Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Edson Luiz dos Santos

Modelos preditivos para Doença Cardiovacular

Curitiba
2021

Edson Luiz dos Santos

Modelos preditivos para Doença Cardiovacular

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Professor Dr. Wagner Hugo Bonat

Curitiba
2021

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Modelos preditivos para Doença Cardiovacular

Edson Luiz dos Santos[1]

Professor Dr. Wagner Hugo Bonat[2]

Resumo

Este estudo foi realizado com o intuito de auxiliar a identificação antecipada de pessoas que tenham tendências de contrair doença cardiovascular.A base de dados utilizada neste estudo, é uma base pública disponível no kaggle, ela possui 918 registros com 11 variáveis explicativas e 1 variável resposta. A primeira parte do processo, foi efetuar uma análise exploratória das variáveis e identificar quais possuem maior influência na causa de doenças cardiovasculares, e quais variáveis podiam ajudar a prever quais pacientes podem desenvolver doença cardiovascular, além de identificar quais são os principais causadores deste tipo de doença. A segunda parte do processo, para responder os fatores da pesquisa, foram utilizadas as metodologias, modelo logístico, arvore de decisão e *random forest*, estas metodologias foram escolhidas por possuírem características de modelos binários, ou seja, modelos de classificação. A terceira parte do processo, foi utilizada para testar a acurácia das técnicas aplicadas, os modelos são capazes de realizar a predição da doença, conforme as características do conjunto de dados. Todos os testes resultaram em uma excelente assertividade, o modelo logístico obteve a melhor performance, atingiu aproximadamente 87%, por se tratar de um modelo mais simples e a melhor acurácia, o modelo logístico foi o escolhido para auxiliar na predição deste tipo de doença.

Palavras-chave: Análise exploratória, Modelo Logístico, Árvore de decisão, *Random Forest*, Acurácia, Predição.

Abstract

This study was carried out with the aim of helping the early identification of people who have a tendency to contract heart disease. The database used in this study is a public database available on kaggle, there are 918 records with 11 explanatory variables and 1 response variable. The first part of the process was to carry out an exploratory analysis of the variables and identify which ones have the greatest influence on the cause of heart disease, and which variables could help predict which patients may develop heart disease, in addition to identifying which are the main causes of this type of disease. The second part of the process, to answer the research factors, were used the methodologies, logistic model, decision tree and random forest, these methodologies were chosen because they have characteristics of classification models. The third part of the process was used to test the accuracy of the applied techniques, the models are able to predict the disease, according to the characteristics of the data set. All tests resulted in excellent assertiveness, the logistic model obtained the best performance, reaching approximately 87%, because it is a simpler model and the best accuracy, the logistic model was chosen to assist in the prediction of this type of disease.

Keywrods: *Exploratory Analysis, Logistic Model, Decision Tree, Random Forest, Accuracy, Prediction.*

I Introdução

Neste artigo, serão apresentados os estudos efetuados, bem como as metodologias utilizadas para obter um resultado que forneça uma boa capacidade de predição da doença cardiovascular. Este estudo tem como objetivo descrever o relacionamento das 11 variáveis explanatórias do conjunto de dados, buscando a criação de uma ferramenta capaz de prever se o indivíduo possui fatores de risco ou poderá desenvolver doença cardiovascular.

Doença Cardiovascular é um problema mundial, aproximadamente 18 milhões de pessoas morrem anualmente por causa desta doença, a qual é a principal causa das mortes, e isto não é diferente para os brasileiros. Indivíduos que possuem a doença ou que tenham alto risco cardiovascular, devido presença de problemas de saúde, como pressão alta, colesterol alto ou diabetes ou até mesmo um conjunto destes problemas.

Estes problemas precisam de uma detecção o quanto antes, a detecção antecipada, poderá evitar ou postergar grande parte destas mortes. Baseado no resultado, o modelo com a prevenção e o tratamento adequado dos fatores de risco e da doença cardiovascular, podemos reverter essa grave situação.

