Heart disease data classification

Gabriel Henriques

Rio de Janeiro, Brazil

Abstract

In order to bring two main fields of research together, medicine and technology, I intend to evaluate a logistic regression algorithm as a method of classification of the proposed dataset, to try and draw conclusions with biological depth and relevance, after the data is thoroughly processed. After analyzing all biomarkers inside the data and define a probability for the prediction, we can assess metabolic pathways that are common amongst patients in cardiovascular disease.

**1. Introduction**

As of 2018, cardiovascular diseases are a critical public health condition. According to the American Heart Association1, 1 of every 3 deaths in the US are from cardiovascular diseases. Healthcare data these days are being broadly explored in order to build models and analysis for better increase success rates in diagnostics and to develop precision treatments through the combination of science and technology, thus, reducing the costs of treatments, healthcare services and medications.

***1.1 Problem Statement***

Healthcare data these days are being broadly explored in order to build models and analysis for better increase success rates in diagnostics and to develop precision treatments, and by trying to find metabolic patterns in patients, therefore, predicting its best course of action.

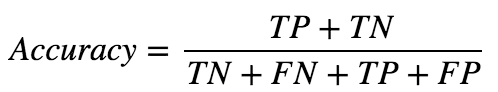
By using logistic regression2 as a classification algorithm, i intend to evaluate the probability occurrence of the dependent variable, which is, in this project’s dataset, whether a patient has cardiovascular disease (v=0) or not (v=1).

The goal of this project is to define a model good enough that will allow us to classify patients with cardiovascular disease and enhance diagnostic precision in future patients and point attributes that would help prevent disease based on the outcome.

***1.2 Metrics***

Standard classification metrics such as, accuracy, precision and recall, will be used in this project.

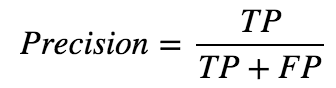
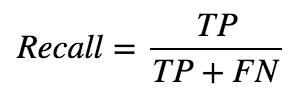
In prediction problems like the one presented in this paper, accuracy is a common evaluation metric. It measures the proportion of correct predictions. The problem with accuracy is that if we have unbalanced datasets or a tendency towards some class, the results can be misleading, and not always a higher accuracy means a positive outcome. But since its very common technique used in binary classifiers, and our dataset is balanced, we are using accuracy as a reference metric to base some of our conclusions if needed. Accuracy formula can be define as below:



*Fig. 1: Accuracy formula[[1]](#footnote-1)*

Precision and recall are powerfull metrics to understand the behavior of classified items, being those positive or negative. Precision tells us how many of those classified as “Positive”, were really a “Positive”. Recall is similar, and gives us a metric of how frequently we classify something as “Positive” or “Negative”, when it really belongs to that class.

If we consider a classifier that has high Recall and low Precision, we would have many observations as being “Positive”, being able to find most of the correct classes, but as higher cost in false positives. On the other hand, if we have a high Precision but a low Recall, we would have lower observations being classified as “Positive”, giving us the ability to identify correctly the “Positive” classes, but at a cost of leaving a lot of “Positive” outcomes pass through without being correctly classified. Precision and Recall formulas can be defined as below:



*Fig. 2: Precision and Recall formula*

**2. Analysis**

**3. Methodology**

**4. Results**

**5. Conclusion**

**6. Improvement**

**7. References**

1. Benjamin, E. J. *et al.* Heart disease and stroke statistics - 2018 update: A report from the American Heart Association. *Circulation* **137,** E67–E492 (2018).

2. Kleinbaum, D. G. & Klein, M. *Logistic regression : a self-learning text*. (Springer, 2010).

1. TP: True positives; TN: True negatives; FN: False negatives; FP: False positives; [↑](#footnote-ref-1)