XXIII SEMEAD
Seminários em Administração

novembro de 2020
ISSN 2177-3866

MODELOS DE MACHINE PARA TOMADA DE DECISÃO NO SISTEMA PÚBLICO DE SAÚDE BRASILEIRO

GUILHERME FERREIRA DA SILVA
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

DAIELLY MELINA NASSIF MANTOVANI
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

MODELOS DE MACHINE PARA TOMADA DE DECISÃO NO SISTEMA PÚBLICO DE SAÚDE BRASILEIRO

INTRODUÇÃO

O Sistema Único de Saúde do Brasil (SUS) enfrenta diariamente diversos empecilhos para um funcionamento pleno e fatores que podem elevar o nível de dificuldade na gestão de recursos e dificultar a maximização de vidas salvas pelo sistema. Para isso, análises e ferramentas precisam ser criadas no intuito de resolver ambos os problemas em diferentes aspectos.

A maior causa de morte de brasileiros anualmente, tendência que também acontece no restante do mundo, são doenças e males cardiovasculares, representando uma taxa de 80,02 a cada 100.000 habitantes em solo nacional, até o ano de 2019, antes da pandemia do Covid-19. Isso significa que todos os anos diversos brasileiros ocupam salas de pronto-atendimento, UTI e Emergência em função de problemas cardiovasculares, gerando altíssimos impactos financeiros na economia do país além de ter grande participação no impacto do orçamento do sistema de saúde.

Desta maneira, conseguir identificar possíveis pacientes desses malefícios antes de seu estado mais grave e realizar uma gestão de filas, identificando os casos mais graves e de maior complexidade, pode causar um grande impacto tanto na economia em médio e longo prazo, como salvar um maior número de vidas.

Diversos modelos de *Machine Learning* estão sendo aplicados em diversos setores, desde predições de risco de crédito para clientes de banco a até análises de como prever futuros focos de infecções de vírus em populações de risco. Assim, este tipo de ferramenta matemática possibilita atualmente tomadas de decisão muito mais embasadas e fornecendo capacidade de realizar ações mais assertivas e rápidas para empresas privadas e organizações públicas.

O objetivo do trabalho é desenvolver um modelo preditivo de associação que identifique pacientes que tem maior ou menor chance de ter algum malefício cardíaco, e identificar possíveis usos e benefícios da utilização de modelos preditivos na saúde pública brasileira.

COMPLEXIDADE PARA O SISTEMA DE SAÚDE NO MODELO POLÍTICO BRASILEIRO

O modelo político Brasileiro é o federalismo, que divide o Estado brasileiro em 3 âmbitos gerais: o Governo federal, que abrange todo o território da união, o estadual que abrange as unidades federativas, e o municipal. Dentro de suas esferas, segundo a constituição de 1988, cada uma delas tem independência administrativa e não estão ou devem estar ligadas hierarquicamente. Ou seja, o governo federal não pode exercer influência sobre as organizações públicas e políticas do governo estadual ou municipal, assim para todas as esferas (MS,2002).

Assim, segundo o Ministério da Saúde, em sua publicação sobre sua constituição e histórico, o modelo federalista pode exercer dificuldades para o planejamento e aplicação de modelos públicos de saúde, devido às peculiaridades da jurisdição, aplicação e modelo legislativo para cada município e estado brasileiro, assim como suas divergências nos quesitos de população geográfica, renda, e desigualdade (MS, 2002).

Antes da criação do SUS o ministério da saúde se concentrava em medidas de prevenção a doenças e promoção de políticas de saúde como campanhas de vacinação e controle de focos de doenças em locais específicos (endemias). Todas as medidas explicitadas tinham caráter universal, não tendo diferenciação por público, região ou caráter regional. Os focos em relação à assistência médica, havia poucos hospitais especializados como nas áreas de psiquiatria e pneumologia, e eram

concentradas em regiões específicas do território da união como o Norte e Nordeste, regiões com menor poder de compra e com IDH abaixo da média do país (MS, 2002).
A maior parte dos esforços na área de assistência médica era proporcionada pelo INAMPS (Instituto Nacional de Assistência Médica da Previdência Social), uma autarquia do Ministério da Previdência e Assistência Social. As ações do INAMPS se limitavam aos seus associados, ou seja, apenas trabalhadores que estivessem sob a legislação da CLT como trabalhadores formais. As demais camadas da população, portanto, não tinham direito a estes hospitais ou serviços públicos de maneira universal, e quando tinham acessos a estes meios eram provenientes de instituições de caráter filantrópico (MS, 2002).
No final dos anos 80 pode-se dizer que os requisitos de utilização de hospitais do INAMPS foram afrouxados, deixando com que cidadãos sem a Carteira de Segurado do instituto pudessem utilizar seus serviços, criava-se assim um sistema de atendimento mais próximo do universal, em um processo que ligava o sistema de saúde do INAMPS com as secretarias responsáveis nos estados brasileiros (MS, 2002).
Em 1990 foi outorgada a lei 8.080 da constituição federal que criava o Sistema único de Saúde (SUS) que tem comando único em cada esfera do governo e esta sob tutela do Ministério da Saúde transferindo o INAMPS do Ministério da Previdência Social. Entre os princípios do SUS destaca-se "universalidade do acesso aos serviços de saúde em todos os níveis de assistência", ou seja, todo o cidadão brasileiro, independente se empregado de maneira legal, classe social, localidade, idade ou gênero pode, por direito, ter acesso ao sistema público de saúde (MS, 2002).
O SUS, como maior órgão dentro do Ministério da saúde é um dos maiores prestadores de saúde pública do mundo, sendo o único que garante a assistência total gratuita para a totalidade da população, sendo declarado por alguns especialistas como o maior sistema de distribuição de renda do país (MS, 2018).
Segundo o Relatório de Gestão do Ministério da Saúde de 2018 tem-se:
- 42.826 Unidades Básicas de Saúde (UBS) em funcionamento no país, sendo 49,3% tendo prontuário eletrônico.
- 70% da população do país utilizando o serviço público de saúde, aproximadamente 147 milhões utilizando do sistema público de saúde.
- 82,7% da cobertura nacional pelo serviço emergencial de saúde (SAMU)
- 345 mil pessoas vivendo com HIV/AIDS com carga viral suprimida e totalmente paga pelo governo (95% de toda população infectada)
- 614 unidades de pronto atendimento (UPA) espalhadas pela união.

