Is it Possible to Predict Heart Disease with Machine Learning?

Business Problem

The term "heart disease" refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart. Decreased blood flow can cause a heart attack.

Is it possible to predict heart disease with machine learning? With the help of the dataset's attributes and applying machine learning techniques, we can identify the indication of the presence of heart disease in a patient. Typically doctors and healthcare companies use similar features and factors to predict this type of Disease with patients with similar lifestyles and family histories. This project focuses on developing a model that can be used to make early predictions about people at high risk of developing heart disease.

Background/History

I selected heart disease data for my first project because it relates to the healthcare industry currently benefiting from data science. I also wanted to focus on heart disease since it is a common disease in my father's family and most Hispanic families. Therefore, I hope to use machine learning to predict the presence of heart disease by common traits.

The selected dataset comes from the Kaggle competition section that used UCI's data of heart disease information. The dataset has fourteen attributes that are based on the profiles of different patients who are dealing with heart disease. This dataset has many helpful data points, including blood pressure, heart rates at other times, and if the patient was a smoker? These data points will help give insights into the factors leading to heart disease.

I plan to use different approaches to predicting the presence of heart disease. Boiling down this project, I see this being a classification problem in predicting heart disease in men and women of different age groups and lifestyles. I plan on using classification algorithms like random forests, ada boost, and logistic regression. Using different approaches will give an insightful image of which one works and the differences in the results. This will give me a better understanding of the results and how to handle them.

I also noticed that the code book has well-documented information for the results of this dataset. However, the info needs some modification to make it understandable for the public. Therefore, in the appendix section of this paper, I will include the original version of the code book. The appendix is located after the references section of this paper.

Data Explanation

I started with the dataset from Kaggle, which was in a CSV format. Then, I used the Pandas package to read and transform the dataset. I have also used the Seaborn and Matplotlib packages to create data visualizations that will help answer my research question. Going back to the dataset, I want to talk about it and give the reader a better understanding. This dataset has 14 variables and 303 rows of data, including anonymized health profiles of past medical patients.

Updating Data Types:

I looked at the data types through the Pandas info function to see if the data types are in formats that will be useful. I noticed that some data types were not in a helpful form, and I would need to correct them. I used the Pandas package to convert sex, target, cp, resting ECG, thal, fasting blood sugar, ca, exang, and slope data types to a definite format. This will make it easier to understand the results through data visualizations. After, I used the Pandas head function to see the first rows of the dataset to check the order of the data.

Updating Dataset's Columns:

I noticed the columns were not matching the code book. For consistency's sake, I wanted to rename the columns to check what was seen in the code book. This will make everything consistent and orderly. Not doing this will result in inconsistent results that would not be helpful to anyone. This section was straightforward since the Pandas package has a function that allowed me to rename the mismatched columns.

The columns that needed to be renamed are listed below:

cp became chest pain

trestbp became restingbp

fbs became fastingbloodsug

exang became angina

oldpeak became exinduced_depression

thalach became max_heartrate

ca became numofmajorvessels

Once the columns were renamed, I noticed some categorical variables needed to be converted to numerical variables. To make it easier for me when I need to use them for model fitting.

Updating Dataset's Variables:

Eight variables are needed to go through this process of conversion. I knew this would cause me to work harder initially, but this process would be beneficial in the long run.

The variables that needed to be converted are listed below:

sex was 0 for Female became Female and 1 for Male became Male

fastbloodsug was 0 for Normal became Normal, and 1 for High became

High

restecg was 0 for Normal became Normal, 1 for Abnormal became Abnormal, and 2 for Very Abnormal became Very Abnormal

angina was 0 for No became No and 1 for Yes became Yes

target was 0 for No Disease became No Disease and 1 for Disease became Disease

the slope was 0 for Upslope became Upslope, 1 for Flat became Flat, and 2 for Downslope became Downslope

chest pain was 0 for Asymptomatic became Asymptomatic, 1 for Non-Anginal became Non-Anginal, 2 for Atypical Angina became Atypical Angina, and 3 for Typical Angina became Typical Angina

After converting the eight variables from numerical to categorical, I wanted to check for anymissing data. Unfortunately, one variable had four missing data areas, and it was the restecg variable. So I decided to drop this one.

