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 and have proven that between Logistic Regression, SVM , and Neural Networks, the later had a better accuracy.

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. I continued by using a repeated 10-fold cross validation with 5 repeats to train the classifier for both Logistic Regression and SVM classifiers. 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 cp, fbs, restecg, slp, caa and thall.

3. Result

Using the Logistic Regression classifier, I was able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	26	5
1	15	44

Statistics

Accuracy	0.7778
Error	0.222
Sensitivity	0.6341
Specificity	0.8980
Precision	0.8387
Recall	0.6341
F1	0.7222

Using the SVM classifier, I was able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	27	10
1	14	39

Statistics

Accuracy	0.7333
Error	0.2667
Sensitivity	0.6585
Specificity	0.7959
Precision	0.7297
Recall	0.6585
F1	0.6923

Using the Neural Networks method of classification, we were able to determine the following:

Confusion Matrix

	Reference	
Prediction	0	1
0	25	3
1	16	46

Statistics

Accuracy	0.7889
Error	0.2111
Sensitivity	0.6098
Specificity	0.9388
Precision	0.8929
Recall	0.6098
F1	0.7246

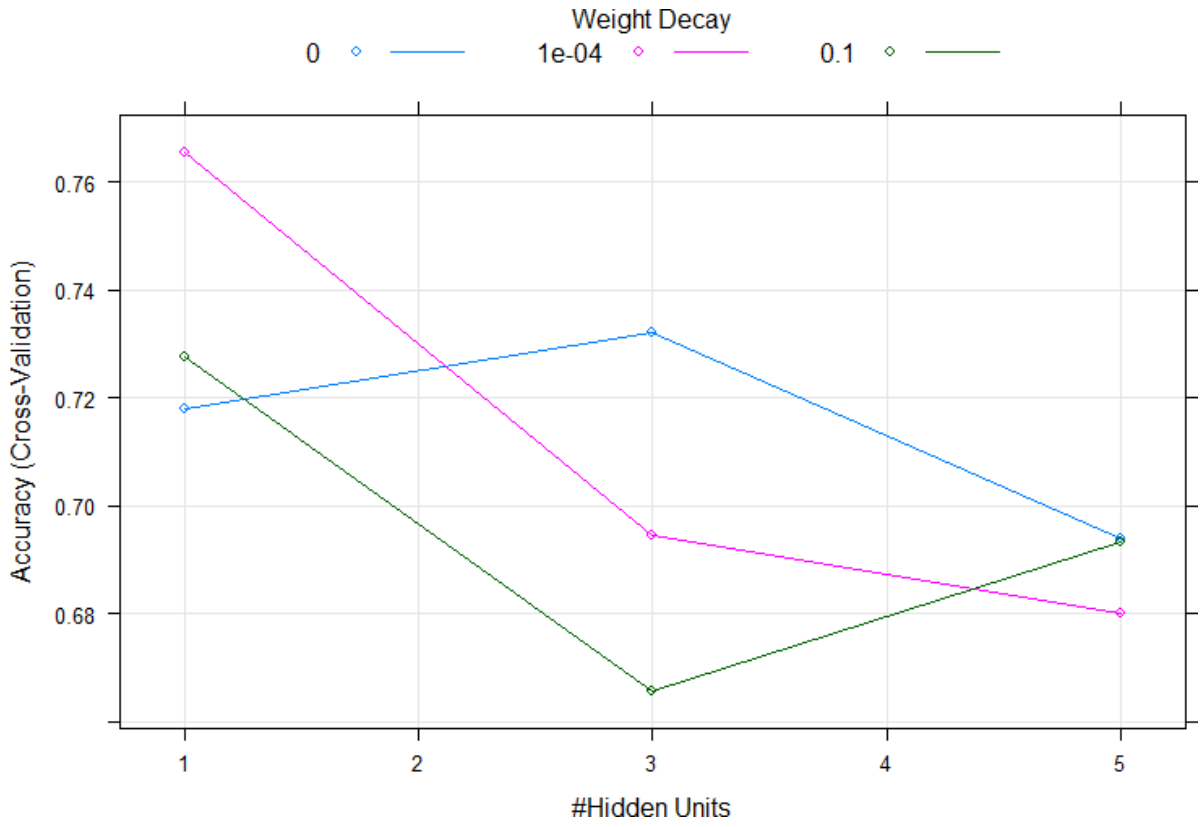

Figure 1: Neural Network Hidden Units

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 three classifiers to determine a model for predicting the chances of a heart attack. The Logistic Regression classifier provided an accuracy of 77.78% which is considerably high in our case, the SVM classifier in turn provided an accuracy comparable to the one before with 73.33%, but the best classifier for the following dataset was the Neural Network classifier with an accuracy rate of 78.89%, which out bested both classifiers in our case.

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