Prediction of Heart Disease Using Machine Learning

Mary Akshara Allam

Department of Engineering and Computer Science CSU, Sacramento
maryaksharaallam@csus.edu

Abstract

Cardiovascular diseases are the most common leading cause of death over the few years. There are several factors for causing heart disease such as unhealthy diet, lack of physical exercise, smoking or drug usage, alcohol contribute to social problems. Some of the risk factors that cannot be determined such as family background and age factor. The main purpose is to develop a heart disease prediction system that predicts whether a patient is suffering from heart disease or not.

Machine Learning models K Nearest neighbors, Support Vector Machines (SVM), are used for Heart disease predictions.

1. Introduction

In the field of healthcare, Machine Learning is widely used in various fields of science like to identify the rare diseases, understanding the patterns to predict a rare disease and so on. According to the survey conducted by World Health Organization, 17.5 million total global deaths occur due to heart attacks and strokes[1]. The application of algorithms and interpretation of the patterns can be helpful in saving numerous people lives by anticipating the condition of the disease in advance. This project is focused on determining whether the patient has a heart disease or not by taking into consideration the UCI dataset. The dataset originally contained seventy six attributes which were collected from four different databases and fourteen attributes are used for our study[2].

The dataset contains 14 attributes such as:

- age - in years
- sex - (1=male, 0=female)
- cp - chest pain type
- trestbps - resting blood pressure in mm Hg
- chol - serum cholesterol in mm/dl
- fbs - fasting blood sugar > 120mm/dl
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina
- oldpeak - ST depression induced by exercise relative to rest
- slope - The slope of the peak exercise segment
- ca - Number of major vessels (0-3) colored by fluoroscopy.
- Thal - 3=normal, 6=fixed, 7=reversible defect
- Target - have disease or not (1=yes, 0=no)

Under progress

2. Methodology

The methodology used for this experiment will be unsupervised learning using K Nearest neighbor Algorithm. Apart from this, Support Vector machine algorithm will be implemented and results are compared between these algorithms.

3. Experiments and Deviations from Proposal

In order to begin experimentation, we loaded the dataset into a Pandas data frame object in python. The dataset contained various tools and options within the data frame that were used from the library itself. A great deal was learned in the process of exploring the various properties that the machine learning toolkits already provided. Since the libraries already contained inbuilt functions, we were able to filter the dataset to the architecture and patterns we wanted that made the observations and predictions very stable. There are no deviations observed in this process from the original project proposal.

4. Dataset Description

The dataset contains 76 attributes, but among them a subset of 14 attributes were chosen for our prediction models. The dataset was extracted from the ML related projects website Kaggle(https://www.kaggle.com/ronitf/heart-disease-uci) and the size of the dataset downloaded is 303 rows and 14 columns.

5. Data Exploration Plan and tools

The data was loaded from csv file into the development environment i.e. python. Few operations were performed to check whether there are redundant values and null values.

Using Exploratory Data Analysis, I was able to figure out how many patients are diseased depending on the age factor. I was able to notice the frequency to heart disease depending the chest pain type. As of now, the tools used were PyCharm and python.

5.1.1 Data Cleaning

The raw data acquired from Kaggle needs to be filtered so that the resulting data set can be used for building the models. In the data, there is one categorical variable which is modified into numeric values using the python library named Pandas. Redundant data were handled by using *duplicate()* and null values were handled by *isnull()*.

5.1.2 Variable Identification

Variables in the dataset were handled using *dtype()* and importing the library NumPy and Pandas.

Some libraries such as scikitlearn are also being looked into for use during the remainder of the project.

6. Data Insights after Data Exploration

Data plots were achieved with the help of some python libraries such as matplotlib as well as seaborn. Many data plots were available for use such as scatterplots, grouped histograms, heatmaps etc.

During the data cleaning process, we were able to plot some graphs that helped us to look into the right direction during the prediction process. The graph in fig-1, represents a plot that was plotted to determine the gender that had more frequency/are more prone to heart diseases

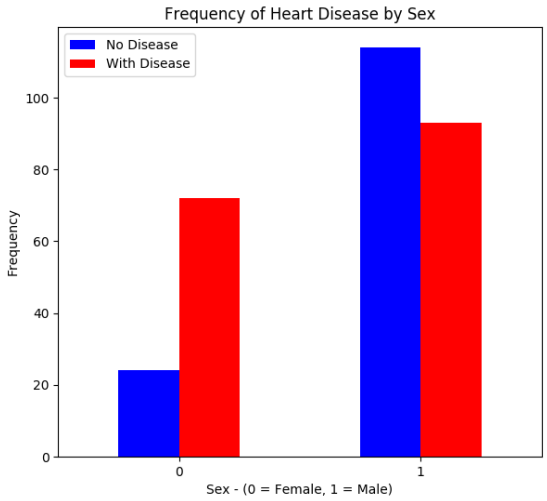

Fig-1

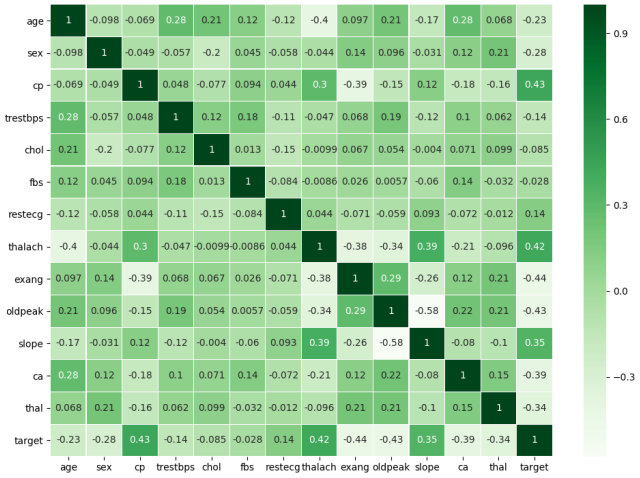

Fig-2

Similarly, we also were able to plot a heatmap that showed all 14 columns and a representation of how each column affected a person with heart diseases as shown in Fig-2.

7. Applied Models

Under progress

8. Graphical User Interface

Under progress

9. Schedule for Remaining Tasks

I am looking forward to completing these steps by week 12, I will have all the pre-processing completed thoroughly and then I will start off with analyzing the algorithms in week 13. Then, I will proceed with applying those models to my dataset to see their prediction accuracy.

By the end of week 13, I plan my documentation of the project to be complete, so that I can go ahead with poster presentation.

10. Conclusion

Under progress

11.References

[1]https://file.scirp.org/Html/14-1560633_88650.htm [Status: Available Accessed April 11th]

[2]https://www.kaggle.com/ronitf/heart-disease-uci [Status: Available Accessed April 8th]