Bruno de Oliveira

Um estudo sobre as características associadas ao desenvolvimento de doenças cardíacas utilizando modelos de regressão logística

Niterói - RJ, Brasil

26 de setembro de 2022

Conteúdo

Resumo

Em geral, as doenças cardíacas evoluem de maneira silenciosa no corpo durante anos, muitas vezes quando o paciente percebe os sintomas já é muito complicado resolver o problema. Detrano et al. (1989) usaram um modelo de função discriminante para estimar probabilidades de doença coronariana angiográfica em 3 grupos diferentes de pacientes. Outro estudo interessante é o de Aha e Kibler (1988), onde eles utilizam algoritmos de aprendizado de máquinas para prever doenças cardíacas. Este trabalho consiste em utilizar regressão logística para avaliar o efeito das covariáveis e o seu poder preditivo na prevalência de doenças cardíacas. Para isso foi utilizado diferentes modelos desta classe, entre eles um modelo utilizando o método LASSO para selecionar variáveis. O Modelo final escolhido possui apenas 5 covariáveis, entre os principais resultado podemos destacar o fato dos homens apresentarem 21% mais chances de desenvolver doenças cardíacas do que as mulheres, além dos indivíduos com talassemia reversível terem duas vezes mais chances de desenvolver doenças cardíacas do que os pacientes com talassemia normal.

1 Introdução

Segundo a Organização Mundial de Saúde (OMS), observa-se um aumento recente na carga de doenças cardiovasculares, principalmente em países de baixa e média rendas4, reflexo do aumento da expectativa de vida e, consequentemente, do maior tempo de exposição aos fatores risco para as doenças crônicas não transmissíveis (MENDIS et al., 2011). As doenças cardiovasculares são atualmente a principal causa de morte nos países em desenvolvimento, e espera-se que continue sendo a causa de mortalidade mais importante no mundo durante a próxima década (ABUBAKAR; TILLMANN; BANERJEE, 2015).

No Brasil, as doenças cardiovasculares são responsáveis por 27,7% dos óbitos, atingindo 31,8% quando são excluídos os óbitos por causa externas, sendo consideradas a principal causa de morte (MANSUR; FAVARATO, 2012). Sabe-se que alguns fatores que aumentam consideravelmente a prevalência de doenças cardíacas são o sexo, a idade, o índice de massa corporal, a presença de altos níveis de gorduras no sangue, a pressão arterial elevada diabetes (TESTON et al., 2016).

Este foi construído em duas etapas, na primeira foi realizada uma análise descritiva do conjunto de dados através de análises gráficas buscando identificar variáveis que tem relação com a prevalência de doenças cardíacas. Na segunda etapa, foi realizado o ajuste de modelos de regressão logística, inicialmente foi ajustado um modelo utilizando a técnica LASSO para selecionar as variáveis com os possíveis melhores poderes preditivos para explicar o desenvolvimento de doenças cardíacas, depois foi ajustado um modelo de regressão logística considerando todo o conjunto de variáveis explicativas presentes no conjunto de dados. Por fim, um modelo reduzido foi ajustado utilizando apenas as variáveis que apresentaram estimativas significantes.

Então, utilizamos medidas de qualidade de ajuste para avaliar os modelos ajustados, essas medidas foram a *deviance*, o *Akaike Information Criterion* (AIC), o *Bayesian information Criterion* (BIC) e os *resíduos de Pearson* r_i.

2 Materiais e Métodos

Nesta seção vamos apresentar a base de dados utilizada e a metodologia referente ao modelo de regressão logística, a técnica de seleção de variáveis LASSO e as medidas estatísticas utilizadas para avaliar a qualidade de ajuste dos modelos ajustados.

2.1 Descrição da base de dados

O conjunto de dados utilizado neste trabalho é composto por 14 variáveis clínicas de referentes à 270 pacientes do banco de dados da cidade de Cleveland nos Estados Unidos. A seguir, é apresentado o dicionário das variáveis utilizadas:

- **Idade:** Idade do paciente.
- **Sexo:** Variável binária que indica o sexo. Categoria de referência é mulher.
- **Tipo de dor no peito:** Variável categórica que indica o tipo de dor no peito do paciente. Categoria de referência é angina típica.
- **Pressão sanguínea em repouso:** Variável numérica. Pressão arterial em repouso medida na chegada ao hospital em mmHg.
- **Colesterol:** Quantidade de colesterol total na corrente sanguínea em mg/dl.
- **Diabetes:** Variável binária que indica se o indivíduo sofria de diabetes.
- **Resultados dos eletrocardiográficos:** Variável categórica que indica o resultado do eletrocardiográfico. Categoria de referência é normal.
- **Frequência cardíaca:** Variável numérica com os valores da frequência cardíaca máxima alcançada pelo paciente.
- **Angina de exercício:** Variável binária que indica se o paciente tem angina induzida por exercício. Categoria de referência é não.

