Report/ White Paper

Heart Disease Prediction

DSC 680

Saurabh Shrestha

Business Problem

Cardiovascular disease (CVDs) are the leading cause of death globally which takes an estimated 17.9 million lives each year based on World Health Organization. They are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. In this project, I will be exploring different Machine Learning approaches for predicting whether a patient has 10-year risk of developing coronary heart disease (CHD)

Background/History

Cardiovascular diseases are a group of disorders involving the heart and blood vessels and one of the leading causes of death globally, according to the American Heart Association. In 2019, cardiovascular diseases took the lives of nearly 18 million people, accounting for 32% of deaths worldwide (World Health Organization, 2021). 85% of these deaths were due to heart attacks and strokes, with 38% among people under the age of 70. Early detection is critical in the treatment and management of cardiovascular diseases; wherein machine learning can be a powerful tool in detecting a potential heart disease diagnosis.

Data Explanation

The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors

- Sex: male or female(Nominal)
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Current Smoker: whether or not the patient is a current smoker (Nominal)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)
- Tot Chol: total cholesterol level (Continuous)
- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- Glucose: glucose level (Continuous)

Methods

First we need to work on if there are duplicates or missing values. Handling missing data is important as many machine learning algorithms do not support data with missing values. For data cleaning and pre-processing I would check and deal with missing and duplicate variables from the data set as these can affect the performance of different machine learning algorithm. Additionally, I would work on creating boxplot and treat outliers. Such as I will remove the outliers as anything over 75% of the maximum value. Additionally, perform Exploratory Data Analytics - Here I would to gain important statistical insights from the data such as distributions of the different attributes, correlations of the attributes with each other and the target variable.

Feature Selection

Since having irrelevant features in a data set can decrease the accuracy of the models applied, I would use the Boruta Feature Selection technique to select the most important features

which is later used to build different models. It tries to capture all the important, interesting features in a data set with respect to an outcome variable.

