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,
rahul.unadkat@student.amity.edu, 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. In the last 15 years, heart diseases have remained the leading cause of death globally. Ischaemic heart disease and stroke are the world's biggest killers, accounting for a combined 15.2 million deaths in 2016. [2]

Given the criticality of the human heart, accurate and timely diagnosis of heart disease is important for heart failure prevention and its treatment. Diagnosis through traditional medical history is considered unreliable in many aspects. Heart disease diagnosis is especially a challenge in developing countries where availability of diagnostic apparatus and shortage of physicians and others resources affect proper prediction and treatment of heart patients. [3] Therefore, to classify heart diseases, non-invasive based methods such as machine learning and data mining have be proven to be reliable and efficient. This particular case study will focus on methods to predict the status of an angiographic disease i.e. the narrowing in the diameter of any major blood vessel by 50% using various machine-learning algorithms.

Introduction

The dataset referred to and used in this case study is a popular dataset used for heart disease prediction – the "Cleveland Heart Disease Dataset." This dataset has been used by various researches and can be accessed from the online data-mining repository of the University of California, Irvine Machine Learning Repository. [4] This dataset was used in this case study for designing a machine-learning-based system for heart disease diagnosis. The Cleveland heart disease dataset has a sample size of 303 patients, 76 features, and includes some missing values.

While this dataset consists of 76 features, most researchers use only 14 of these features and this case study follows the same path of feature selection. The significance of the selection of these 14 parameters are described as follows.

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.
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.
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.

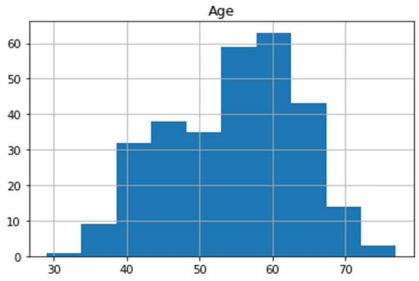
5. Serum Cholesterol: Arteries are narrowed by high levels of low-density lipoprotein also known as the “bad” cholesterol. On the other hand, high levels of high-density lipoprotein lowers the risk of a heart attack.
6. Fasting Blood Sugar: Sugar levels may rise if there is not enough secretion of insulin by the pancreas or if the body does not 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 beats 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: There are multiple conditions associated with ST depression. Some of these include hypokalemia, cardiac ischemia, and medications such as digitalis.
12. Number of major vessels: Major blood vessels are analysed through fluoroscopy colouring.
13. Thalassemia: This is an inherited haemoglobin disorder resulting in chronic haemolytic anaemia that typically requires life-long transfusion therapy.
14. Diagnosis of heart disease: This is the parameter that has to be predicted by the machine learning model.

Review of System

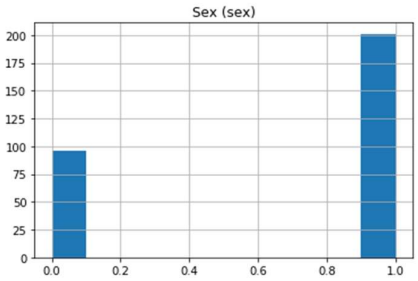
An exploratory data analysis was carried out on the Cleveland heart disease dataset. Out of 303 records, 6 records contained missing values in the *ca* and *thal* attributes. Instead of imputing the missing values, these records were simply removed from the dataset resulting in a total of 297 total records.

Each attribute was represented through a histogram to further understand their distribution in accordance with the data set description. [5] Appropriate data pre-processing steps were applied to this dataset. [6] The exploratory data analysis is summarized as follows.

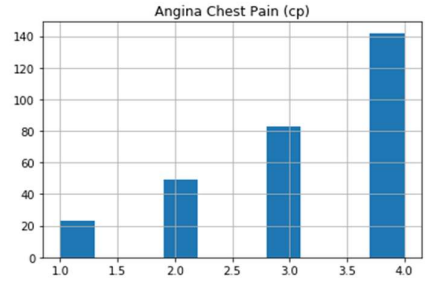
1. Age: This feature specifies the age of the individual. The average age in the dataset is 54.54 years.