- **Depressão do ST:** Variável numérica. Depressão do ST induzida por exercício em relação ao repouso.
- **Declive do ST:** Variável categórica que indica a inclinação do segmento ST no pico do exercício. Categoria referência é inclinação ascendente.
- **Número de vasos principais:** Variável numérica. Número de vasos principais coloridos por fluoroscopia (0-3).
- **Talassemia:** Variável categórica que indica o tipo de talassemia do paciente. Categoria de referência é normal.
- **Doença cardíaca:** Variável binária que indica se o paciente tem Doença cardíaca.

2.2 Modelo de Regressão Logística

Esse tipo de modelo é amplamente usado para analisar dados envolvendo respostas binárias ou binomiais. Como o objetivo deste trabalho é identificar as variáveis que estão associadas ao indivíduo ter desenvolvido uma doença cardíaca, podemos definir uma variável Y_i que indica se o i-ésimo paciente da base de dados tem doença cardíaca,

$$Y_i = \begin{cases} 1 & \text{, se o i-ésimo indivíduo tem doença cardíaca,} \\ 0 & \text{, caso contrário;} \end{cases}$$

Como essa variável é do tipo binária, podemos utilizar a distribuição Bernoulli para modelar a mesma, ou seja, $Y_i \sim Bernoulli(\pi_i)$. Logo a probabilidade de um indivíduo ter doença cardíaca é $P(Y_i = 1) = \pi_i$ e a de não ter doença cardíaca é $P(Y_i = 0) = 1 - \pi_i$.

O modelo de regressão logística é um tipo de modelo linear generalizado (MLG) que possui esse nome devido ao fato dele usar a função de ligação logito, para ligar o parâmetro π_i a um preditor linear composto pelas variáveis explicativas e pelo coeficientes associados a essas variáveis, que são os parâmetros a serem estimados pelo modelo.

Em geral o modelo de regressão logística é representado pela equação

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta, \qquad Y_i \sim Bernoulli(\pi_i), i = 1, \ldots, N \tag{2.1}$$

em que $x_i = (1, x_{i1}, \ldots, x_{ip})^T$ é o vetor de variáveis explicativas do i-ésimo indivíduo e $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros estimados pelo modelo.

2.3 Modelo LASSO

A regressão de LASSO é o que chamamos de método de regressão penalizado, usado frequentemente em aprendizagem de máquinas para selecionar um subconjunto de variáveis. O LASSO, faze uso de uma estrutura de penalização na função de log-verossimilhança $l(\beta; y)$ que encolhe a estimativa dos coeficientes β não significativos para zero. Este método foi proposto Tibshirani (1996) e suas estimativas são obtidas a partir da seguinte equação:

$$arg\max_{\beta}\left\{ l(\beta; y) - \lambda \sum_{i=1}^{p} |\beta_i| \right\}, \tag{2.2}$$

em que λ é um parâmetro que controla o quanto o termo de penalização afeta as estimativas, ou seja, λ grande resulta em uma alta penalização e λ pequeno resulta em uma penalização mais baixa. Este parâmetro λ é estimado por meio de validação cruzada.

2.4 Medidas de qualidade de ajuste

Para avaliar a qualidade de ajuste e a capacidade de predição dos modelos afim de se identificar o modelo que melhor ajustou aos dados foram consideradas a deviance, a estatística qui-quadrado de Pearson, o AIC e o BIC. Para avaliar a capacidade preditiva dos modelos foi medida a acurácia, que foi obtida calculando a proporção de classificações corretas realizadas pelo modelo, ou seja, a proporção de vezes que o modelo classificou uma pessoa que tinha doença cardíaca, quando ela realmente tinha doença e classificou como não tendo doença cardíaca, um indivíduo que realmente não tinha doença cardíaca.