Ainda sob tutela do Ministério da saúde existe a farmácia popular, que além de fornecer gratuitamente remédios para doenças específicas para a população carente, como diabetes ou problemas respiratórios , ainda conta com o auxílio no envio de remédios de tratamentos caros para quem não tem acesso ou poder aquisitivo para um tratamento particular, como é o caso já citado de pacientes contagiados com o vírus HIV, tendo o tratamento inteiro coberto pelo sistema público (MS, 2018).
Com esses números o Ministério da Saúde conta com 62.000 servidores públicos e teve seu orçamento medido para o ano de 2018 em R$ 131 bilhões de reais, tendo gasto 118,3 bilhões do valor total em ações e serviços públicos de saúde. O gasto do governo federal com seu programa de saúde pública foi equivalente a 43% de todo o gasto com saúde pública no país, sendo os outros 57% divididos em autarquias municipais e estaduais (MS, 2018).

A principal complexidade do modelo e da gestão da política pública de saúde brasileira é a relação do sistema único e universal que interliga os 3 âmbitos da esfera pública, assim tendo sua ação variável ao seu meio; a grande população que é atendida, complexidade na distribuição de renda entre as camadas sociais, raciais e geográficas da sociedade, dificultando no processo de logística e criação de novas unidades de atendimento para as pessoas mais necessitadas do serviço, e a alocação de recursos para áreas de pouca população, como órgãos, equipamentos médicos, profissionais na área de saúde e remédios nas farmácias populares (MS, 2002).
Com isso, faz-se necessárias ferramentas que auxiliam na gestão e na criação de oportunidades de melhorias no processo do serviço de saúde. Sob esse preceito, em 2015, foi criado o SEI (Serviço Eletrônico de Informações) como um dos produtos do PEN (Processo Eletrônico Nacional) que teve o objetivo de diminuir drasticamente a utilização de papel dentro da área de gestão do SUS, assim como redução de perdas de documentos em pastas, ganhos com agilidade de processos via ERP e de escopo, automatizando e facilitando para os funcionários a análise financeira, de custos e folha de pessoal de maneira eletrônica (MS, 2018).
Ainda não foram realizadas pesquisas pelo Ministério da Saúde sobre o desempenho do serviço e próximos passos. O SEI ainda é classificado no site como projeto piloto e está em fase de implementação no SUS (MS,2018).

CENÁRIO DAS PRINCIPAIS CAUSAS DE MORTE NO PAÍS E FATORES DE RISCO

Segundo o departamento de informação e análise Epidemiológica, na secretária de vigilância em Saúde, no estudo de 2019, as principais causas de morte no país no ano de 2017 foram, em ordem decrescente: Doença Isquêmica do Coração, Doença cerebrovascular, infecção das vias aéreas, e Alzheimer (SVS,2019).
Pode-se verificar segundo o departamento, que a principal causa de morte foi a Doença Isquêmica do coração, a mesma persiste como a de maior mortalidade desde 1990 (SVS,2019) apesar de uma grande queda na sua taxa, indo de 169,1 por 100 mil habitantes em 1990 para 80,02 em 2017, uma redução de 53%, sendo a doença com maior taxa de redução no período. Tanto a primeira quanto a segunda maior causas de mortes no Brasil estão enquadradas no grupo de doenças cardiovasculares. Uma descrição breve desse grupo de doenças pode ser encontrada na página da Organização Pan-Americana de Saúde (OPAS) em português em uma publicação de 2017 (OPAS,2017):

> "As doenças cardiovasculares são um grupo de doenças do coração e dos vasos sanguíneos e incluem: Doença coronariana – doença dos vasos sanguíneos que irrigam o músculo cardíaco; Doença cerebrovascular – doença dos vasos sanguíneos que irrigam o cérebro; Doença arterial periférica – doença dos vasos sanguíneos que irrigam os membros superiores e inferiores; Doença cardíaca reumática – danos no músculo do coração e válvulas cardíacas devido à febre reumática, causada por bactérias estreptocócicas; Cardiopatia congênita – malformações na estrutura do coração existentes desde o momento do nascimento; Trombose venosa profunda e embolia pulmonar – coágulos sanguíneos nas veias das pernas, que podem se desalojar e se mover para o coração e pulmões. Ataques cardíacos e acidentes vasculares cerebrais geralmente são eventos agudos causados principalmente por um bloqueio que impede que o sangue flua para o coração ou para o cérebro. A razão mais comum para isso é o acúmulo de depósitos de gordura nas paredes internas dos vasos sanguíneos que irrigam o coração ou o cérebro. Os acidentes vasculares cerebrais também podem ser causados por uma hemorragia em vasos sanguíneos do cérebro ou a partir de coágulos de sangue. A causa de ataques cardíacos e AVCs geralmente são uma