Analysis

This section will display a few visualizations to help me understand the dataset. As mentioned earlier in this paper, I used the Seaborn and Matplotlib packages to create these visualizations. The first visualization was a histogram that displayed the number of men and women who might or might not have the presence of heart disease in their bodies seen in Figure 1. Looking at the histogram's legend, blue represents the presence of heart disease. Orange represents no presence of heart disease.
When looking at the male section of the visualization, there is a clear difference between men who have the presence of heart disease and the men who do not. In the female area of the visualization, there is also a clear difference between women who have the presence of heart disease and the women who do not. This was a good base for me to see how many people had the presence of heart disease between the two genders. From this histogram, I noticed that men are more likely to have the presence of heart disease compared to women.

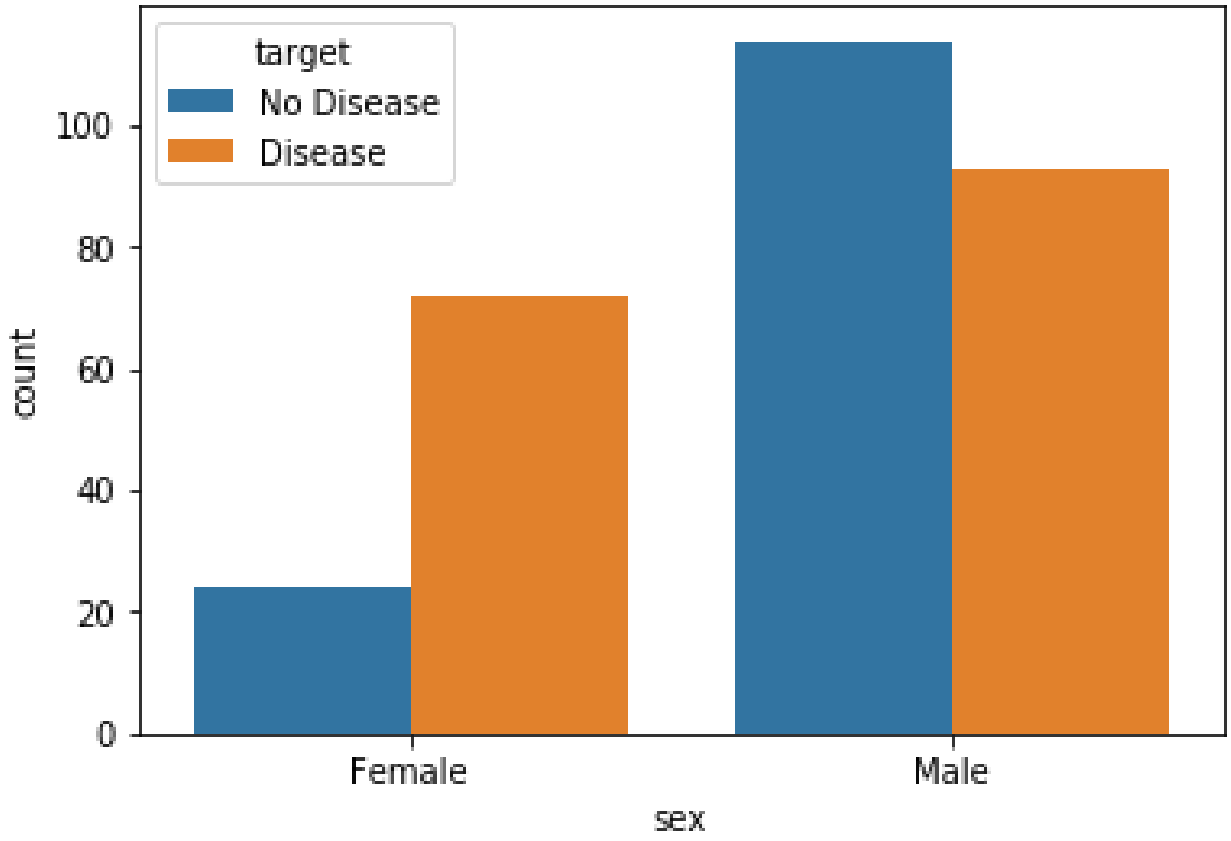

Figure 1: Histogram of Males and Females who might have heart disease or not.