A deviance no caso do modelo de regressão logística, pode ser calculada da seguinte forma

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{(n_i - y_i)}{(n_i - \hat{y}_i)}\right)\right], \tag{2.3}$$

em que $\hat{y}_i = \hat{\pi}_i$,já que no contexto desse problema $n_i = 1$, para todo i. As probabilidade estimada de π_i são obtidas a partir da equação

$$\hat{\pi_i} = \frac{\exp(x_i^T \hat{\beta})}{1 + \exp(x_i^T \hat{\beta})} \tag{2.4}$$

Com base nas estimativas $\hat{\pi_i}$, pode-se adotar um critério para classificar se um indivíduo tem doença cardíaca ou não. Existem diversas técnicas que permitem escolher o critério mais interessante para fazer essa classificação. Porém nesse trabalho o critério escolhido

foi o mais simples e natural, que é apresentado a seguir:

$$C_i = \begin{cases} 1 & \text{, se } \hat{\pi}_i > 0,5; \\ 0 & \text{, se } \hat{\pi}_i \leq 0,5. \end{cases}$$

Deste modo, pode-se avaliar a acurácia dos modelos ajustados, avaliando a proporção de observações que foram classificadas corretamente.

A estatística qui-quadrado de Pearson pode ser calculada usando a sua relação com os resíduos de Pearson, $X^2 = \sum_{i=1}^{N} r_i^2$, onde

$$r_i^2 = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{2.5}$$

O *AIC* e o *BIC* são as últimas medidas de qualidade de ajuste utilizadas na comparação dos modelos, essas medidas são baseadas na verossimilhança, com penalização baseada no número de parâmetros e no número de observações. Essas estatísticas são dadas por:

$$AIC = -2l(\hat{\beta}; y) + 2p, \tag{2.6}$$

$$BIC = -2l(\hat{\beta}; y) + p\log(N). \tag{2.7}$$

em que p é o número de parâmetros estimados pelo modelo e N é o número de observações.

3 Análise dos Resultados

Na Figura 1 é possível notar que que a variável Diabetes, não parece influenciar na chance da pessoal ter ou não doença cardíaca, já nas variáveis Angina induzida por exercício e Sexo, pode-se notar uma relação em quem tem angina ou é do sexo masculino com ter doença cardíaca.

