Predicting Heart Diseases Using Machine Learning

Name: Rawan Ibrahim Mohamed Ibrahim ElShenieky
Student ID: 21091000188
Faculty and Department: Software Engineering
Lecturer's Name: Dr. Mohammed Al-Hubaishi
Course Code: CEN 122(2)
Programming Language: Python
Assignment Number: 6
Date Submitted: 24th of May 2022

Introduction

The heart is one of the most vital organs in human's body. It is responsible for various function important for the stability of health status for each one of us. If we take a look at the body and how it functions, we will see that it all works in sync according to a very specific order that we cannot control. Therefore, we must make sure that this working system does not get disrupted by any means, especially for an important organ like the heart.

From that point, we used the development of technology- specifically the department of machine learning- to help us predict diseases that might affect the heart. With some attributes put into consideration we can come up with a system that can help us be alarmed whenever we reach a risky point when it comes to heart diseases. There are several types of heart disease there in the world; however Coronary Artery Disease (CAD), and Heart Failure (HF) are the most common heart diseases that are present. [1]

After doing research in both the medical and the technical fields, I present to you the summary of how we can use the science and data we have nowadays in favor of the heart. In the following sections, the system will be analyzed thoroughly. Firstly by analyzing the attributes the system depends on to induct a conclusion. Not to mention, there will also be other attributes to be discussed other than the ones in the dataset used. Secondly, the code written will be inspected with explanations of each code block; to understand each output produced.

Analyzing the attributes

- Age: age in years

- Sex: 1-male, 0-female
- CP: chest pain type
- 1: Typical angina - chest pain related decrease blood supply to the heart
- 2: Atypical angina - chest pain not related to heart
- 3: Non-anginal pain - typically esophageal spasms (non heart related)
- 4: Asymptomatic - chest pain not showing signs of disease

- BP: Blood pressure in mm Hg on admission to the hospital

- Cholesterol: serum cholesterol in mg/dl
- above 200 is cause for concern

- FBS: fasting blood sugar > 120 mg/dl
- 1 = true; 0 = false
'>126' mg/dL signals diabetes

- EKG results: resting electrocardiographic results
- 0: normal
- 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

- Max HR: maximum heart rate achieved
- By subtracting the age from 220

- Exercise Angina: exercise included angina (1-yes, 0-no)

- ST depression: induced by exercise relative to rest
- looks at stress of heart during excercise
- unhealthy heart will stress more

- Slope of ST: the slope of the peak exercise ST segment
- 1: Upsloping - better heart rate with excercise (uncommon)
- 2: Flat - minimal change (typical healthy heart)
- 3: Downsloping - signs of unhealthy heart

- Number of vessels floru: number of major vessels (0-3) colored by flourosopy

- Thallium: thallium stress result
- 3: normal
- 6: fixed defect
- 7: reversible defect

- Heart Disease: The chance of presence of heart disease
- 1: Presence of heart disease
- 0: Absence of heart disease

More Attributes

The aforementioned attributes are not the only we can depend on in general. Since each one of those attributes itself gets affected by another factors and behaviors, therefore we can conclude another relations between them and causing heart diseases.

The psychological health is without a doubt as important as the physical one. Not only because it has a direct and great connection with our physical health, but also for the fatal consequences one might face because of it. In this situation, the heart is not an exception, it is getting affected by our psychological health. That's why as surprising as it may sounds, traffic is a problem that affects our heart; due to the fact that it causes stress. [2]

Staying up late in the night all disturbs the heart rate and the hormones of the body. The researchers say shift work has a bad impact on the body's circadian rhythm, which is known as your internal clock. Also, getting less than 6 hours of sleep a night on routine basis may raise your risk of having higher blood pressure and cholesterol.[2]

Moreover, it is found that hepatitis C may cause inflammation of the body's cells and tissues, including those in the heart. Again, this inflammation is no good for the heart to function properly.[2]

Heredity, which is when members of a family pass traits from one generation to another through genes, is an uncontrollable factor of causing heart diseases. Unfortunately, the family history with heart disease affects the upcoming generations of the family as it goes on. It is proved that this history increases the chance to get a heart disease.[3]

Analyzing the code

Section One: Libraries and Analysis

(A)
I started by importing the libraries we are going to use to execute the code. Some of the libraries belonged to panda data frame; for analysis. Others (and most of them) were of sklearn, for implementing machine learning into the system. Also, uploading the dataset took place in that section.
(B)
I previewed the histogram of all the columns available in my dataset as a sort of analysis. I adjusted the size of the histograms to be clear and larger than the default size.

Section Two: Machine Learning

(A)
Then comes the step of splitting the data to two unequal parts: test (1/3 of the dataset) and train (2/3 of the dataset). Also I set X variable to be all the independent factors of the data, while y variable is the dependent variable (Heart Disease).
(B)
Our target in here is to know the score, which is how far each factor affects getting heart disease. We trained the system with the X_train and y_train data to be used later for the decision tree.
(C)
In here, we plotted the scores vs the features or the factors in my dataset. It is seen that factor number 3 (Chest pain type) has shown a score of approximate 0.77 which is the highest score achieved by the system. Meaning, that when creating a decision tree, the chest pain type will be our root (the starting point)
(D)
I set the features of the dataset as the headers of my dataset, to be used later when drawing the decision tree. I fit the classifier where comes the machine learning part to my train values. I also set the size of the tree to be visible.
(E)
Lastly, I wanted to check the accuracy of the decision tree classifier model, so we ran a test to figure out the accuracy_score. The result was satisfactory with 0.7 (70%) accuracy.

References

1. https://ieeexplore.ieee.org/document/9155586 [Predicting Heart Disease at Early Stages using Machine Learning: A Survey] (2-4 July 2020)
2. https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c [Predicting presence of Heart Diseases using Machine Learning] (Feb 12, 2019)
3. https://www.cdc.gov/heartdisease/risk_factors.htm [Know Your Risk for Heart Disease] (Dec 9, 2019)
4. https://pythonguides.com/scikit-learn-decision-tree/ [Scikit learn Decision Tree] (Dec 16, 2021)
5. https://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html#sphx-glr-auto-examples-tree-plot-unveil-tree-structure-py [Understanding the decision tree structure]