> combinação de fatores de risco, como o uso de tabaco, dietas inadequadas e obesidade, sedentarismo e o uso nocivo do álcool, hipertensão, diabetes e hiperlipidemia (OPAS, 2017)"

Isso demonstra uma forte evolução nas políticas públicas de combate e prevenção a estas doenças. As pessoas com maior disposição a sofrer com essas doenças são pacientes com hipertensão, diabetes e casos anteriores na família (OPAS, 2017). Os principais fatores de risco para pessoas com pré-disposição a esse tipo de mal são tabagismo, dietas não-balanceadas, obesidade, sedentarismo e consumo excessivo de álcool (OPAS,2017).

Outros fatores explorados pela tese de mestrado da Doutora e Professora, da Universidade de São Paulo, Kelli Oliveira (2004) são divididos em dois grupos: os fatores individuais e os fatores comunitários. Os Fatores de Risco Individuais para as doenças cardiovasculares são: Idade, Sexo, Nível de Instrução, composição genética, Tabagismo, Hábitos alimentares, sedentarismo, nível de colesterol no sangue, e obesidade (aqui sendo considerado o nível mais elevado de IMC do indivíduo). Já os fatores de risco Comunitários são a situação econômica, desemprego, composição familiar, clima, poluição do ar, práticas, normas e valores e nível de infraestrutura do ambiente (no original, urbanização).

Assim, é possível verificar que o grau de escolarização, acesso à informação e o atendimento médico inadequado estão muito mais relacionados à prevenção dos outros fatores de risco, assim, diminuindo as chances de mortalidade e ocorrência da doença. Assim faz-se de extrema importância identificar quais os fatores de risco mais impactantes na mortalidade e acentuação das doenças cardiovasculares, identificando os pesos das variáveis, ajudando a entender as maneiras mais expressivas tanto para prevenção quanto para o tratamento deste malefício.

Diversos estudos correlacionando modelos estatísticos preditivos, modelos de *machine learning* e demais ferramentas estão sendo realizados para entender como podem auxiliar a saúde, não apenas no âmbito público, por todo mundo.

Em 2019, por exemplo, a professora Helen Geremias dos Santos, professora na Faculdade de Saúde Pública da Universidade de São Paulo, fez a comparação de 5 modelos preditivos e analisou sua eficácia com base na amostra obtida do estudo "Saúde Bem-estar e Envelhecimento", os resultados foram satisfatórios, sendo os modelos com melhor desempenho os de redes neurais, seguidos por dois modelos de Regressão Logística (Santos, 2019).

No estudo de Santos (2019), tentou-se predizer a variável depende "óbitos em até 5 anos" de idosos, no município de São Paulo. A amostra contou 2.808 casos e a base foi dividida na proporção de 70% da amostra em treino e 30% da amostra em validação. O estudo ainda realizou a observação de que os modelos preditivos podem auxiliar no sistema de saúde, até mesmo em um processo de triagem, mas devem-se usar os resultados com cautela e realizar calibragem com o decorrer do tempo. Segundo a autora,

> "Em relação à qualidade do risco predito, mesmo que um modelo não apresente boa calibração, ainda pode ser útil em aplicações específicas, como para determinado ponto de corte da probabilidade predita ou para cenários em que a sensibilidade ou a especificidade seja mais relevante. Por exemplo, em casos de utilização de determinado modelo para uma atividade de triagem, em que o objetivo é discriminar indivíduos com risco muito baixo para determinado desfecho. Embora esse modelo possa não apresentar calibração perfeita, ou seja, possa resultar em estimativas incorretas do risco para indivíduos (...)" (Santos, 2019).

Em âmbito internacional, Jenna Wiens e Erica S. Shenoy discutem outra área de aplicação de *machine learning* na área da saúde. Em agosto de 2017, ambas realizaram uma análise para entender como estas técnicas poderiam auxiliar no ramo da Epidemiologia. Em seu artigo, elas debatem que modelos estatísticos de predição corroboram com uma melhoria, que antes da criação e de sua utilização não seriam possíveis, a análise de diversas variáveis independentes em bases com muitos dados ao mesmo tempo, podendo assim identificar possíveis influenciadores na variável dependente que antes, pela falta de poder de processamento e na dificuldade de se obter os dados, não eram possíveis de serem analisadas.

Apesar de verificarem e confirmar o êxito destas técnicas, ambas ressaltam que os modelos que se dão melhor com grandes quantidades de dados, como redes neurais, tendem a ter melhores resultados, e que devem ser a primeira escolha, caso fontes de dados como essas, estejam disponíveis. Ressaltam, porém, a necessidade da compreensão que as variáveis dependentes e independentes, mesmo em modelos mais complexos e auditáveis, não podem ser relacionados com relações de causa e efeito, e sim, ger a oportunidade para possíveis ações de prevenção de Epidemiologistas com base nas mesmas e não tentando impedir a ocorrência das variáveis em questão (Wiens & Shenoy, 2017).