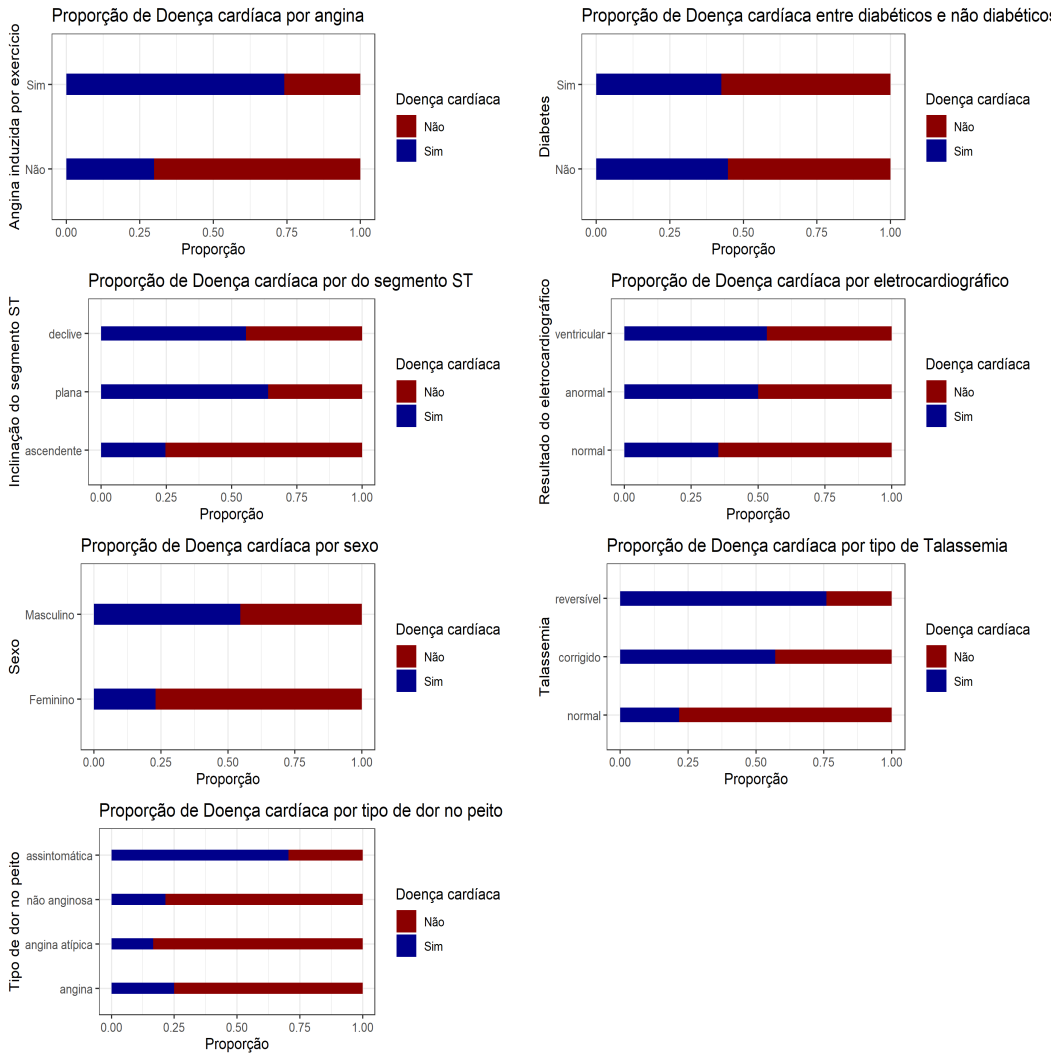

Figura 1: Variáveis explicativas categóricas pela variável resposta.

A categoria inclinação plana do segmento ST apresentou alguma relação a prevalência de doença cardíaca, assim como a categoria declive da mesma variável, apesar de ser uma proporção menor do que a citada anteriormente, entre os tipos de dor no peito, os indivíduos com dor assintomática, assim como, os pacientes com talassemia reversível e corrigida apresentaram relação proporção maiores nos indivíduos com doença cardíaca. Os resultados dos eletrocardiográficos mostram que com exceção daqueles pacientes que tiverem resultado mostrando provável hipertrofia ventricular esquerda.

Na Figura 2, é possível observar a pressão arterial dos indivíduos se assemelha entre os que foram diagnosticados com doença cardíaca ou não, entretanto a distribuição dos pacientes com doença cardíaca é maior que a dos que não tem, que por sua vez é mais concentrada em torno de sua média, apesar de possuir um *outlier* a mais do que os do outro grupo. Nas variáveis Idade, Depressão do segmento ST, Colesterol e número de de vasos principais coloridos por fluoroscopia apresentaram valores médios maiores nos grupos de pacientes com doença cardíaca, a idade por exemplo, apresentou média de quase 53 anos no grupo de pessoas que não tem doença cardíaca, enquanto os pacientes com doença cardíaca apresentaram idade média de mais de 56 anos, números bem próximos, o que pode indicar que não existe uma forte relação entre essa variável e os pacientes com doença cardíaca nessa base de dados. O Colesterol foi outra variável que não apresentou grandes diferenças médias nos dois grupos, assim como a idade, esse resultado pode indicar que essa variável nem tem muito poder de preditivo no nosso estudo. Por fim, a variável Frequência cardíaca, apresentou uma média maior nos indivíduos do grupo que não possui doença cardíaca. Neste presente trabalho foram ajustados 3 modelos, um modelo de regressão logística com todas as variáveis explicativas disponíveis na base de dados (Modelo completo), um outro modelo regressão logística considerando apenas as variáveis com efeitos significativos na variável resposta (Modelo reduzido) e por último um modelo usando o método de seleção de variáveis LASSO (Modelo LASSO). Para realizarmos os ajustes dos modelos, dividimos aleatoriamente a base de dados em duas partes, uma parte com 75% dos dados para treino e a outra com 25% dos dados para teste, a amostra treino ficou com 203 observações enquanto a amostra teste ficou com 67 observações.

Na Tabela 2 é apresentado as medidas de qualidade de ajuste de todos os três modelos ajustados para comparar a adequabilidade dos modelos, afim de se escolher o modelo melhor ajustado para esses dados.