Os dados serão analisados utilizando modelos supervisionados e serão avaliados com os modelos logístico [1], arvore de decisão [2], e *random forest* [3]. Inicialmente a base foi dividida em 2 partes, sendo 80% para efetuar o treino e 20% para testar a efetividade. Os modelos serão ajustados pela base de treino e a acurácia será avaliada com a base de teste. A base de dados possui 12 variáveis e 918 registros, sendo uma a variável resposta, em uma análise superficial já é possível

[1]Edson Luiz dos Santos, eluiksantos@gmail.com.
[3]Professor Dr. Wagner Hugo Bonat - DEST/UFPR.

perceber que a grande maioria dos casos, ocorre em pessoas do sexo masculino.
A seção 2 deste artigo, apresenta uma descrição detalhada da base de dados utilizada neste estudo. Na seção 3, encontramos informações sobre as metodologias aplicadas no conjunto de dados para obter o resultado. A seção 4, possui informações sobre os resultados atingidos com a aplicação das metodologias. A seção 5 demonstra uma comparação entre as metodologias aplicadas, e os resultados obtidos.

2 Conjunto de dados

A base de dados foi capturado e está disponível no Kaggle, possui 12 variaveis, Doença Cardiovascular é a variável resposta, as demais variáveis são: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak e ST_Slope.

Descrição variáveis	Nome original	Sigla
idade	Age	ida
sexo	Sex	sex
tipo dor no peito	ChestPainType	tdp
pressão sanguínea em repouso	RestingBP	psr
colesterol	Cholesterol	col
açúcar no sangue em jejum	FastingBS	asj
resultados eletrocardiograma em repouso	RestingECG	rer
frequência cardíaca máxima alcançada	MaxHR	fcm
angina induzida por exercício	ExerciseAngina	aie
oldpeak = ST - depressão do segmento ST	Oldpeak	dsST
a inclinação do segmento ST de pico do exercício	ST_Slope	isST
Doença cardíaca	HeartDisease	doc

Tabela 1: Esta tabela possui a descrição das variáveis do conjunto de dados, a partir deste ponto, as variáveis serão referenciadas por sua sigla.

2.1 Tratamento dos dados

Neste estudo foi identificado que algumas variáveis devem ser transformadas em variáveis categóricas, as variáveis sex, tdp, rer, asj, aie e isST, foram as variáveis que sofreram esta alteração. A variável col foi excluída deste estudo, ela possui muitos registros com valor zero, principalmente para indivíduos que possuem a doença, isto poderia distorcer os resultados.

2.2 Analise exploratória

Análise das medidas estatísticas das variáveis quantitativas.

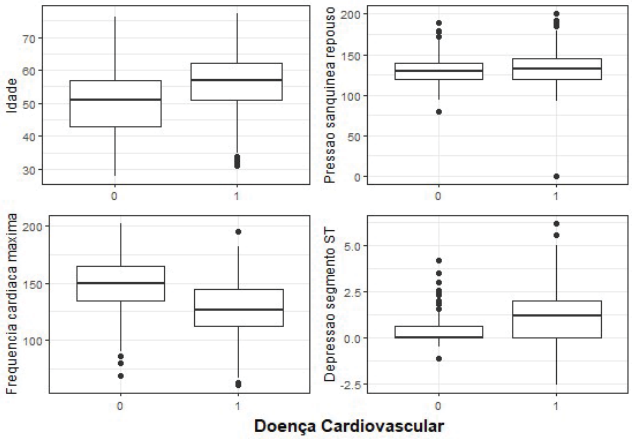

Figura 1: Boxplots de ilustração das medidas estatísticas.

Os gráficos da figura 1, apresentam a distribuição da doença cardiovascular por idade. Esta visão demonstra que as doenças ocorrem em pessoas com idade mais avançada. Medidas de pressão quando o indivíduo está em repouso (psr), apresenta uma pressão media aproximada, porém quando existe a doença podemos observar alguns picos de pressão sanguinea. Frequência cardíaca máxima (fcm), note que os pacientes com a doença, possuem uma frequência cardíaca menor. Em uma entrevista com um especialista da área, ele informou que possivelmente estes pacientes tomam uma medicação para controlar a frequência cardíaca, a documentação da base não possui essa informação. Variação no resultado do eletrocardiograma no momento da depressão do segmento ST (dsST).

Análise da distribuição percentual das convariáveis classificatórias.