PROCEDIMENTOS METODOLÓGICOS

Este estudo pode ser considerado exploratório-descritivo. Segundo Sekaran e Bougie (2009), existem quatro principais motivos para a utilização do método exploratório, ou o problema é altamente complexo, ou não há muita pesquisa sobre o tópico explorado, pode-se também não ter uma pesquisa realmente fundamentada com conclusões claras sobre o modelo explorado ou, por último, não há teoria base suficiente para confirmar pesquisas prévias segundo o método científico (Sekaran & Bougie, 2009).

Já a pesquisa descritiva tem como principal função descrever um fato ou problema, um exemplo seria determinar qual a média de aprovação de um político em uma certa região do país ou determinar a idade média de pessoas com problemas cardíacos na cidade de São Paulo. Outra forte expressão e importância deste tipo de pesquisa é tentar correlacionar duas variáveis que podem apresentar uma ocorrência de causa-efeito (Sekaran & Bougie, 2009).

Vale ressaltar que este tipo de pesquisa pode verificar a correlação das variáveis entre si mas não demonstrar relação de causa e efeito, isso ocorre pois a mesma se limita a utilizar e coletar dados para a amostragem de maneira matemática e para respostas diretas e simples. O estudo seguiu duas etapas:

- Pesquisa Exploratória – Levantamento bibliográfico para obtenção de dados secundários com a finalidade de entender o problema: O modelo de saúde pública do Brasil é complexo, assim, ferramentas de *Machine Learning* ou outros modelos estatísticos, principalmente os modelos preditivos, podem auxiliar em sua gestão? Quais as variáveis adequadas para criar um modelo preditivo?
- Pesquisa Descritiva - Verificar as variáveis explicitadas na pesquisa exploratória, checar suas correlações com o nível de urgência do caso do paciente e entender quais as mais importantes em casos de urgência assim possibilitando a criação de um modelo que maximize a "produtividade" dos recursos públicos de saúde para o atendimento da população.

A maior causa de mortes no país, mas também no mundo, são problemas cardiovasculares, que apesar de uma grande diminuição nas últimas décadas, contínua sendo a maior causa de mortes desde a década 1990, quando o estudo começou. Para a elaboração do algoritmo foram usadas

bases cedidas pelo IBGE referente à Pesquisa Nacional de Saúde de 2103, cujos dados foram coletados entre 2012 e 2013, sendo esses os dados mais atuais disponíveis.
A linguagem R foi escolhida como opção de programação. Diante do objetivo do estudo e natureza das variáveis, considerou-se adequado aplicar algoritmos de classificação (classificar os pacientes por urgência). Nesse sentido foram aplicados dois algoritmos, árvores de decisão (*decision trees*) e regressão logística. As *decison trees* criam modelos de classificação com bases nas variáveis independentes utilizando fatores lógicos de segregação, utiliza modelos semelhantes a "if X do Y" para decidir qual grupo cada indivíduo pertence (Lantz, 2013). A Árvore de Decisão, retorna o processo de decisão que o algoritmo tomou, assim, o usuário consegue verificar se o processo faz sentido e se deve ser modificado ou aprimorado para uma melhor utilização.
Para a obtenção de uma base de dados de confiança e com boa assertividade foram conferidas e contactadas três principais possibilidades de órgãos públicos, e/ou, filantrópicos da área de pesquisa ou saúde. Foram eles: DATASUS, IBGE, e a OPAS, obtendo-se retorno apenas do IBGE.
Após a obtenção da base não consolidada foi necessário mapear quais variáveis puderam e quais não puderam ser compostas para a criação do algoritmo. Após a varredura no dicionário, também enviado pelo IBGE, foram selecionadas as variáveis de entrada do modelo.

Variável	Mensuração
Diagnóstico cardíaco (variável dependente)	0= não, 1=sim
Número de moradores no domicílio	Número de moradores no domicílio entrevistado
Sexo	1= Masculino, 2 = Feminino
Idade	Idade do cidadão entrevistado no momento da Pesquisa
Escolaridade	Qual o curso mais elevado frequentado pelo entrevistado variando entre 1= "Alfabetização" e 12= "Doutorado"
Número de empregos	Número de trabalhos simultâneos o entrevistado durante a entrevista
Renda do paciente	Somatória dos salários dos trabalhos que o entrevistado tinha durante a entrevista
Regularidade consumo alcóolico	Número de dias por semana que o entrevistado consumia bebidas alcoólicas normalmente
Regularidade atividade física	Número de dias por semana que o entrevistado pratica atividade física
Fumante	Se o entrevistado fuma algum produto derivado de tabaco: 1= Sim, Diariamente; 2= Sim, menos que diariamente; 3 = Não
Hipertensão	Já foi diagnostica com pressão alta? 1="Não"; 2="Apenas durante a gravidez"; 3="Sim"
Diabetes	Já foi diagnostica com Diabetes? 1="Não"; 2="Apenas durante a gravidez"; 3="Sim"
Colesterol alto	Já foi diagnostica com Diabetes? 1="Não"; 2="Apenas durante a gravidez"; 3="Sim"
IMC	Peso em kg medido pela última vez do entrevistado/ (altura medida pela última vez do entrevistado) ^2