Nota-se que o modelo reduzido apresentou menores AIC e BIC, já o modelo completo apresentou a menor deviance, um resultado natural, já que este modelo é o que considera

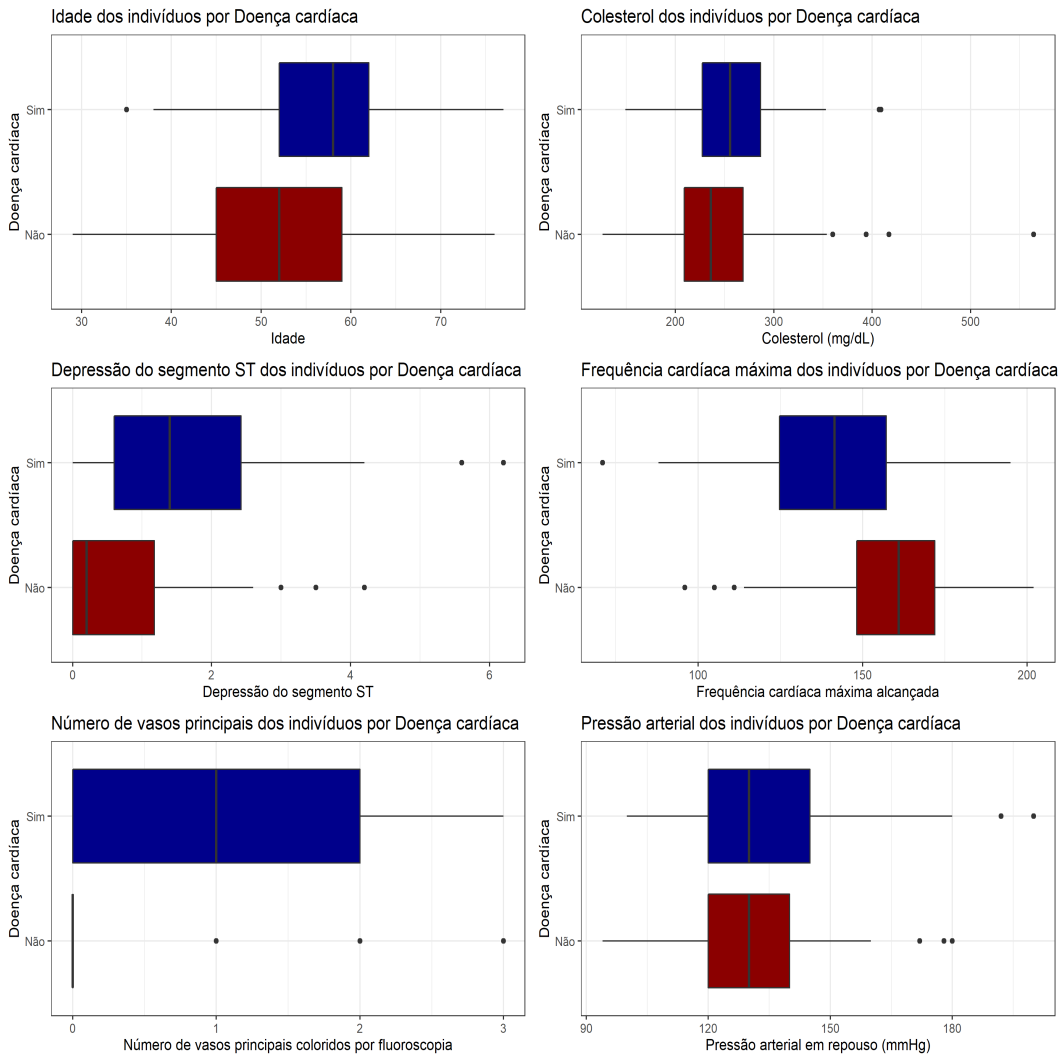

Figura 2: Variáveis explicativas numéricas pela variável resposta.

Modelos	Deviance	Pearson	AIC	BIC	Acurácia
Modelo completo	**107,100**	303,387	145,100	191,485	**74,63%**
Modelo reduzido	119,615	270,409	**137,615**	**167,434**	70,15%
Modelo LASSO	134,132	**114,282**	143,568	189,952	73,13%

Tabela 1: Estatísticas de qualidade de ajuste e acurácia da predição dos pacientes com doença cardíaca pelos três modelos ajustados.

o maior conjunto de variáveis explicativas. O modelo LASSO apresentou o menor valor da qui-quadrado de Pearson. Então por ter apresentado o menor AIC e BIC, além de um valor qui-quadrado de Pearson menor do que o do modelo completo e ser um modelo mais parcimonioso, já que tem um conjunto menor de variáveis explicativas que o modelo completo, o modelo reduzido foi o escolhido como modelo final.