The following visualization is a box plot that gave me more insights into the data between males and females, continuing with the dataset that compares males and females with high blood sugar levels and who have normal blood sugar levels. The reason for this comparison was that high blood sugar levels are one of many indicators that lead to heart disease. The box plot that compares which gender might have high blood sugar levels can be seen in Figure 2.

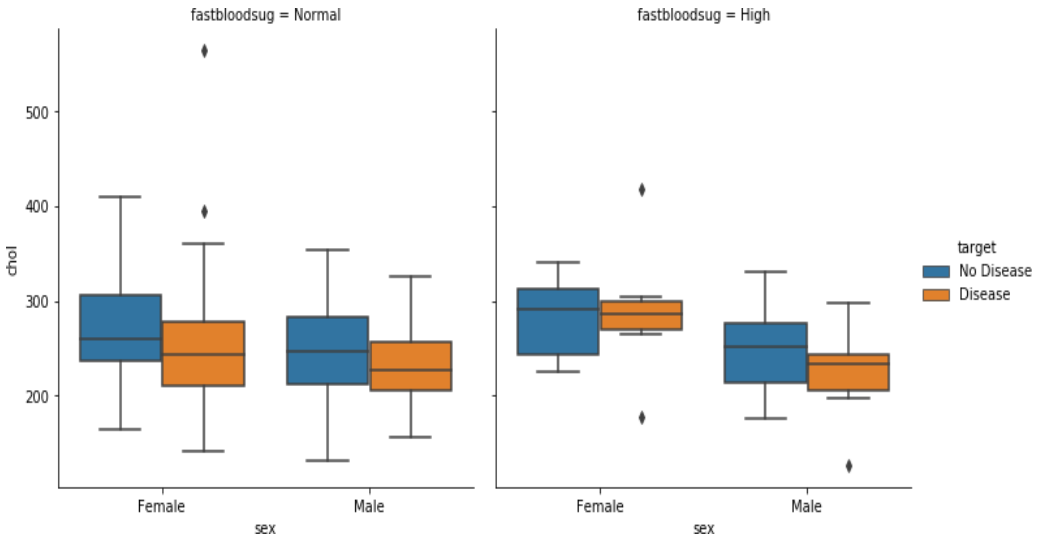

Figure 2: Box Plot of Males and Females with high blood sugar and normal levels with heart disease.

It seems that there is an overall distribution of Males and Females who have the presence of heart disease or not that also have high and regular blood sugar levels.

The violin plot, which is an unusual plot, gave me a better understanding of how heart rates are affected by the presence of heart disease or not. This was an injunction with effects between males and females. As seen in Figure 3, the higher the heart rate could lead a person having the presence of heart disease. The violin plot also displays men having high heart rates with the presence of heart disease exceeding the number of women who also have elevated heart rates with the presence of heart disease.

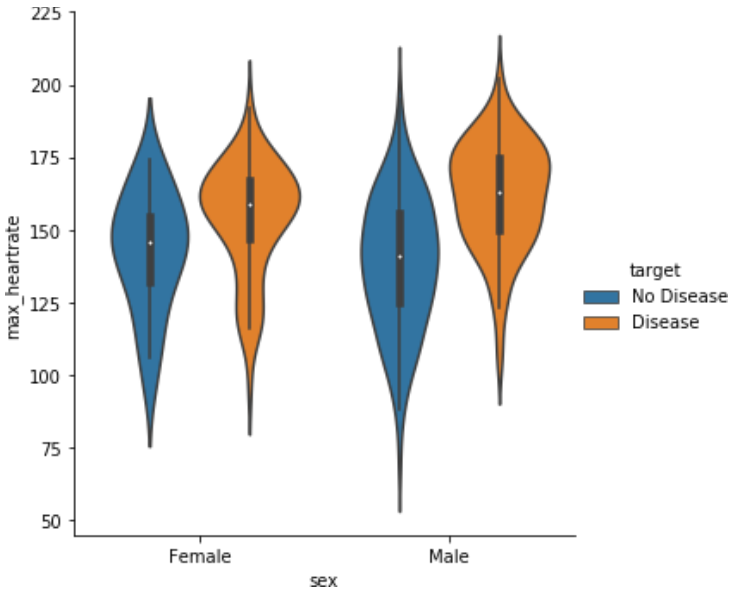

Figure 3: Violin Plot of Males and Females with high heart rates with heart disease.