Quadro 1. Variáveis de entrada no modelo

Inicialmente, planejou-se utilizar como variável dependente o "óbito do paciente", contudo, em função da indisponibilidade dessa informação na base de dados utilizada, foi necessário redefinir a variável resposta. Como previamente descrito, as duas maiores causas de morte no Brasil se deram

por problemas relacionados diretamente ou indiretamente a doenças cardiovasculares, assim sendo, fez-se a opção para a continuidade do modelo de se usar a variável “Algum médico já lhe deu o diagnóstico de doença do coração tal como infarto, angina, insuficiência cardíaca ou outra?”, como dependente. As respostas disponíveis para esta pergunta, na PNS 2013 foram “Sim”, “Não” e “Não Aplicável”. Desta maneira, as respostas passam a ser mais objetivas e independem de quando foram diagnosticadas.

Essa alteração foi possível pois, como a intenção original era analisar casos de falecimento por estes tipos de males, as variáveis independentes puderam ser utilizadas para auxiliar na previsão do diagnóstico e auxílio no tratamento e prevenção destas doenças.

O início da análise em R dos dados se deu em 3 etapas: A extração/obtenção de dados, a transformação/tratamentos destes dados, e a execução do modelo de *Machine Learning*, incialmente escolhido pelo algoritmo de árvore de decisão. Este processo se assemelha ao modelo de estruturação de dados conhecido por ETL (*Extract, Transform, and Load*) normalmente utilizado nas ferramentas de *Business Intelligence*. A utilização desta ordem nos processos tende a diminuir o excesso de repetições de etapas previamente realizadas de maneira desnecessária.

Inicialmente fio realizado o tratamento dos dados. Retiraram-se os erros, dados faltantes e valores que deveriam ser desconsiderados, como, por exemplo, os valores “NA” na coluna de “Número de empregos” e os valores “.” na coluna “Altura (em cm)”.

Após a retirada destes valores houve uma perda significativa no número de linhas em nosso *dataframe*, indo de 205.546 linhas para 41.578 linhas. Uma redução de aproximadamente 79,8% de toda a base de dados registrada pela PNS 2013. A variável com maior responsabilidade de perda de dados foi nossa variável dependente, por apresentar muitos dados faltantes. Além disso, transformaram-se valores que apareciam como *strings* em valores numéricos, converteram-se “.” em zeros e demais processos necessários de tratamento de base.

Após o tratamento foram verificados os tipos e classes das variáveis com a função *str* padrão do R. Assim obtivemos os valores na figura 1:

```
'data.frame':   41578 obs. of  14 variables:
 $ Número de moradores no domicilio               : int  4 3 3 3 3 1 1 2 3 1 ...
 $ Sexo                                           : int  1 2 1 2 1 1 1 1 1 2 ...
 $ Idade                                          : int  35 32 53 36 44 27 29 37 33 55 ...
 $ Nível de escolaridade                          : int  5 8 5 9 8 5 8 5 8 8 ...
 $ Número de empregos                             : num  1 0 1 1 1 1 1 1 1 1 ...
 $ Renda do paciente                              : num  700 0 2000 1200 2000 900 1200 741 1100 2000 ...
 $ Regularidade de consumo alcóolico              : int  1 1 1 1 1 1 1 1 1 3 ...
 $ Regularidade de prática de exercícios por semana: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Fumante                                        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Já foi diagnosticado com hipertensão           : int  3 3 3 3 3 3 3 3 3 3 ...
 $ É diagnosticado com Diabetes?                  : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Foi diagnosticado com colesterol alto          : int  2 2 2 2 2 2 2 2 2 2 ...
 $ Cardiaco                                       : Factor w/ 2 levels "Não","Sim": 1 1 1 1 1 1 1 1 1 1 ...
 $ IMC                                            : num  59.5 55 78 60.6 70.4 66 80 77.3 89 66 ...
```

Figura 1. Variáveis independentes e suas classes.

Também foram transformadas as variáveis em fatores, que facilitam a maneira como o código de árvore decisão funciona, os transformamos utilizando a função *as.factor()*.

TESTE DO ALGORITMO DE ÁRVORE DE DECISÃO E CONCLUSÕES INICIAIS

A divisão da base em duas, assim como deixar a base de maneira aleatória é essencial para a utilização do algoritmo de árvore de decisão. A principal utilização do modelo árvore de decisão, assim como boa parte de outros algoritmos e modelos de *Machine Learning* é a sua capacidade de previsão e agrupamentos para novos casos do material estudado. Como o exemplo deste estudo, o