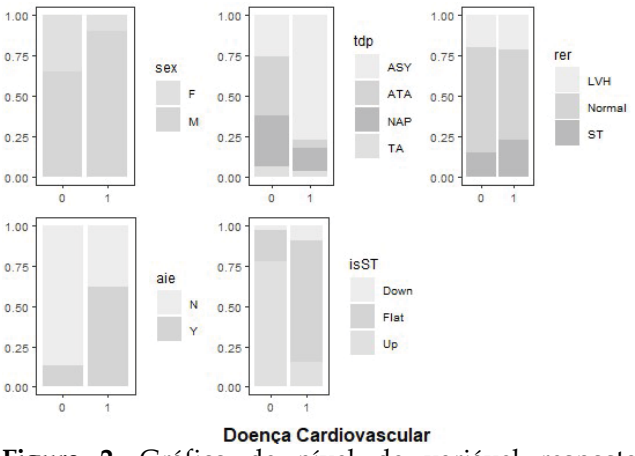

Figura 2: Gráfico de nível de variável resposta e covariáveis.

Representam a distribuição percentual das covariáveis classificatórias.
Os gráficos da figura 2, demonstram que os indivíduos do sexo masculino, possuem uma tendência maior de contrair a a doença cardiovascular. Quando que o tipo de dor no peito ASY, possui uma enorme influência na causa da doença cardiovascular. Distribuição do resultado do eletrocardiograma, quando o indivíduo está em repouso. Distribuição da doença por tipo de dor no peito, quando

a dor no peito é induzida pelo exercício. Maior incidência ocorre quando a inclinação do segmento ST de pico do exercício é igual a Flat.

3 Metodologia

Com o objetivo de efetuar uma predição de indivíduos doentes ou sãos, com base nas variáveis do conjunto de dados Doença Cardíaca, foram utilizadas três metodologias diferentes, visando a identificação da melhor técnica para efetuar a predição da doença.
Para aplicar as técnicas, a base foi dividida em duas partes, sendo uma base para treino e uma base para teste, a base de treino possui 80% do total de registros, o restante, 20%, será utilizado para efetuar os testes de eficiência dos modelos.
A função glm do software R [*1*] foi utilizada para ajustar o modelo logístico, a família de distribuição utilizada, foi a Binomial, esta família é utilizada para dados binários. O link Logit foi o link escolhido para a família Binomial.
A regressão logística é uma poderosa técnica estatística, e tem como objetivo, criar um modelo de predição de resultado, a partir de uma série de variáveis explicativas, mostrando a relação entre os recursos e calculando a probabilidade de um determinado resultado [*8*].
A regressão logística é usada no aprendizado de máquina (ML), auxiliando no processo previsões precisas. É semelhante à regressão linear, exceto que, em vez de um resultado gráfico, a variável resposta é binária, 0 ou 1.
Existem dois tipos de mensuráveis, as variáveis explanatórias e a variável resposta, a qual é o alvo a ser atingido.
Esta técnica é fortemente utilizada na área da medicina, visando a identificação de indivíduos doentes ou sãos, e será utilizada para efetuar a predição de Doença cardiovascular, a partir do conjunto de dados.
Para o modelo 2, a função rpart do software R [*2*], foi utilizada para aplicar a metodologia de modelo de árvore de decisão.
Uma árvore de decisão usa uma estrutura de árvore para representar um número de possíveis caminhos de decisão e um resultado para cada caminho, esta técnica gera um gráfico no formato de árvore, demonstrando visualmente as condições e as probabilidades para chegar aos resultados desejados.
O algoritmo utilizado para a representação visual da árvore, é do grupo de aprendizado de máquina supervisionado, sua aplicação serve para classificação (variável-alvo categórica) ou para regressão (variável-alvo contínua), funcionando para variáveis de entrada e saída categóricas e contínuas.
Este algoritmo efetua várias divisões dos dados, gerando subconjuntos, de tal forma que os subconjuntos vão ficando cada vez mais puros, ou seja, contendo menos classes ou apenas uma da variável resposta.
Elas possuem fácil entendimento e interpretação, o processo por onde chegam em uma previsão é completamente visível [*9*].
Para o modelo 3, a função *random forest* [*3*] do software foi utilizada.
Random forest, o próprio nome explica como este algoritmo funciona. Ele cria muitas árvores de decisão, de maneira aleatória, gerando uma floresta, sendo que para a construção de cada uma dessas árvores, os dados não são utilizados em sua totalidade. Algumas amostras dos dados serão selecionadas de maneira aleatória, por um método de reamostragem que permite amostras repetidas na seleção. Para criação dos nós das árvores, também existirá uma etapa aleatória, onde algumas variáveis serão selecionadas de forma randômica.
Sua principal característica, é a combinação de diferentes modelos para chegar em um único resultado. Essa característica torna esses algoritmos mais robustos e complexos, levando a um maior custo computacional que costuma ser acompanhando de melhores resultados [*3*].
Para avaliar a precisão dos modelos, algumas técnicas serão utilizadas.
A área sob a curva ROC é uma medida de qualidade preditiva do modelo. A curva que se forma após plotarmos valores no gráfico, o melhor teste é quando os valores estão próximos de 1, isto indica modelos com elevada capacidade preditiva [*10*]. A matriz de confusão é uma tabela de fácil entendimento, facilmente podemos identificar os quatro tipos de classificação do modelo de classificação binário [*11*]. A diagonal principal da matriz de confusão [*11*] demonstra os resultados preditos corretamente, quando estes valores são comparados com a variável resposta do conjunto de dados. A acurácia é a precisão do modelo, ela apresenta o percentual de poder preditivo do modelo [*10*].