Methods

I experimented with three types of models for this project. I have selected the Random Forest Classifier in combination with Grid Search with Cross-Validation. This combination gave me the best results for identifying the parameters to validate the model. The parameters are found with a grid search with four selectors: criterion, max_depth, max_features, and n_estimators. The results of my best parameters were criterion was best with gini, max_depth was 4, max_features was auto, and n_estimators was 10. What does this all mean?

The criterion has a selection between entropy and gini, which changes your results and accuracy. The max_depth has a selection from 4 to 8, and measuring it can change your results. Finally, the max_features has a selection of auto, sqrt, and log, which changes your results depending on the choice of max_features.
The n_estimators have a select range from 1 to 10. Once Grid Search gave me the best parameters, I calculated the prediction with the test split and the area under the curve score. My score came to be 88 percent accurate, which is excellent. Next, I validated my score again with Grid Search and Random Forest Classifier, as seen in Figure 4.

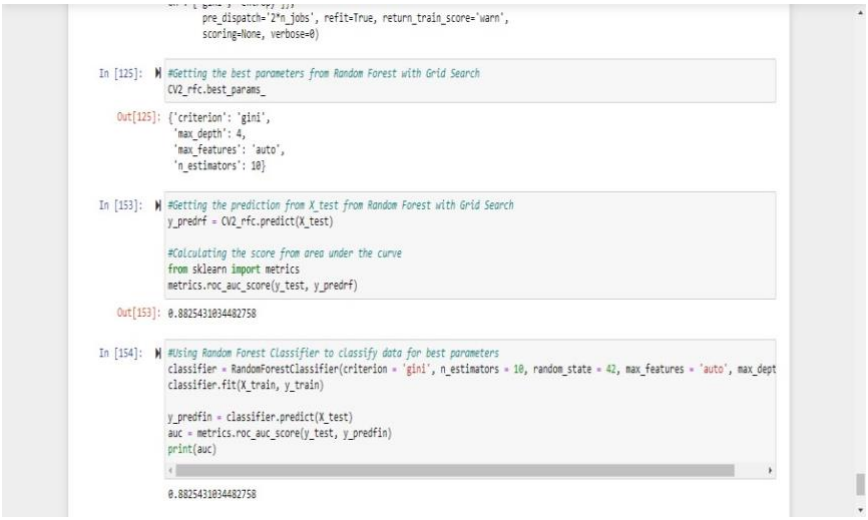

Figure 4: Validating the Random Forest Model with Grid Search.

As I mentioned earlier, this model has a high accuracy rate of 88%. The high accuracy rate could be due to the size of the selected dataset, which, compared to others, could be small. Either way, the model shows that it reached a high accuracy rate.

Conclusion

The focus of this project was to fit a model that could predict the presence of heart disease based on anonymized health profiles of past medical patients. I experimented with three classification models to find and analyze the best parameters for making predictions. Out of the three classification models selected for this project, the Random Forest Classifier with Grid Search was the best, with an accuracy rate of 88%. In addition, since the size of the selected dataset was relatively small, the experiment models could predict with high accuracy rates.

Limitations

I had to clean the data to filter a few observations from the available data in the Database. Unfortunately, this leads to a reduction in the valuable data points, which might have reduced the regression model coverage.

Challenges

The data available in the dataset was collected and cleaned for Kaggle exercise use. However, there are a limited number of instances of data. This is a limiting factor in predicting the target variable by creating an overfitting model.

Future Uses

With more data, these models can be extended to predict the heart attack outcome more effectively.

Recommendations

Based on the data available, this model predicts the heart attack outcome with reasonable accuracy. But this model should be regressed again when more data is available.

