opposing viewpoints. There has been a lot of research done surrounding different health factors and heart disease, however, when comparing many of them to each other, I often find contradictions in the literature. For example, this paper from ncbi shows that there is a lack of evidence to support the common assumption that high cholesterol levels affect cardiovascular (heart) disease rates. Furthermore, this Harvard article seems to find a middle ground, stating that new studies say cholesterol isn't as bad as we used to think, but we should still stay away from high amounts of certain foods. Moreover, this Mayo Clinic page says that cholesterol directly correlates with heart disease and its consumption should be limited.

Dataset

I plan to analyze a refined version of the Cleveland health dataset, which includes information about heart disease. I chose to work on this dataset because of how interested I have been in the medical/health realm lately. Also, because of HIPAA medical data like this is hard to come by. This could be a particularly useful dataset to help us understand what factors contribute most to heart disease. I hope to learn how to make a model to predict whether or not someone may have heart disease, and what factors contribute most to heart disease.

This dataset was sourced from Kaggle, and after searching for a Cleveland dataset that doesn't have as many issues, I found this one. The Kaggle usability score for this dataset is 8.82, falling short of a 10 because it is not updated frequently, however, that's not really applicable for this dataset anyway.

- The original source is: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
- The Kaggle link is: https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci
    - The Kaggle license links to Reddit API terms: https://www.reddit.com/wiki/api

This dataset is suitable for my needs because it has binary heart disease outcomes and factors that may contribute to said heart disease in the dataset, however, more observations would be beneficial to my work. Ethical considerations should always be made, one being that medical data is valuable, yet fragile, and nothing should be misconstrued. Furthermore, when I make a classification model, it needs to have as high of accuracy as possible, as such medical models could be the difference between life and death.

Overview and Research Questions:

The goal of this project is to create an ML model that will be able to identify and classify "people" with heart disease given certain health states. The data analysis portion of this project is to be able to find what attributes contribute the most towards heart disease. Visualizations will be made along with the model that gives insight into the distribution of some of the attributes.

Some research questions/hypotheses for this project include:

- What impact do cholesterol levels have on heart disease rates?
- What role do different resting blood pressure levels play on heart disease rates?
- What role can machine learning techniques play in the medical realm?

and

- Higher cholesterol levels lead to higher rates of heart disease.
- Heart disease is more frequent in males rather than females.

Methodology

I plan to investigate these phenomena by first taking into consideration the descriptive statistics of the data set. This will give me information about the significance of some of the attributes and the dataset as a whole, as well as provide more possibly needed background information. From there, I will subset and filter the data to show visualize the impact and relationship of different attributes on heart disease rates. Bar charts, scatterplots, treemaps, etc. will be utilized in this case. For the model, I plan to make several iterations, attempting to find what ML variation will provide the most accurate results without the risk of overfitting. These ML classification models may include SVM, logistic regression, Naive Bayes, clustering like KNN, and decision trees/random forests. After this, I will utilize different stats packages like sklearn, etc. to measure the accuracy of the models (confusion matrix or cross-validation). I will output the results in the form of a visualization or table when possible.

Findings

For a little introductory data analysis, I created some visualizations of the data to show correlations between age and heart disease, as well as sex and heart disease. The sex vs heart disease graph did not output any conclusive findings; it seems like heart disease rates are about the same in males vs females (at least from this dataset). On the other hand, age seems to be a significant factor in heart disease rates. The visualization output showed that the people who were diagnosed with heart disease were (on average) older than those who didn't.

To my surprise, my findings on cholesterol were inconclusive for the most part. In this dataset, the average person has a cholesterol reading of about 247mg/dl. Between above and below-average cholesterol readings in this dataset, those with above-average showed a ~54% chance of heart disease, and those who were below-average showed a ~40% chance of heart disease. This number gives us a little bit of information when thinking about a correlation between the two, however, neither the correlation table nor the logistic regression output table seem to show cholesterol having much impact on heart disease.

I then analyzed exercise-induced angina and resting blood pressure. After performing analysis techniques on exercise-induced angina, it was found to be a factor in heart disease rates. The logistic regression, as well as the visualizations that were created both showed having exercise-induced angina increases heart disease rates. On a similar note, it was found that higher resting blood pressure also increased the rates of heart disease among those who were included in this study. The visualizations played a key role in this, with the violin plots showing a greater median resting blood pressure in those who had heart disease vs those who didn't.

Moreover, for this dataset, machine learning techniques were not the best on this dataset. None of the models achieved greater than 87% accuracy, which is alright, but when talking about medical data, the highest accuracies should be prioritized (without overfitting). It is likely that with more data the models would have performed much better.

Discussion

There are big factors that could be considered a limitation of this analysis. To start, the dataset consisted of only 297 observations (rows). This isn't really that much data on a real-world note, however, it is sufficient for educational purposes. Of course, thicker data would have also been preferred, as getting some of the qualitative data like how a participant's daily lifestyle (exercise, eating habits, etc.) would have played a big role in this analysis.

As for the implications of this analysis, it confirmed what more recent literature has been saying, that cholesterol levels aren't as important as people may have believed in the past (in relation to heart disease). Furthermore, it may inspire individuals to look out for their blood pressure, and research ways to keep it down, as well as research for ways to limit exercise-induced angina.

It is also possible that this analysis/study could inspire others to perform analyses like this on other diseases. Although the machine learning models in this analysis did not perform the best, it could inspire others to gather thicker and more complete data to further test how machine learning fairs in a medical environment.

Conclusion

To conclude, in relation to heart disease, my analytical techniques showed a correlation with some features in this dataset and no correlation with others. To answer the research questions I came up with,

- Cholesterol levels don't seem to have much of an impact on heart disease rates, however, there did seem to be a very slight correlation between high cholesterol and high heart disease rates.

- Higher resting blood pressure results in a higher likelihood of being diagnosed with heart disease, based on this dataset.
- Machine learning in medicine is still up and coming. In this dataset, machine learning methods did not perform well enough to be directly considered for this medical topic, more research will need to be performed.

Works Cited

Soliman, Ghada A. “Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease.” *Nutrients*, 16 June 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6024687/.

“Cholesterol and Heart Disease: The Role of Diet.” *Harvard Health*, 15 Nov. 2021, www.health.harvard.edu/cholesterol/cholesterol-and-heart-disease-the-role-of-diet.

“High Cholesterol.” *Mayo Clinic*, 11 Jan. 2023, www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800.