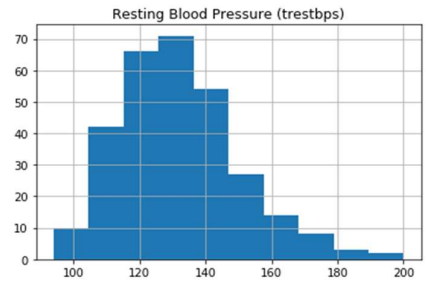
2. Sex: This feature specifies the gender of the individual using the mapping: {1=male, 0=female}. The dataset has 201 male individuals and 96 female individuals.



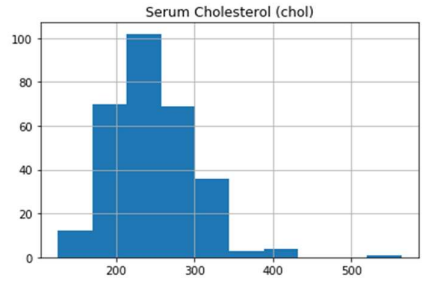
3. Angina (Chest Pain): This feature classifies the type of chest-pain experienced by the individual using the mapping:{1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptotic}.



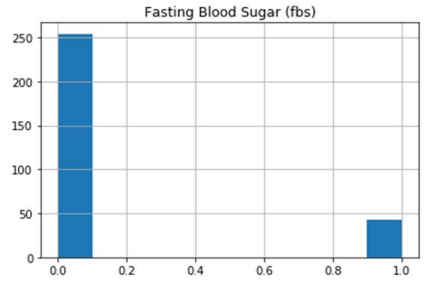
4. Resting Blood Pressure: This displays the resting blood pressure value of an individual in mmHg (unit).



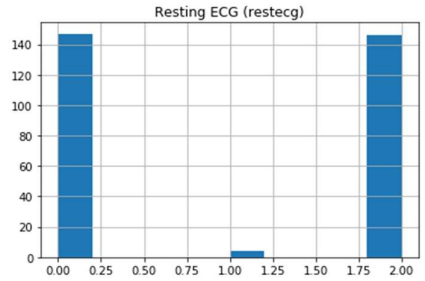
5. Serum Cholesterol: This feature displays the serum cholesterol in mg/dl (unit).



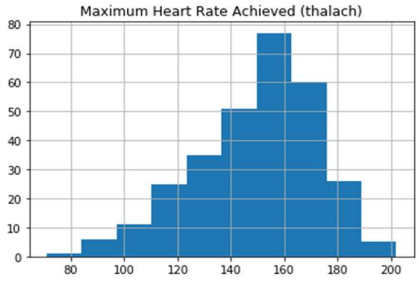
6. Fasting Blood Sugar: This feature compares the fasting blood sugar value of an individual with 120mg/dl. The values are mapped as {If fasting blood sugar > 120mg/dl then:1 (true) else : 0 (false)}.



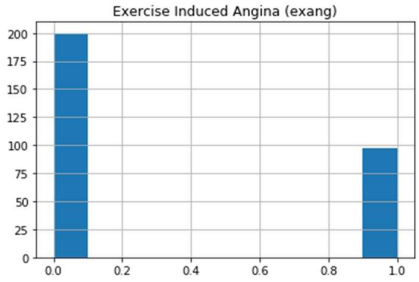
7. Resting ECG : This feature displays resting electrocardiographic mapped as {0=normal, 1=having ST-T wave abnormality, 2=left ventricular hypertrophy}.