4 Resultados

Regressão Logística: Os modelos foram ajustados com 80% dos registros do conjunto de dados, a função glm do software R [*1*], foi utilizada para aplicar a metodologia de modelo logístico.
A fórmula inicial utilizada foi aplicada com todas as 11 variáveis do conjunto de dados, a partir deste primeiro modelo, a função stepwise [*8, 12*] foi utilizada para identificar a melhor fórmula que será utilizada nesta pesquisa. O resultado apresentou as variáveis ida, sex, tdp, asj, aie, dsST, isST, sendo estas, as variáveis preditoras que serão utilizadas no modelo logístico.
O modelo logístico apresentou resultados eficientes, sendo capaz de efetuar uma ótima predição.

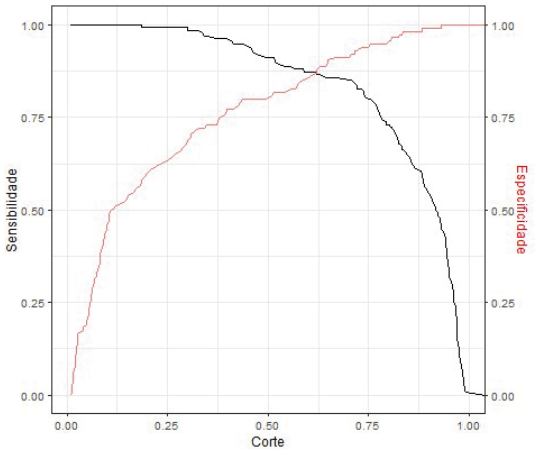

Figura 3: Apresenta o melhor ponto de corte.

O modelo apresentou uma área sob a curva de aproximadamente 95%
A curva ROC foi utilizada para identificar a curva ótima, o ponto que possui o valor mais próximo de 1.
A área sob a curva é de 95,69, o modelo proposto apresenta um intervalo de confiança de 95%, entre 93,57 e 97,82.

Valor Previsto/Valor Real	Positivo	Negativo
Positivo	100	18
Negativo	15	116

Tabela 2: Matriz de confusão.

Valor acurácia
A matriz diagonal principal apresenta os acertos do modelo preditivo (100 + 116 = 216).
A matriz diagonal secundaria apresenta os erros do modelo preditivo (15 + 18 = 33).
O Valor da acurácia é apurado pelo total da diagonal principal, dividido pelo total das diagonais, 216 / 249, esta divisão resulta em uma acurácia de 86,75%.

Arvore de decisão:

Os modelos foram ajustados com a base de treino contendo 80% dos registros do dataset, a função rpart do software R [2], foi utilizada para aplicar a metodologia de modelo de árvore de decisão.

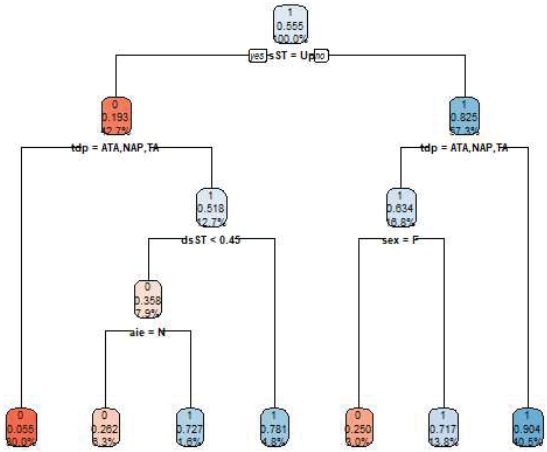

Figura 4: Estrutura da árvore de decisão.