algoritmo tem como função agrupar os pacientes em prováveis casos de risco de doenças cardíacas ou em pacientes com menor risco de terem doenças cardíacas.
Desta maneira se faz se extrema importância dividir a base em treino, onde o algoritmo verificará os casos, e moldará seus valores para classificar as pessoas nestes dois grupos, e a base de teste, onde serão comparadas as classificações feitas pelo algoritmo com a realidade para entender se o modelo pode trazer um benefício de ganhos de resultado ou predição correta.
Utilizar uma mesma base para treino e predição pode gerar o problema sobre ajuste de bases além de dificultar a percepção de casos não contemplados na amostra (Lantz, 2013). Isso significa que se pode criar um modelo praticamente perfeito para a sua amostra, porém caso ela esteja com algum viés ou erro, esse erro será totalmente transferido ao algoritmo. A divisão da tabela de dados tem como objetivo reduzir esse ajuste específico da tabela e verificar como algoritmo se sai com casos novos. Assim, dividimos nosso treino e performance em uma base de treino com 95% das linhas e uma de teste com 5% das linhas.
Deixar a base com uma ordem aleatória é essencial para evitar que a ordem dos dados se torne um viés que prejudique de forma agressiva o modelo criado. Imaginemos que a PNS 2013 tenha sido realizada por unidade federativa, ao dividir o *dataframe* poderíamos estar excluindo algum estado do modelo, e caso algum fator daquele estado seja necessário para uma melhor explicação de resultados, não teríamos englobado no cálculo classificatório (Lantz, 2013).
Para utilizar essas duas etapas utilizamos as funções *set.seed* e *order.* Após essa etapa verificamos as proporções da nossa variável dependente para entender se sua proporção tanto na base consolidada quanto nas bases de treino ou de teste se mantém próximas, evitando vieses e permitindo a construção do modelo. Observou-se nas bases consolida, de teste e treino proporções semelhantes de casos nas categorias da variável dependente (~95% de casos sem diagnóstico de problemas cardíacos).
Utilizando o algoritmo C5.0 e rodando a função, a resposta resultante do algoritmo da árvore de decisão foi apenas uma *"root"*, isso significa que, para a análise da base em específico, o modelo de árvore de decisão (na função c5.0), o erro das classificações não "compensa" a determinação de um caso como "**Sim**". Assim o algoritmo determina que é mais simples e diminui o erro ao classificar todos os casos como "Não", deixando o erro do modelo com mesmo o percentual de amostras com algum malefício cardíaco já diagnosticado.
Isso significa que, para esses parâmetros, o modelo de árvore de decisão não nos auxilia na classificação, e assim, não seria útil em sua aplicação no SUS.
Para que o caso fosse melhor explicitado foi realizado um segundo teste, agora com o algoritmo de árvore de decisão da biblioteca *rpart*, que tem uma customização de seus parâmetros mais simples, além de ser mais popular devido a outras funções além de sua *Decision Tree*. Com esse novo algoritmo obtivemos resultados diferentes, como ilustra a figura 2.

