**Predicting Presence of Heart Disease using Machine Learning**

Neeraj Katewa, Rahul Unadkat, Siddhartha Dutta

Department of Computer Science and Engineering

Amity School of Engineering & Technology

Amity University, Mumbai

[neeraj.katewa@student.amity.edu](mailto:neeraj.katewa@student.amity.edu),

[rahul.unadkat@student.amity.edu](mailto:rahul.unadkat@student.amity.edu), [siddhartha.dutta1@student.amity.edu](mailto:siddhartha.dutta1@student.amity.edu)

Problem Statement

Heart disease or cardiovascular disease (CVD) consists of a number of conditions that influence the heart and its operations apart from just heart attacks.[1] These diseases cause the heart to be unable to pump the required amount of blood to parts of the body to fulfil normal functionalities, often leading to a heart failure. Heart disease and its associated failure remains a widespread complex and deadly disease. These diseases have remained the leading causes of death globally in the last 15 years. Ischaemic heart disease and stroke are the world’s biggest killers, accounting for a combined 15.2 million deaths in 2016.[2]

Accurate and on time diagnosis of heart disease is important for heart failure prevention and treatment. The diagnosis of heart disease through traditional medical history has been considered as not reliable in many aspects. This is especially a challenge in developing countries where are availability of diagnostic apparatus and shortage of physicians and others resources which affect proper prediction and treatment of heart patients.[3] To classify the healthy people and people with heart disease, non-invasive based methods such as machine learning and data mining are reliable and efficient. This case study will focus on methods to predict the status of angiographic disease (diameter narrowing > or < than 50% in any major vessel.) using various machine learning algorithms.

Introduction

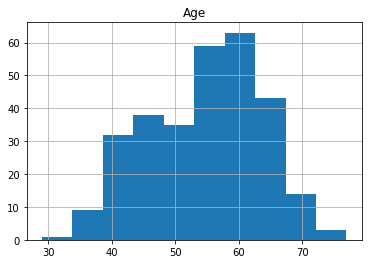
The dataset used in this case study is from the UCI Machine Learning Repository. [4] A popular dataset for heart disease prediction. The “Cleveland heart disease dataset 2016” is used by various researchers and can be accessed from online data mining repository of the University of California, Irvine. This dataset was used in this research study for designing machine-learning-based system for heart disease diagnosis. The Cleveland heart disease dataset has a sample size of 303 patients, 76 features, and some missing values.

Out of the 76 features, only 14 of them were chosen and the reasons are explained ahead.

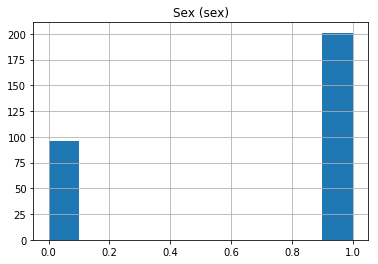
1. Age: Looking at the development of heart diseases, age plays a very important role where the risk increases by approximately three times with each decade
2. Sex: Women who are in the pre-menopausal stage have less risk of heart disease than men. More research is still being carried out to compare risks once past menopause. Also males with diabetes have a lesser risk than women with the same.
3. Angina(chest pain): This arises when there is not enough oxygen-rich blood for the heart muscle. It can be described as squeezing or pressure in the chest. It could a affect other parts like shoulders,neck,jaw,back or the arms.
4. Resting Blood Pressure: High blood pressure tends to damage arteries that are connected to the heart and along with that, other conditions like obesity that cause high blood pressure increase the risk even more.
5. Serum Cholesterol: Arteries are narrowed by high levels of low-density lipoprotein also known as the “bad” cholesterol. On the other hand, high levels of high-density lipoprotein lowers the risk of a heart attack.
6. Fasting Blood Sugar: Sugar levels may rise if there isn’t enough secretion of insulin by the pancreas or if the body doesn’t respond to insulin like it should, which leads to a higher risk of an attack.
7. Resting ECG: The USPSTF concludes with average certainty that the possible harms of screening with resting ECG maybe equal to or may exceed the benefits
8. Maximum Heart rate achieved: Increase in heart rate by around 10 bits per minute has shown to increase the risk of cardiac death by approximately 20%.
9. Exercise induced angina: The pain caused by angina can differ from mild to severe and could spread to other parts as well.
10. Peak exercise ST segment: The ST Segment is a section of the ECG between the end of the S wave and the beginning of the T wave. A horizontal or downsloping ST segment indicates a higher likelihood of multivessel disease.
11. ST depression induced by exercise to relative rest:
12. Number of major vessels:
13. Thal: is an inherited hemoglobin disorder resulting in chronic hemolytic anemia that typically requires life-long transfusion therapy.

Review of System

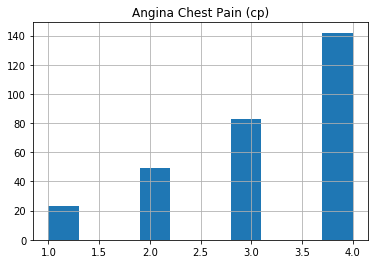
1. Age: Specifies the age of the individual. The average age in the dataset is 54.54 years.



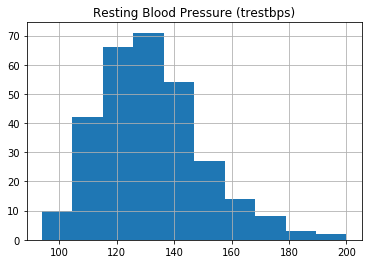
1. Sex: Specifies the gender of the individual with the following mapping: 1 = male and 0 = female. The dataset has 201 male individuals and 96 female individuals.