Esta árvore confirma que a variável Inclinação do segmento ST (isST), deve ser classificada como o primeiro critério para identificar a doença, quando o registro desta variável possui o valor UP, temos uma probabilidade de 19%(0,193) de que o indivíduo não terá a doença, esta característica é encontrada em 42% dos indivíduos do conjunto de dados.
Quando o tipo de dor no peito (tdp) é diferente de ATA, NAP e TA, a probabilidade do paciente não ter a doença é de 5,5% (0,055), isto ocorre para 30% dos indivíduos.
Quando o tipo de dor no peito (tdp) é igual a ATA, NAP e TA, a probabilidade do indivíduo obter a doença é de 51% (0,518), isto ocorre para 12,7% dos indivíduos.
A próxima variável é a Depressão do Segmento ST (dsST), quando o resultado apresenta um valor menor que 0,45, a probabilidade do indivíduo ter a doença é de 78%(0,781), isto ocorre para 4,8% dos indivíduos
Quando o resultado da variável Depressão do Segmento ST (dsST), for maior que 0,45, a probabilidade do indivíduo não ter a doença é de 35%(0,358), isto ocorre para 7,9% dos indivíduos.
Quando o valor da variável dor induzida no peito (aie) for igual a N, a probabilidade do indivíduo ter a doença é de 72% (0,727), isto acontece para 1,6% dos indivíduos.
Quando o valor da variável dor induzida no peito (aie) for diferente de N, a probabilidade da doença não acontecer é de 26% (0,262), isto ocorre para 6,3% dos indivíduos.
Para a variável inclinação do segmento ST (isST) diferente de UP, a probabilidade da doença cardíaca acontecer é de 82% (0,825), isto ocorre para 57% dos indivíduos.
O próximo nível apresenta a variável tipo de dor no peito (tdp), quando os valores forem iguais a ATA, NAP e TA, a probabilidade da doença acontecer é de 90% (0,904), isto ocorre para 40,5% dos indivíduos.
Ainda quando o Tipo de dor no peito (tdp) forem iguais a ATA, NAP e TA, existe a probabilidade de 63% (0,634) da doença acontecer, 16,8% dos indivíduos, sendo que a probabilidade da doença acontecer quando a variável Sexo (sex) for igual a F, é de 71% (0,717), isto ocorre para 13,8% dos indivíduos.

Quando o valor da variável Sexo (sex) for diferente de F, a probabilidade da doença não acontecer é de 25% (0,250), isto acontece para 3% dos indivíduos.

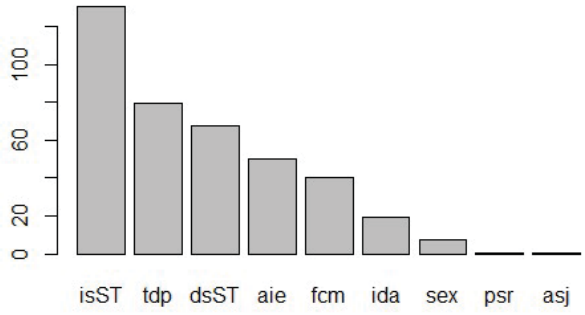

Figura 5: Análise da Importância das variáveis

A ilustração apresenta as variáveis em ordem decrescente, conforme sua importância.

Valor Previsto/Valor Real	Positivo	Negativo
Positivo	87	26
Negativo	15	123

Tabela 3: Apresenta a matriz de confusão com o resultado do modelo Árvore de decisão.

Valor acurácia
A matriz diagonal principal apresenta os acertos do modelo preditivo (87 + 123 = 210).
A matriz diagonal secundaria apresenta os erros do modelo preditivo (15 + 26 = 41).

O Valor da acurácia é apurado pelo total da diagonal principal, dividido pelo total das diagonais, 210 / 251, esta divisão resulta em uma acurácia de 83,66%.

Random forest

A terceira metodologia avaliada foi a r*andom forest*, onde as mesmas variáveis foram utilizadas, 80% dos registros foram utilizados como base de treino, o ajuste do modelo foi efetuado com a função *random forest* [3].