Implementation Plan

With the Currently given parameters, this model can be launched to predict the heart attack condition. However, as mode data becomes available, the model must be redone to ensure there is no slippage due to data.

Ethical Assessment

Largely this model has been built based on the stand ord data set, with a limited data set. Therefore, there could be missing analysis or incorrect interpretation due to incomplete data collection for other vaccine manufacturers.

So, users of the model need to be careful in inferring outcomes and applying the actions in real-world scenarios.

References

[1] Bhanot, K. Predicting presence of Heart Diseases using Machine Learning. February 12, 2019. https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c (Article on using machine learning to predict heart disease)

[2] Green, W. Machine Learning with a Heart: Predicting Heart Disease. July 31, 2018. https://medium.com/@dskswu/machine-learning-with-a-heart-predicting-heart-disease-b2e9f24fee84 (Article on using machine learning to find the presence of heart disease)

[3] Muthuvel, Marimuthu & Sivaraju, Deivarani & Ramamoorthy, Gayathri. (2019). Analysis of Heart Disease Prediction using Various Machine Learning Techniques. https://www.researchgate.net/publication/330981991_Analysis_of_Heart_Disease_Prediction_using_Various_Machine_Learning_Techniques (Article on the analysis of different machine learning techniques on heart disease predictions)

[4] Nashif, S., Raihan, Md.R., Islam, Md.R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. https://www.scirp.org/html/14-1560633_88650.htm (Article on detecting heart disease traits with machine learning)

[5] European Society of Cardiology. (2019, May 13). Machine learning overtakes humans in predicting early death or heart attack: Machine algorithm uses 85 variables to calculate risk in individuals. https://www.sciencedaily.com/releases/2019/05/190513081412.htm (Article on the benefits of using machine learning to predict heart disease)

[6] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707. https://ieeexplore.ieee.org/document/8740989 (Article on using hybrid machine learning techniques on predicting heart disease)

[7] Sharma, H. (2017). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. https://www.semanticscholar.org/paper/Prediction-of-Heart-Disease-using-Machine-Learning-Sharma/d0a5d4b8e8da3ee2a6bf8ac5d44196fb0365cf1c?p2df (Article on the survey results of the effectiveness on using machine learning for predicting heart disease)

[8] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2018, Article ID 3860146, 21 pages, 2018. https://www.hindawi.com/journals/misy/2018/3860146/ (Article on using a hybrid framework for using machine learning techniques to predict heart disease)

[9] Ronit. Heart Disease UCI Dataset. Kaggle. N.D. https://www.kaggle.com/ronitf/heart-disease-uci (Dataset of UCI's heart disease information)

[10] Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020). https://www.ijert.org/heart-disease-prediction-using-machine-learning (Article on predicting heart disease with machine learning techniques)

Appendix

Code Book of Kaggle/UCI Heart Disease Dataset

This Database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that ML researchers have used to date. For example, the "goal" field refers to the presence of heart disease in the patient. It is an integer-valued from 0 (no company) to 4.

Attribute Information:

age

sex

chest pain type (4 values)

resting blood pressure

serum cholesterol in mg/dl

fasting blood sugar > 120 mg/dl

resting electrocardiographic results (values 0,1,2)

maximum heart rate achieved

exercise-induced angina

oldpeak = ST depression induced by exercise relative to rest

the slope of the peak exercise ST segment

number of major vessels (0-3) colored by fluoroscopy

thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

The patients' names and social security numbers were recently removed from the Database and replaced with dummy values. One file has been "processed," that one containing the Cleveland database. All four unprocessed files also exist in this directory.

Acknowledgments

Creators:

Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D.,

Ph.D. Donor:

David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Inspiration

Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

10 Questions

1. Why are men more likely to have heart disease compared to women?
2. What are the common traits for women developing heart disease?
3. Could a person's lifestyle lead to developing heart disease?
4. Could heart disease be a result of unhealthy eating habits?
5. Does having high blood sugar levels lead to heart disease?
6. Does having a higher heart rate be a result of heart disease?
7. What causes men with heart disease to have higher heart rates?
8. Does having chest pains be a sign of heart disease?
9. Does having high cholesterol levels be a sign of heart disease?
10. What can people do to prevent developing heart disease?