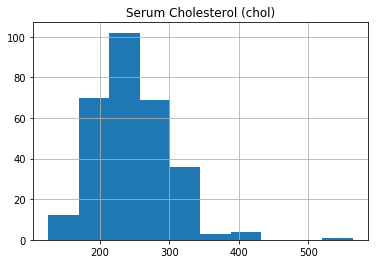
1. Angina (Chest Pain): displays the type of chest-pain experienced by the individual using the following format :1 = typical angina 2 = atypical angina 3 = non — anginal pain 4 = asymptotic



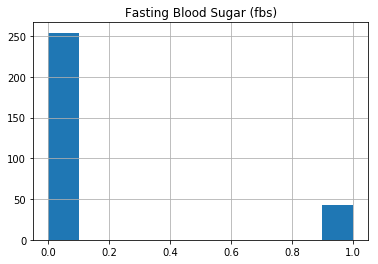
1. Resting Blood Pressure: displays the resting blood pressure value of an individual in mmHg (unit)



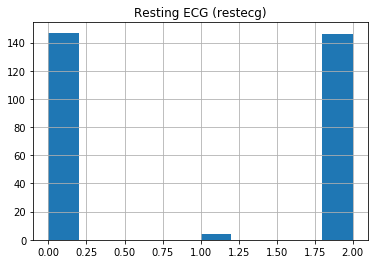
1. Serum Cholesterol: displays the serum cholesterol in mg/dl (unit)



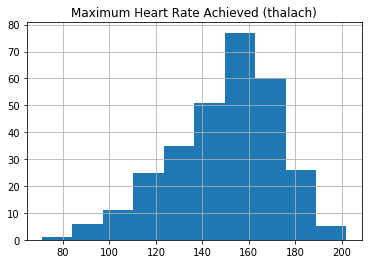
1. Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) [43] else : 0 (false) [254]



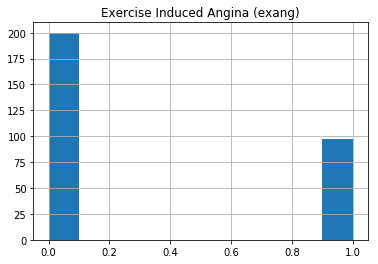
1. Resting ECG : displays resting electrocardiographic results 0 = normal (147) 1 = having ST-T wave abnormality (4) 2 = left ventricular hyperthrophy (146)



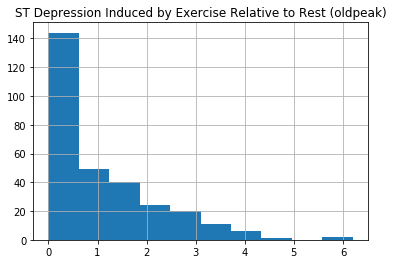
1. Max heart rate achieved : displays the max heart rate achieved by an individual. (Mean: 149.5993265993266)



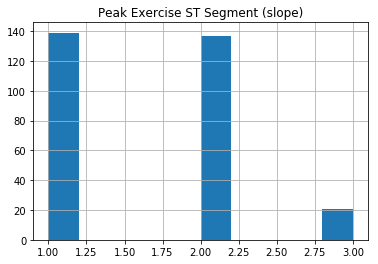
1. Exercise induced angina :  
   1 = yes (97)  
   0 = no (200)



1. ST depression induced by exercise relative to rest: displays the value which is an integer or float.

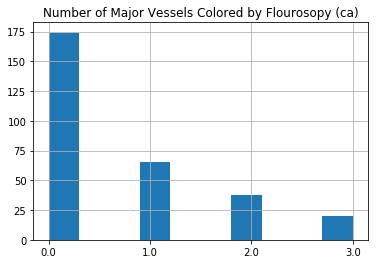


1. Peak exercise ST segment : 1 = upsloping (139) 2 = flat (137) 3 = downsloping (21)

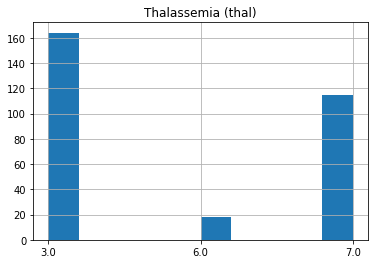


1. Number of major vessels (0–3) colored by flourosopy : displays the value as integer or float.

0: 174, 1: 65, 2: 38, 3: 20



1. Thal : displays the thalassemia : 3 = normal (164) 6 = fixed defect (18) 7 = reversible defect (115)



1. Diagnosis of heart disease : Displays whether the individual is suffering from heart disease or not : 0 = absence (16) 1, 2, 3, 4 = present (137)

Dataset Info

Analysis

Log Reg

Decision Criteria’s

Include Confusion Matrix

Conclusion

Conclude

References

[1] [A. Methaila, P. Kansal, H. Arya, and P. Kumar, “Early heart disease prediction using data mining techniques,” in Proceedings of Computer Science & Information Technology (CCSIT-2014), vol. 24, pp. 53–59, Sydney, NSW, Australia, 2014.]

[2] [<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>]

[3] [<https://www.scirp.org/html/5-9601148_35396.htm>]

[4] [https://archive.ics.uci.edu/ml/datasets/Heart+Disease]