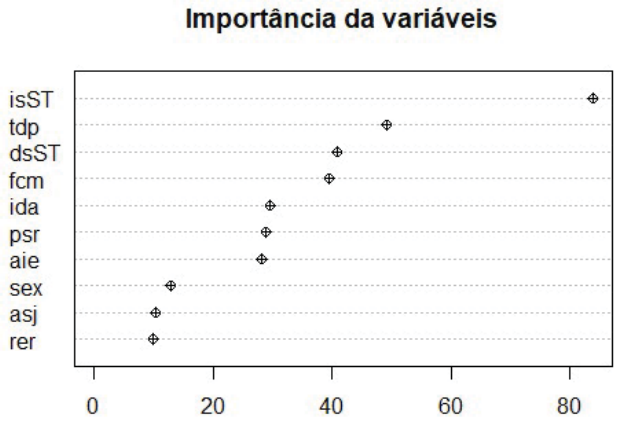

Figura 6: Análise da Importância das variáveis

A ilustração apresenta as variáveis em ordem decrescente, conforme sua importância.

Matriz de Confusão

Valor Previsto/Valor Real	Positivo	Negativo
Positivo	78	12
Negativo	24	115

Tabela 4: Apresenta a matriz de confusão com o resultado do modelo *random forest*.

Valor acurácia

A matriz diagonal principal apresenta os acertos do modelo preditivo (78 + 115 = 193).
A matriz diagonal secundaria apresenta os erros do modelo preditivo (24 + 12 = 36).O Valor da acurácia é apurado pelo total da diagonal principal, dividido pelo total das diagonais, 193 / 229, esta divisão resulta em uma acurácia de 84,27%.

5 Comparação

Comparação das 3 matrizes de confusão e acurácia em ordem decrescente.

Metodologia	Previsto/Real	Pos	Neg	Acurácia
Regressão logística	Positivo	100	18	86,75%
	Negativo	15	116	
Random forest	Positivo	78	12	84,28%
	Negativo	24	115	
Árvore de decisão	Positivo	87	26	83,67%
	Negativo	15	123	

Tabela 5: Comparação da acurácia entre as metodologias.

Na tabela 5, podemos visualizar que a eficácia do modelo logístico foi ligeiramente superior aos demais modelos.

Discussão

Durante o desenvolvimento algumas limitações foram encontradas. A base de dados não possui documentação. Muitos registros da variável colesterol estavam com o valor zerado, por este motivo foi decidido retirar esta variável para a aplicação das metodologias.
O objetivo inicial foi analisar as variáveis e identificar as que iriam auxiliar na predição de doenças cardiovasculares, devido muitos registros estarem com o valor zerado, a variável colesterol não foi utilizada no ajuste dos modelos. Todas as demais variáveis foram utilizadas para as técnicas de árvore de decisão e *random forest*.
No de modelo Logístico, foi aplicado a função stepwise, esta função retornou a melhor formula para esta técnica.
A acurácia dos 3 modelos foi satisfatória, todas acima de 83%, dentre as 3 metodologias aplicadas, o modelo logístico foi o que obteve melhor resultado, com 86,75% de assertividade.

Agradecimentos

Tenho um enorme agradecimento ao Professor Wagner Hugo Bonat, o qual foi meu professor durante o período da especialização e orientador durante o periodo de desenvolvimento deste estudo, também agradeço a minha esposa Adriana Kovalski, sempre apoiando e incentivando durante todo o período deste curso.

Referências

[1] Dobson, A. J. (1990) *An Introduction to Generalized Linear Models. London: Chapman and Hall.*

[2] Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees. Wadsworth.*

[3] Breiman, L. (2001), *Random Forests, Machine Learning* 45(1), 5-32.

[4] ARNULF, Grübler (1990). *The Rise and Fall of Infrastructures: Dynamics of Evolution and Technological Change in Transport* (em inglês). [S.l.]:

Physica-Verlag. p. 305.

[5] https://pt.wikipedia.org/wiki/Árvore_de_decisão. Acessado em 05 de maio de 2022.

[6] *Tidy Modeling with* R MAX KUHN AND JULIA SILGE, *Version* 0.0.1.9010 (2022-04-29)

[7] Hosmer, David W., e Stanley Lemeschow. 2000. *Applied Logistic Regression. 2 ed. New York: Wiley*

[8] Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S. New York: Springer.*

[9] *Introduction to Machine Learning with R: Rigorous Mathematical Analysis*, O'Reilly Media. 2018.

[10]https://www.sanarmed.com/curva-roc-isso-realmente-faz-diferenca-colunistas/. Acessado em 25/05/2022.

[11]https://ajuda.datarisk.io/knowledge/o-que-%C3%A9-matriz-de-confus%C3%A3o/.Acessado em 25/05/2022.

[12]Hastie, T. J. and Pregibon, D. (1992) *Generalized linear models. Chapter 6 of Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.