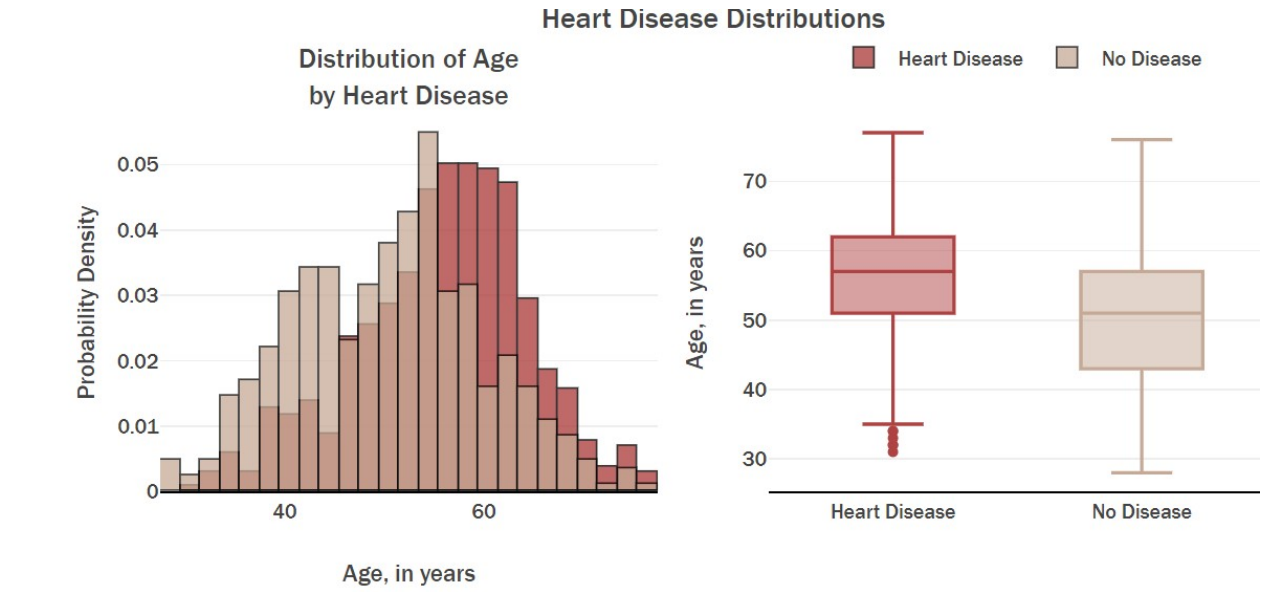
Modeling

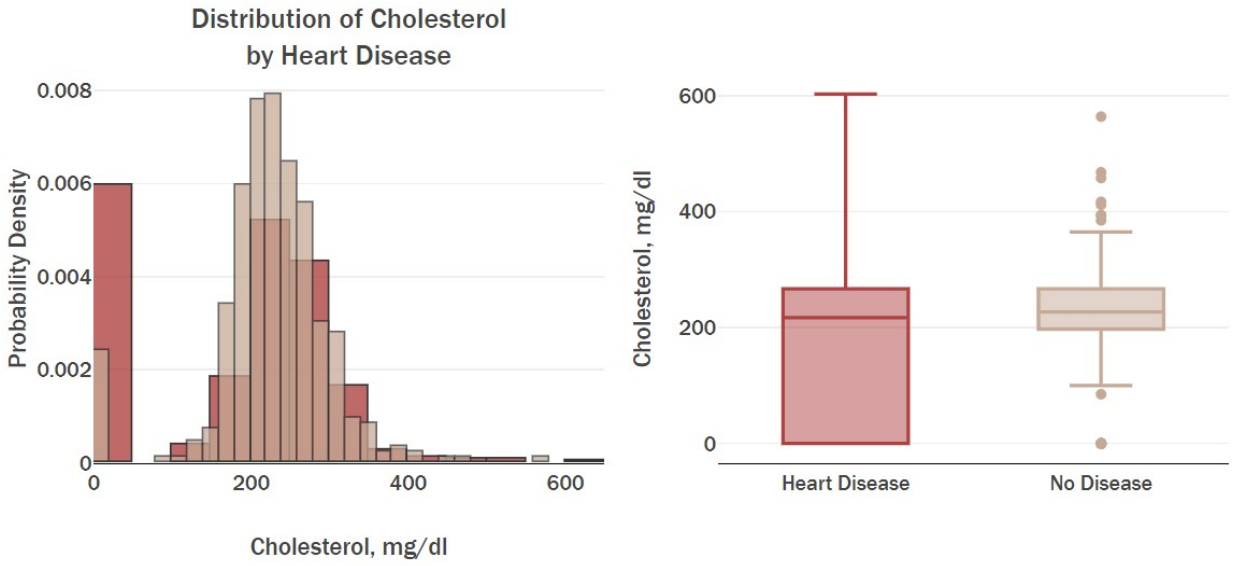
For the machine learning algorithm, I would scale the features to speed up the training of the classifiers and then split the data into a training and test set at a ratio of 0.8 to 0.2 respectively.

Several algorithms could be used for machine learning model. For this project I would chose following models: Logistic regression because it models the probability of a data point belonging to a particular class and assigns this point the appropriate label based on a chosen threshold. K-Nearest Neighbor because the algorithm assumes that similar data points that are near each other will belong to the same class.

After training each model and tuning their hyper-parameters using grid search, I would evaluate and compare their performance using the following metrics: **The accuracy score:** which is the ratio of the number of correct predictions to the total number of input samples, The F1 Score: Which is defined as the weighted harmonic mean of the test's precision and recall, The Area under the ROC Curve (AUC): Which provides an aggregate measure of performance across all possible classification thresholds.