Na Tabela são apresentadas as estimativas dos coeficientes do modelo final, nela é possível notar que todas as variáveis apresentaram efeito significativo.

Variável	Estimativa	P-valor	Razão de Chance	$IC_{95\%}$
Intercepto	-4,486	<0,001	-	-
Angina induzida por exercício	1,758	0,001	5,802	$[2,025; 17,990]$
Dor no peito assintomática	1,865	0,031	6,456	$[1,277; 40,324]$
Número de vasos principais	1,791	<0,001	5,995	$[3,142; 12,905]$
Sexo	1,211	0.037	3,357	$[1,106; 10,982]$
Talassemia reversível	2,487	<0,001	12,021	$[4,283; 37,612]$

Tabela 2: Estimativas dos coeficientes, p-valor, razão de chances e intervalo de confiança de 95% para as variáveis significativas.

Além disso, podemos observar que os indivíduos do sexo masculino tem mais de duas vezes mais chances de ter doença cardíaca do que as mulheres, também podemos ver que indivíduos que tem angina induzida por exercício tem quase seis vezes mais chances de ter doença cardíaca do que os que não tem. Observa-se que, as chances de um paciente com talassemia reversível ter doença cardíaca é doze vezes maior do que os pacientes com talassemia normal, enquanto que, pacientes que apresentaram dor no peito assintomática tem mais de seis vezes chances de ter doença cardíaca do que pacientes. Por fim, um indivíduo com dois vasos principais coloridos por fluoroscopia tem quase seis vezes mais chances de ter doença cardíaca do que o individuo com um vaso colorido apenas.

4 Conclusões

Normalmente a idade e o sexo são fatores associadas ao o indivíduo ter doenças cardíacas, como mostra Teston et al. (2016), mas nesse estudo a idade não apresentou efeito significativo para a prevalência de doenças cardíacas, talvez isso possa estar relacionada a características especificas dessa base de dados.

O objetivo desde trabalho consistia em utilizar regressão logística para avaliar a relação das características de um paciente com ele possuir ou não doença cardíaca, afim de identificar os fatores mais relevantes associadas ao indivíduo essa doença. Para isso foram considerados três modelos, o modelo completo, o reduzido e o LASSO. Para selecionar o melhor modelo, utilizamos medidas de qualidade ajuste e o poder preditivo do modelo foi estimado usando acurácia, estudos como o de Aha e Kibler (1988) apresentam métodos de classificação mais sofisticados e adequados para as análises feitas aqui nesse trabalho do que o próprio método aqui utilizado.

O modelo final escolhido foi o reduzido, que considerava apenas as variáveis, Angina induzida por exercício, Tipo de dor no peito, Número de vasos principais coloridos pelo fluoroscopia, Sexo e Talassemia. A partir deste modelo foi observado uma chance maior de ter doença cardíaca em indivíduos do sexo masculino, ou que tem talassemia reversível, angina induzida por exercício e dor no peito assintomática e cada vaso colorido pela fluoroscopia a chance do individuo ter doença cardíaca aumenta em quase seis vezes. Esses resultados são consistentes se comparados a outros estudos já apresentados pelos demais autores, evidenciando a importância entre estes fatores e a prevalência de doença cardíaca.

Referências

ABUBAKAR, I.; TILLMANN, T.; BANERJEE, A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet*, v. 385, n. 9963, p. 117–171, 2015.

AHA, D.; KIBLER, D. Instance-based prediction of heart-disease presence with the cleveland database. *University of California*, v. 3, n. 1, p. 3–2, 1988.

DETRANO, R. et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, Elsevier, v. 64, n. 5, p. 304–310, 1989.

DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: CRC press, 2018.

MANSUR, A. d. P.; FAVARATO, D. Mortalidade por doenças cardiovasculares no brasil e na região metropolitana de são paulo: atualização 2011. *Arquivos brasileiros de cardiologia*, SciELO Brasil, v. 99, p. 755–761, 2012.

MENDIS, S. et al. *Global atlas on cardiovascular disease prevention and control*. [S.l.]: World Health Organization, 2011.

TESTON, E. F. et al. Fatores associados às doenças cardiovasculares em adultos. *Medicina (Ribeirão Preto)*, v. 49, n. 2, p. 95–102, 2016.