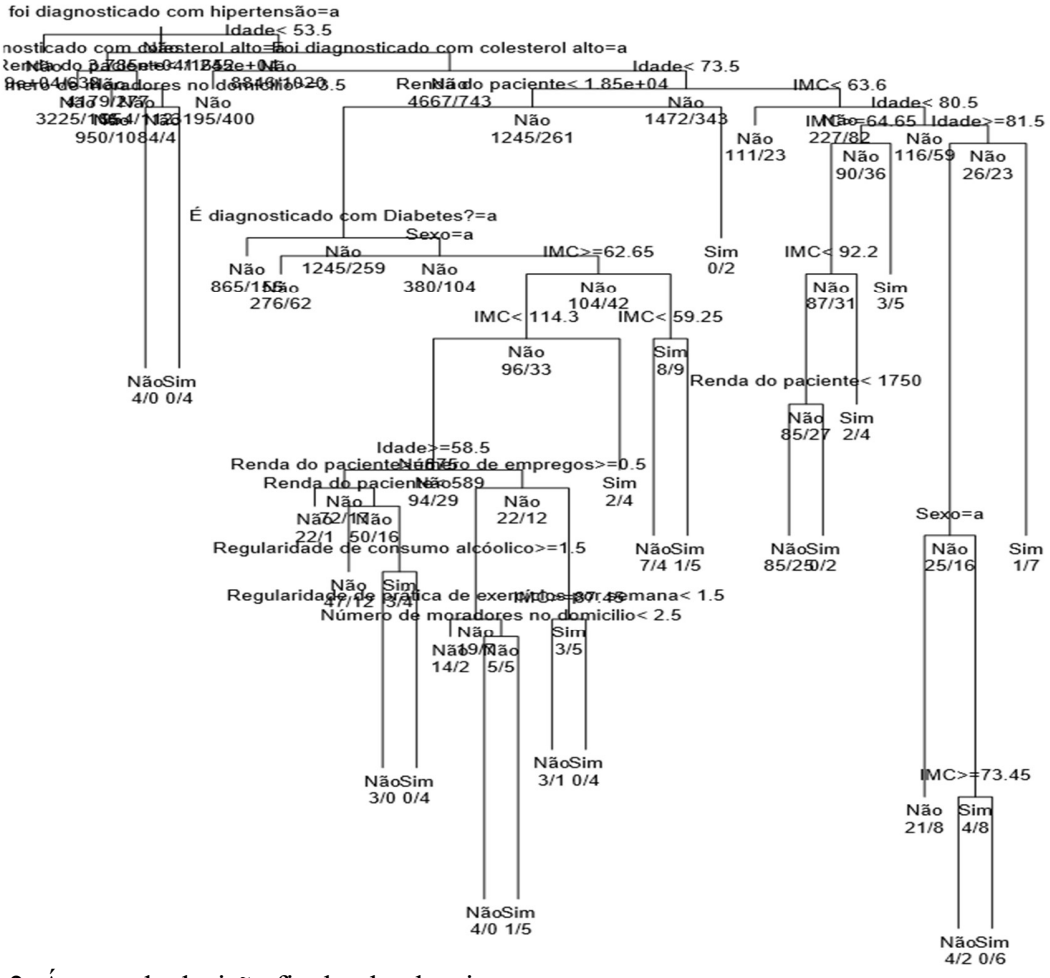


Figura2. Árvore de decisão final pelo algoritmo *rpart*

Ao verificar a consistência dos galhos obtidos pela árvore, foram encontrados alguns problemas. Além do número grande de galhos e suas complexidades, pode-se observar algumas “falhas” no modelo. Ao verificar as respostas do galho cuja pergunta é: A idade é menor que 73.5 anos? Caso a resposta seja não (a pessoa é mais velha) e o paciente não for diagnosticado com diabetes e a renda dele for maior que o valor de 8500 reais ele será classificado como fora do risco, ou seja, com menor probabilidade de ter ou sofrer de um malefício cardíaco.

Ao verificarmos a literatura (Oliveira Kelli, 2004) vemos que idade é um fator que aumenta o risco de se ter um ataque cardíaco ou outra doença cardiovascular. Utilizá-lo como galho de exclusão de risco poderia fazer com que um paciente mais velho, apenas por não ter sido diagnosticado com diabetes e ter um poder aquisitivo maior, fosse excluído do grupo de risco, o que seria uma falha.

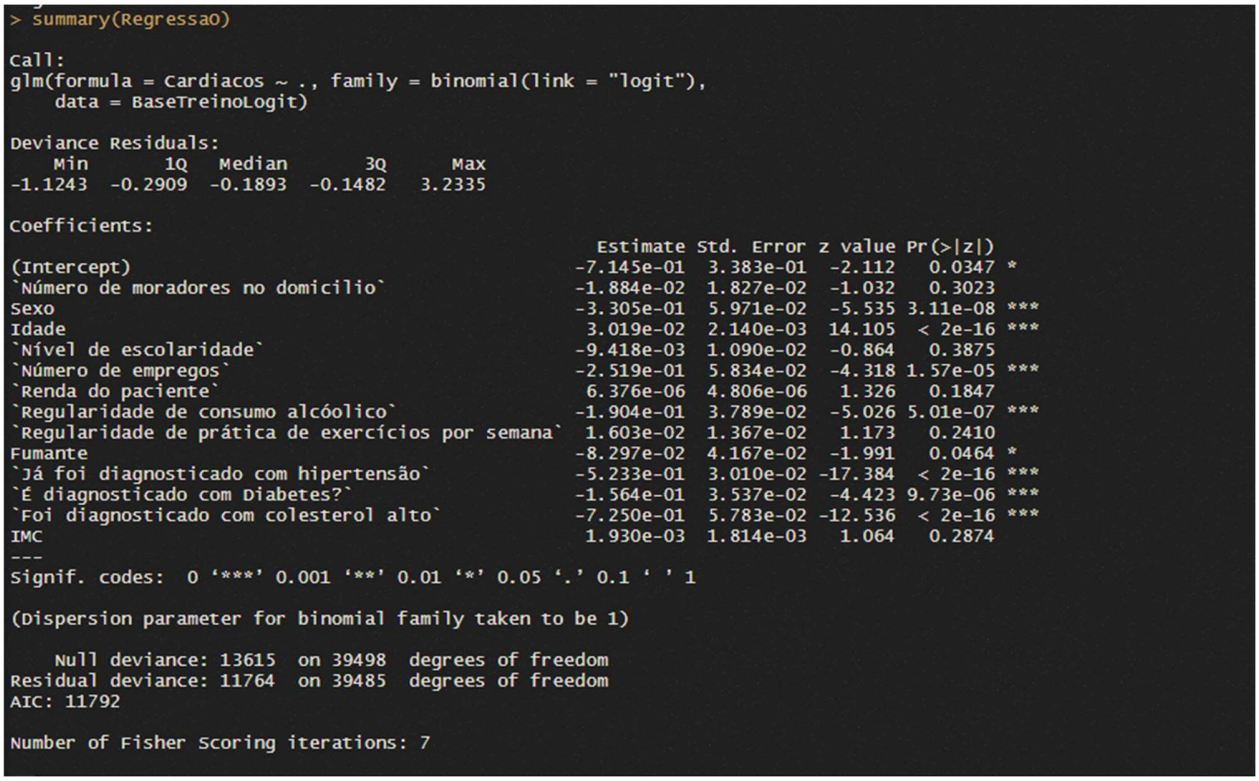
Esse exemplo não demonstra um erro da árvore de decisão em si, apenas ressalta suas limitações do algoritmo. Quando o modelo foi selecionado para a tomada de decisão, a pergunta que exigia uma resposta era se o paciente teria maiores chances de morrer ou não, porém, com a mudança da variável dependente para se o paciente tem maior ou menor chance de ter um mal do coração, a resposta determinante resultado de uma árvore de decisão deixa de ter seu efeito proposto.

