The ability to predict the chances of having a heart attack

Abstract

One of the deadliest health problems that people are unable to predict or see coming is a heart attack. A heart attack has many tell tales, but people often do not realize them before it is too late. This data gives us an outlook on the potential factors leading to a heart attack, most of the time the factors that lead to a heart attack are related to the blood flow to the heart and potential things that could affect it. The effects of a heart, depending on severity, could lead to serious illness, disability, and a lower quality of life. In this report, I have done different types of classifications based on factors that could potentially lead to a heart attack.

1. Introduction

A human being has always been able to feel when something is wrong with its health before it becomes serious, but when it comes to heart attacks it is a little too late because the effects damage the person for life. Many studies have been done to assess the human health and predict heart attacks. The seriousness of heart attacks should be known by many, according to the Center for Disease Control and Prevention (CDC), approximately 805,000 Americans have heart attacks each year. This number should give us an idea about the seriousness surrounding this disease. The following report is an in-depth study of a dataset presenting the factors that lead to a heart attack (Health Dataset). The following report gives a study on multiple classification methods that give us the ability to predict the potentiality of a heart attack.

2. Data and Methods

The study follows a dataset taken from Kaggle representing a “Heart Attack Analysis & Prediction Dataset”. The dataset is composed of 303 rows and 14 attributes, these attributes consist of 4 quantitative attributes and 10 categorical attributes. The categorical variables in the dataset are: sex (Gender), cp (Chest Pain), fbs (Fasting Blood Sugar), restecg (Resting Electrocardiographic), exng (Exercise Angina), oldpeak (Previous Peak), slp (Slope), caa (Number of Vessels), thall (Thal Rate), output (Target Variable). The numerical variables in the dataset are: trbps (Resting Blood Pressure), chol (Cholesterol), thalachh (Maximum Heartrate). It is important to note that the variables in the dataset are all very important and depending on the classification methods used we will identify which ones hold more value compared to other variables. The age attribute is a numeric variable with values ranging from 29 to 77, the sex

attributes is a categorical variable with two factors 0 and 1, cp is a categorical variable with four factors 0, 1, 2, 3, trtbps is a numerical variable with a range of values of 94 to 200, chol is a numerical variable with a range of values of 126 to 564, fbs is a categorical variable with two factors 0 and 1, restecg is categorical variable with three factors 0, 1, 2, thalachh is a numerical variable with a range of values of 71 to 202, exng is a categorical variable with two factors 0 and 1, oldpeak is categorical variable with 139 factors, slp is a categorical variable with three factors 0, 1, 2, caa is a categorical variable with five factors 0, 1, 2, 3, 4, thall is a categorical variable with four factors 0, 1, 2, 3, and finally output is a categorical variable with two factors 0 and 1.

I decided to partition the data in a 70/30 using the holdout method for the training and testing dataset. And continued by processing the data using all the classification models. Before any classification and with further analysis of the data, I removed some attributes to increase the accuracy and overall quality of the prediction. I deleted oldpeak, slp, caa and thall.

3. Result

Using the K-Nearest Neighbor Classifier with one model, I was able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	32	5
1	9	44

Statistics

Accuracy	0.8444
Error	0.1556
Sensitivity	0.7805
Specificity	0.8980
Precision	0.8649
Recall	0.7805
F1	0.8205

Using the K-Nearest Neighbor Classifier with repeated 10-fold cross validation, I was able to determine the following:

Confusion Matrix

	Reference	

Prediction	0	1
0	32	5
1	9	44

Statistics

Accuracy	0.8444
Error	0.1556
Sensitivity	0.7805
Specificity	0.8980
Precision	0.8649
Recall	0.7805
F1	0.8205

Using the K-Nearest Neighbor Classifier with .632 bootstrap, I was able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	32	5
1	9	44

Statistics

Accuracy	0.8444
Error	0.1556
Sensitivity	0.7805
Specificity	0.8980
Precision	0.8649
Recall	0.7805
F1	0.8205

With the three methods we were able to determine a good accuracy rate of 84.44%, an error rate of 15.56%, a sensitivity rate of 78.05%, a specificity rate of 89.80%, a precision rate of 86.49%, a recall rate of 78.05%, and an F1 measure of 82.05%, which evidently gives a good prediction model.

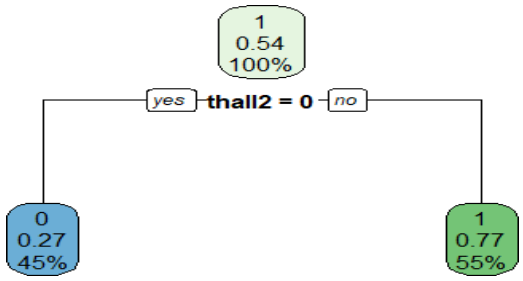
Using the Decision Tree using the CART algorithm and cross validation, we were able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	32	9
1	9	40

Statistics

Accuracy	0.8000
Error	0.2000
Sensitivity	0.7805
Specificity	0.8163
Precision	0.7805
Recall	0.7805
F1	0.7805



Using the Naive Bayes Classifier using cross validation, we were able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	30	6
1	11	43

Statistics

Accuracy	0.8111
Error	0.1889
Sensitivity	0.7317
Specificity	0.8776
Precision	0.8333
Recall	0.7317
F1	0.7792

4. Conclusion

In summary, it is important to note that depending on the attributes you use and the attributes you remove the accuracy of the predictive model will vary as these attributes play an important. I used different classification models with each giving a different result for the predictive model. I used all three classifiers to start; the entire training set, cross validation, and bootstrapping, where all three proposed the same results, with an accuracy rate of 84.44% which is high. The Decision Tree gave an unlikely tree, with an accuracy rate of 80.00% which is also high. The Naïve Bayes classifier on the other hand gave an accuracy rate of 81.11% which is considerably high. With all the analysis we could conclude that the first three classifiers gave a better predictive model in the case predicting the likelihood of a heart attack.

5. References

Cdc. (2021, February 26). Million hearts® costs & consequences. Retrieved April 06, 2021, from https://millionhearts.hhs.gov/learn-prevent/cost-consequences.html#:~:text=Heart%20disease%20and%20stroke%20can,speech%20difficulties%2C%20and%20emotional%20problems.

Heart attack symptoms, risk, and recovery. (2021, January 11). Retrieved April 06, 2021, from https://www.cdc.gov/heartdisease/heart_attack.htm

Rahman, R. (2021, March 22). Heart attack analysis & Prediction Dataset. Retrieved April 06, 2021, from https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset