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.

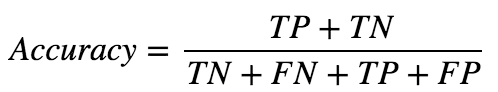
By using logistic regression2 as a classification algorithm, i intend to evaluate the probability occurrence of the dependent variable, which is, in this project’s dataset, whether a patient has cardiovascular disease (v=0) or not (v=1).

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:

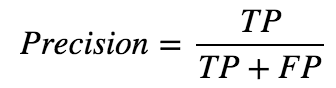
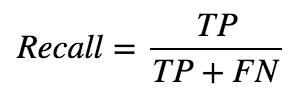


*Fig. 1: Accuracy formula[[1]](#footnote-1)*

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**

***2.1 Data Exploration***

The dataset used in this project contains data from 4 different cities but only the Cleveland dataset is considered due to consistency and to the fact that the other datasets lost many of data points. It consists mainly of 303 instances (patients), and 14 attributes. The “goal” attribute is the #14, which refers to the presence or absence of heart disease in patients. Below a representation of all the attributes inside the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Attribute | Description | Type |
| 1 | age | Age in years | int |
| 2 | sex | Female or Male | bin |
| 3 | cp | Chest pain type | cat |
| 4 | trestbps | Resting blood pressure (mm Hg) | con |
| 5 | chol | Serum cholesterol (mg/dl) | con |
| 6 | fbs | Fasting blood sugar (< 120 mg/dl or > 120 mg/dl) | bin |
| 7  8  9  10  11  12  13  14 | restecg  thalach  exang  oldpeak  slope  ca  thal  num | Resting electrocardiography results  Max. heart rate achieved during thalium stress test  Exercise induced angina (yes or no)  ST depression induced by exercise relative to restinclinação do Slope of peak exercise ST segment  Number of major vessels colored by fluoroscopy  Thalium stress test resultHeart disease status: num. of major vessels with >50% narrowing | cat  con  bin  con  cat  int  cat  int |

***Table 1:*** *Patients atributes inside the Cleveland’s dataset. The #14 attribute representes the predict variable, or the label, for the classification problem. Column “type” indicates if the atribute is binary (bin), integer (int), categorical (cat) or continuous (con).*

The dataset is well **balanced**, containing 139 patients (46%) with disease, 160 without (54%), for the target attribute. The dataset has many important features about the patients metabolism and their respective diagnosis of heart disease. With that I will be able to apply machine learning algorithms and statistical tools to draw conclusions on both perspectives: technology and biology.

So we have an idea of how the data is dispersed amongst patients, in *table 2* we have a small sample of patients and some key attributes from a biological standpoint:

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Sex | Cholesterol | Label |
| 63 | Male | 233 | < 50% |
| 37  41 | Male  Female | 286  250 | > 50%  < 50% |

***Table 2:*** *Small sample of patients with some key attributes. The #14 attribute is the dependent variable, which is the label that we will model our algorithm to predict.*

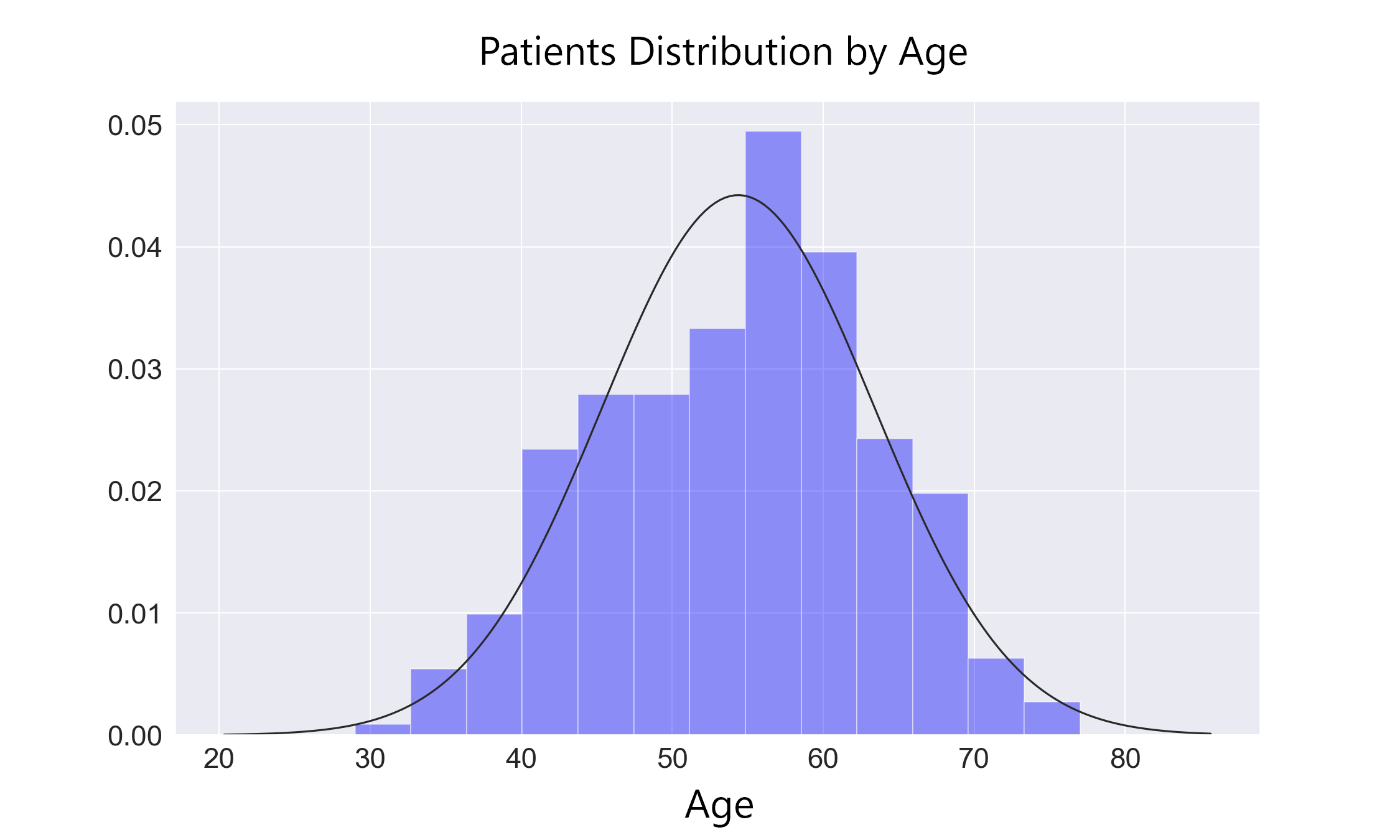
After analyzing the dataset, it was generated a statistics report considering 3 important continuous variables so we could have an idea of how the data was dispersed and if there was any tweaks to be made in the preprocessing part that we could spot right now. Below is the *table* 3 containing the statistics report visualization:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Age | Blood Pressure | Cholesterol |
| median | 54 | 131,7 | 245,6 |
| std | 9,02 | 17,6 | 48,51 |
| min val  max val | 29  77 | 94  200 | 126  417 |
| 25% | 48 | 120 | 211 |
| 50%  75% | 55,5  61 | 130  140 | 240,5  274,75 |

***Table 3:*** *Statistical report of the entire dataset, containing 3 variables as headers for better visualization.*

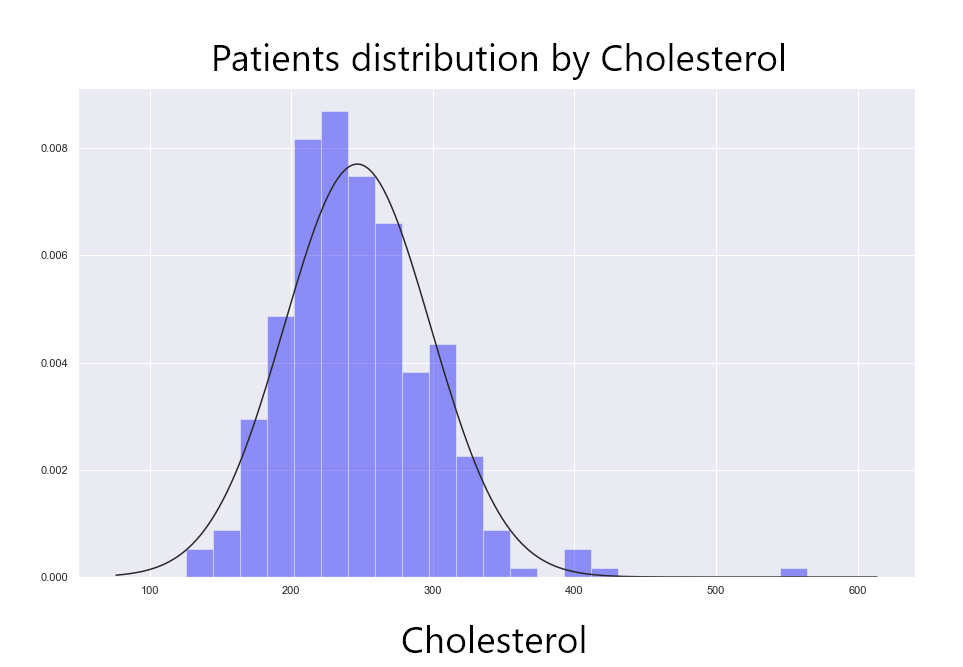
***2.2 Exploratory Visualization***

Below we can see the distribution of patients in the dataset, regarding their age. Figure 3 also plots a gaussian curve so we can assess whether the data is normally distributed.



*Figure 3: Age attribute with a gaussian curve*

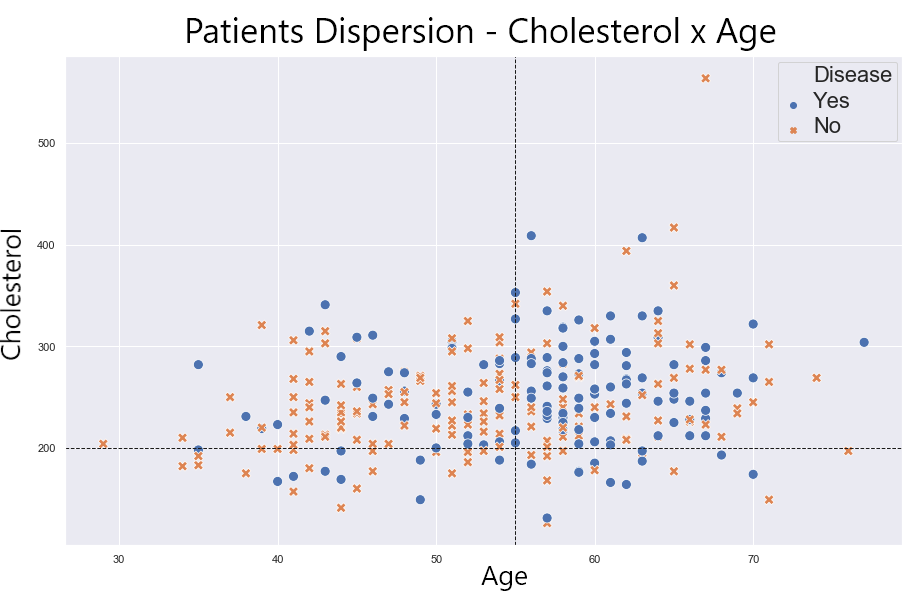
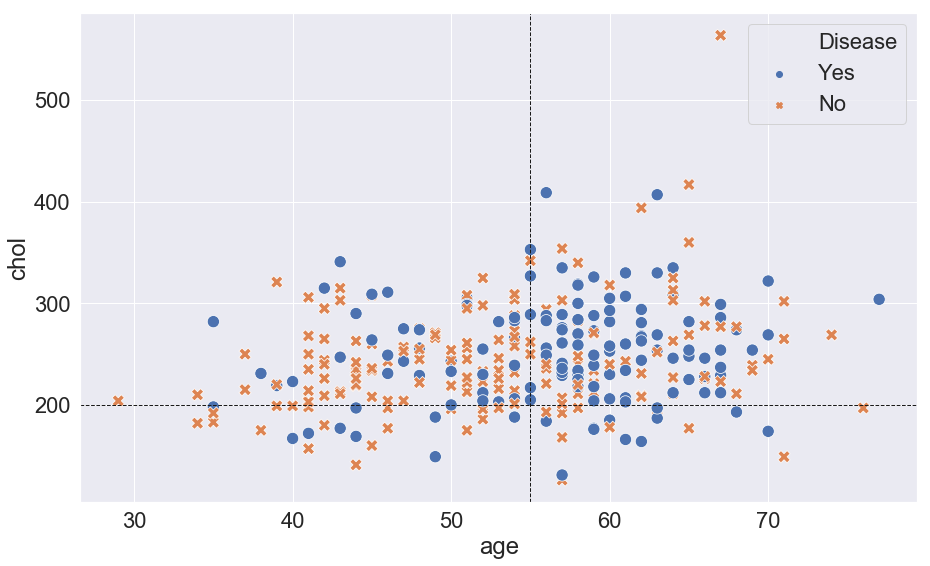
Another important attribute is the Cholesterol. Through *figure 4* we can see how the distribution of the patients by cholesterol is handled.



*Figure 4: Cholesterol attribute with a gaussian curve*

It is possible to notice a mild right skewness, probably due to a few outliers to the right. We will be addressing that on the preprocessing chapter.

In order to better visualize the full dispersion of patients with regards of the two attributes we analyzed on the first plots (figure 3 and 4), it was created a scatter visualization (figure 5), that used as legend the presence of heart disease on those patients:



*Figure 5: Scatter plot of patients.*

*Figure 5: Dispersion of patients with disease*

It is important to observe how the data behave. Here we can see that most of the patients with presence of heart disease (label = yes), are located within the top quadrant to the right, meaning the we may have a predictable behavior for that group.

***2.3 Algorithms and Techniques***

It was decided to use a Logistic Regression (LR) classifier due to its behavior of generating probabilities for the outcomes, which is very welcome on most health related prediction problems.

By classifying our heart disease dataset, we will be generating a model that learns the difference between patients with presence of heart disease against patients with absence, considering its attributes (blood pressure, heart rate, fasting blood sugar and etc). Logistic regression is designed for this purpose (classification), and it is most useful for understanding the influence of several independent variables on a single outcome variable.

Since the Cleveland dataset has limited statistics, I won’t try to predict all the categories of the heart disease status feature, but only whether or not heart disease is present (num>0 or num=0). The dependent variable higher than 0 will be transformed into 1, turning this into a binary LR problem. Using LR provides the advantage to generate a probability estimate for the prediction.

The classifier can be tuned in a few parameters, but we will be using **random state**, and **penalty**, as two parameters that are very related to how our accuracy score behaves.

The model will be trained with a split set applying cross-validation and considering values of 80% of the dataset for training and 20% for testing.

***2.4 Benchmark***

As a benchmark model for my classification problem, it is intended to use a single Naïve Bayes classifier in comparison to the Logistic Regression results and try to perform better with the proposed method.

It is considered to be a reasonable number for accuracy to be around 75% to 80% for this dataset, considering the number of samples and features.

**3. Methodology**

***3.1 Data Preprocessing***

Most of the preprocessing methods are centered on the python project inside the **pre\_process.py** file. Some of the preprocessing inside the notebook was made along the way specifically with the purpose of generating well formatted charts.

The following steps are considered:

* Reading the data and assigning column names to each feature
* Clean the data encoding the dependent variable
* Split the features and the dependent variable for better use
* Prepare the data for charts visualization, tuning names and values
* Prepare the data for csv export (logistic numbers and benchmarks)
* Encoding some features
* Looking for outliers and removing if necessary

***3.2 Implementation***

***3.3 Refinement***

**4. Results**

***4.1 Model Evaluation and Validation***

***4.2 Justification***

**5. Conclusion**

***5.1 Free-form Visualization***

***5.2 Reflection***

***5.3 Improvement***

**6. References**

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