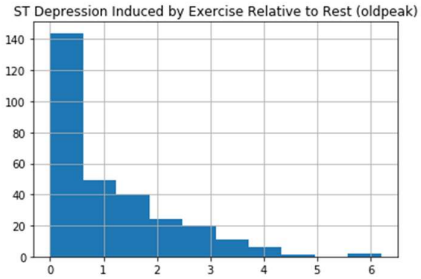
8. Max heart rate achieved: This feature displays the max heart rate achieved by an individual. The average max heart rate achieved is 149.59.



9. Exercise induced angina: This feature represents whether exercised caused an angina mapped as {1=Yes, 0=No}.

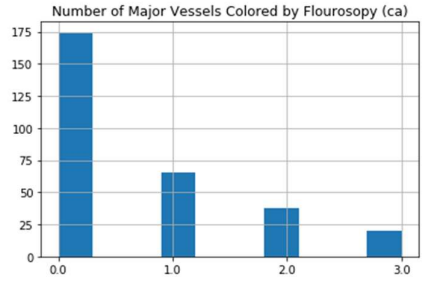


10. ST depression induced by exercise relative to rest:

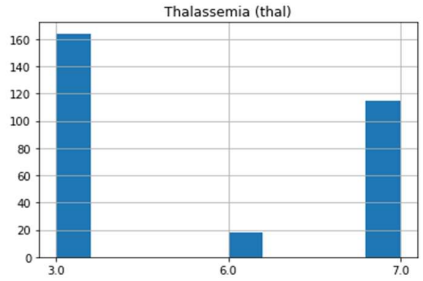


11. Peak exercise ST segment: Mapping = {1=upsloping, 2=flat, 3=downsloping}.

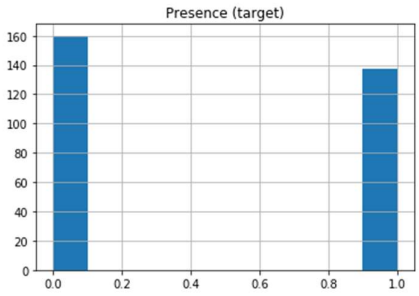
12. Number of major vessels (0–3) coloured by fluoroscopy:



13. Thal: This feature displays the thalassemia mapped as: {3=normal, 6=fixed defect, 7=reversible defect}.



14. Diagnosis of heart disease: Displays whether the individual is suffering from heart disease or not mapped as {0=absence, 1=present}. There are 160 records for no heart disease and 137 records with presence of heart disease.



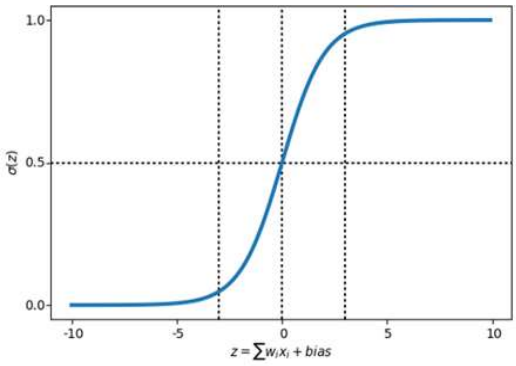
Analysis

The inherent problem domain of the heart disease prediction is binary classification i.e. whether the heart disease is present (1) or not (0). Among the popular machine-learning algorithms, logistic regression was used as the classification algorithm in this case study. [7] The logistic regression model is based

on the concept of probability. A more complex cost function than the one of linear regression is used, known as the 'sigmoid function' which is given as:

$$f(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function maps any real value into another value between 0 and 1 i.e. a probabilistic value, as shown by the sigmoidal curve.