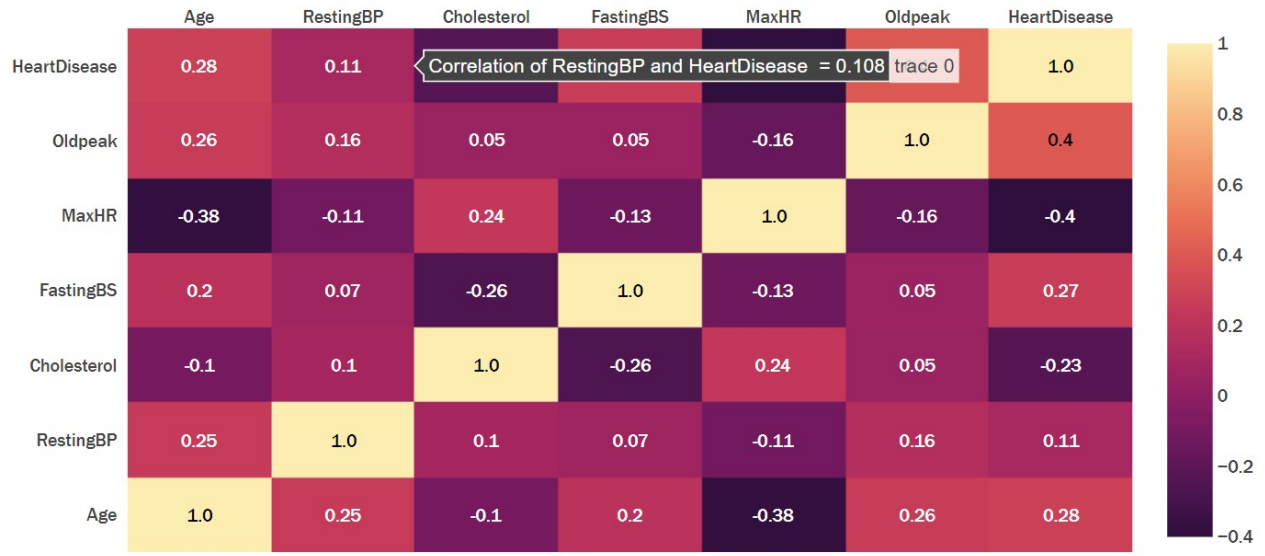
Analysis (with Illustrations)





Observations

- **Age:** In patients with heart disease, there is a smaller spread in the boxplot with the majority of patients between the ages of 51 to 62. There are also a few younger outliers in this group that are below the lower whisker. In patients without heart disease, there is a slightly wider variation in age that is more evenly distributed, with no outliers. The majority of patients in this group are within a younger age range of 43 to 57.
- **Cholesterol:** The distribution of cholesterol appears to be skewed to the right, especially in patients with heart disease where there are a large number of observations missing cholesterol levels that were entered as 0. These values will be addressed in the data cleaning section.



Observations

- **Correlation Matrix**: Based on the correlations and scatterplots, HeartDisease has the strongest positive association to OldPeak (correlation = 0.4) and the strongest negative association to MaxHR (correlation = -0.4). There is also a moderately strong relationship between Age and MaxHR of -0.38. As age increases, heart rate tends to decrease.

Conclusion

A number of aspects of this case study is critical to the success of the project. First, understanding the project, its descriptive variables and if necessary create innovative features to use for modeling which can make the difference between failure and success. Based on this project and the model create, it can be used as a simple screening tool and all that we need to do is to input ones: age, BMI, systolic and diastolic blood pressures, heart rate and blood glucose levels after which the model can be run and it outputs a prediction such as printing simple text “You are not at risk.”

Assumptions and Limitations

I didn’t know that there we maximum null values in one particular feature or that I had to work with so many categorical features. For this project the limitation is that the dataset was created by combining different datasets already available independently. Since this was medical dataset, I had to learn some variables which were new to me. I had to go through dictionary, google searches to learn about the features. For limitation, it was that there are limited variables to go

through to find the heart failure.

Challenges

Issues which could be faced while working on this project would be:

1. Outliers/Skewness in data.
2. Prediction model not returning expected results.
3. Learning curve.
4. Not enough impactful variables present in dataset.
5. Data size too small to apply prediction algorithm.

Recommendations

I would suggest this project would be helpful primarily to hospitals and clinics. Additionally, this project could be useful to biotech companies as well. With the help of analysis, hospital, clinics and biotech companies can determine what type of patients are likely to have heart disease.

Implementation Plan

The implementation plan contains the below steps to accomplish the project.

- Business Understanding
- Data Preparation
- Exploratory Data Analysis
- Data preparation for modelling
- Modelling
- Model Evaluation

Ethical Assessment

The reason for ethical assessment is to figure out what to do next. Ethics means the consideration of behaviors that will do good and assessment is what you do to figure out what you or someone else knows about doing a particular thing—and then figuring out what the next step is. For this project, I researched when working with the variables, followed the rules and regulations of ethics such as how could my system negatively affect impact patients? who is the most vulnerable and why?