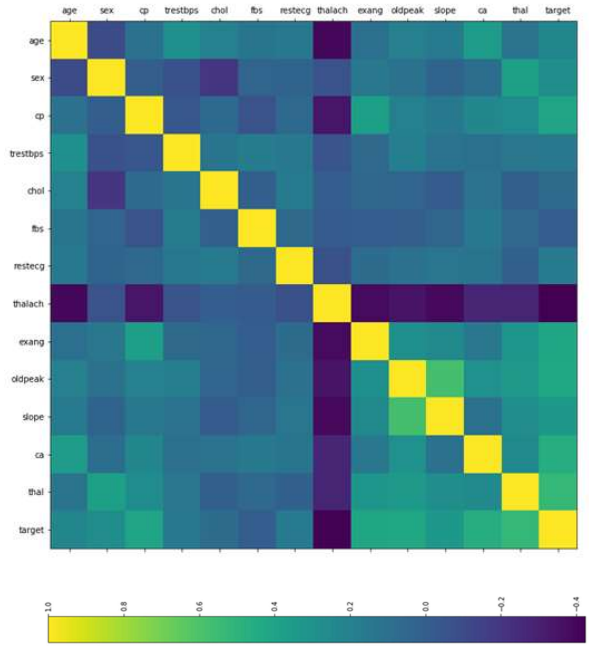
Sigmoid Function Graph

Three major assumptions were dealt with in the consideration of this model:

1. Absence of outliers in the data. This was demonstrated through the histograms shown above.

2. Dichotomous nature of dependent variable. The dependent variable is the prediction of heart disease diagnosis which is binary in nature (yes or no).

3. Low correlation between independent variables. This can be assessed by a correlation matrix among the predictors as shown below.



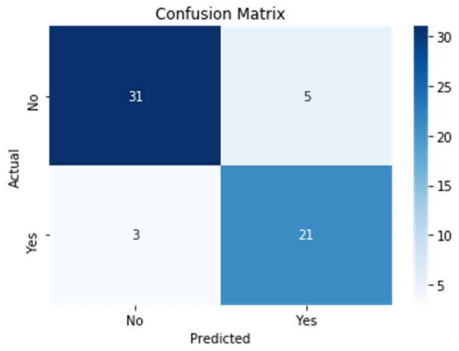
Visualizing the correlation matrix it can be seen that there are no features that are highly correlated with each other either i.e. both: a negative correlation and a positive correlation is not evident.

It can also be seen that there is no single feature that has a very high correlation with the target value. Also, some of the features have a negative correlation with the target value where as some have a slightly positive correlation.

Hence, logistic regression was chosen to be the machine-learning classification algorithm to prepare the model for the heart disease prediction problem.

Decision Criteria's

The dataset was divided into a training set and testing set in the ratio 80:20. The LogisticRegression class of the sklearn.linear_model was used for training purposes. The results of comparing the predicted values with the true values of the testing set can be summarized through the confusion matrix below.



From the above confusion matrix, the following metrics can be calculated.

1. Accuracy Score

$$\frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$= \frac{21 + 31}{21 + 5 + 31 + 3} = 86.67\%$$

2. Recall Score: The recall is intuitively the ability of the classifier to find all the positive samples.

$$\frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{21}{21 + 3} = 87.50\%$$

3. Precision Score: The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

$$\frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{21}{21 + 5} = 80.77\%$$

Conclusion

As the world moves to a more data-centric environment, it is important to harness this data for tasks such as prediction. In this case study, we have seen a useful classification application to the prediction of the presence of a heart disease. Using a basic logistic regression algorithm, the model was able to achieve an accuracy score of 86.67%. There is further scope in improving this model through feature engineering, data augmentation. There has been extensive research on this dataset involving different classification algorithms to develop a more optimal and accurate prediction model. [8]

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
[5] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names
[6] https://github.com/siddharthapdutta/ML-Lab-Sem6/blob/master/Case%20Study/Data%20Preprocessing.ipynb
[7] https://github.com/siddharthapdutta/ML-Lab-Sem6/blob/master/Case%20Study/Predicting%20Presence%20of%20Heart%20Disease%20using%20Machine%20Learning.ipynb
[8] Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms” in Hindawi, vol. 2018, article ID 3860146.