Outro fator limitante do uso da árvore de decisão é a base em que ela é construída. Apesar de ser uma base consistente e grande, a PNS 2013, pode ter falta de casos específicos. Não é verdade que poucos pacientes acima de seus 74 anos com mais de 8500 reais de renda e sem diabetes tem algum problema cardiovascular, porém, a realidade é que nenhum destes casos está na base de dados, enviesando a árvore de decisão.
Assim, como a pergunta e a variável foram alteradas com a entrega da base, também devemos mudar o algoritmo de resposta, um algoritmo que sofra menos com os vieses das bases e que não seja determinante, mas sim, classificatório, desta maneira, podemos analisar os resultados não em grupos de cardíacos ou não, mas analisar dentre os pacientes quem tem mais chances de ter um mal desses e quem tem menos probabilidade. Com este cenário a proposta de mudança foi para o algoritmo de Regressão Logística.

REGRESSÃO LOGÍSTICA E SEUS RESULTADOS NO DATAFRAME

O modelo de regressão logística é uma regressão múltipla que não necessita que suas variáveis independentes assumam uma característica de distribuição normal, além disso, o resultado de sua predição é a probabilidade de um evento ocorrer, sendo um valor entre 0 e 1 (Hair Jr, 2009). Assim, podemos utilizar o modelo de regressão logística como predição tomando os seguintes critérios:
- A variável dependente é se o paciente foi diagnosticado com alguma doença cardíaca (1) ou se não foi (0).
- A resposta para cada cidadão pesquisado é a probabilidade de o mesmo ser classificado como provável paciente cardíaco, podendo assim classificar de maior para menor os casos.

Para que a regressão pudesse ser feita, o primeiro passo foi alterar as variáveis dependentes, de Sim e Não, para 1 e 0 respectivamente. Além disso foi realizada uma cópia da base de dados para que um modelo não interferisse no outro. Então dividimos novamente a base em duas, além de deixarmos sua ordem aleatória e verificarmos se a proporção de valores da variável dependente permanece similar em ambas as divisões (treino e teste), o que foi respeitado. A figura 3 apresenta os resultados do modelo.



```
> summary(RegressaO)

Call:
glm(formula = Cardiacos ~ ., family = binomial(link = "logit"),
    data = BaseTreinoLogit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1243  -0.2909  -0.1893  -0.1482   3.2335

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -7.145e-01  3.383e-01  -2.112   0.0347 *
`Número de moradores no domicilio`          -1.884e-02  1.827e-02  -1.032   0.3023
Sexo                                        -3.305e-01  5.971e-02  -5.535 3.11e-08 ***
Idade                                        3.019e-02  2.140e-03  14.105  < 2e-16 ***
`Nível de escolaridade`                     -9.418e-03  1.090e-02  -0.864   0.3875
`Número de empregos`                        -2.519e-01  5.834e-02  -4.318 1.57e-05 ***
`Renda do paciente`                          6.376e-06  4.806e-06   1.326   0.1847
`Regularidade de consumo alcóolico`         -1.904e-01  3.789e-02  -5.026 5.01e-07 ***
`Regularidade de prática de exercícios por semana`  1.603e-02  1.367e-02   1.173   0.2410
Fumante                                     -8.297e-02  4.167e-02  -1.991   0.0464 *
`Já foi diagnosticado com hipertensão`      -5.233e-01  3.010e-02 -17.384  < 2e-16 ***
`É diagnosticado com Diabetes?`             -1.564e-01  3.537e-02  -4.423 9.73e-06 ***
`Foi diagnosticado com colesterol alto`     -7.250e-01  5.783e-02 -12.536  < 2e-16 ***
IMC                                          1.930e-03  1.814e-03   1.064   0.2874
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13615  on 39498  degrees of freedom
Residual deviance: 11764  on 39485  degrees of freedom
AIC: 11792

Number of Fisher Scoring iterations: 7
```

Figura 3. Coeficientes de regressão do modelo inicial.

Verifica-se, portanto, que número de moradores no domicílio, Nível de escolaridade, Renda do paciente, Regularidade da prática de exercícios por semana, e IMC são variáveis que, na base do IBGE, não são significantes, e podem ser retirados do modelo. A variável “Fumante” é significante ao nível de 5%, apresentando valor p próximo de alfa, desta maneira foi retirada do modelo, assim como as variáveis não significantes. Destaca-se que na base de dados original o código 1 foi atribuído à resposta não, por isso hipertensão, diabetes e colesterol aparecem com coeficientes negativos. A figura 4 apresenta o modelo ajustado e com a codificação reajustada para que as respostas sim tivessem código 1. A regularidade de consumo alcoólico, que se refere ao número de dias na semana que o entrevistado consome este tipo de bebida, por sua vez, teve que ser retirada do modelo, pois apresentava inconsistências com a referência bibliográfica devido aos problemas da amostra em questão, que variavam apenas entre os consumos de 1 a 3 dias por semana, não contendo casos de pacientes que consomem, por exemplo, todos os dias ou casos em que não bebem em nenhum dia da semana, o que enviesa toda a construção do modelo.

```
Call:
glm(formula = Cardiacos ~ ., family = binomial(link = "logit"),
    data = BaseTreinoLogit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1065  -0.2917  -0.1889  -0.1497   3.1877

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -6.251258   0.167292 -37.367  < 2e-16 ***
Sexo                               -0.278302   0.055234  -5.039 4.69e-07 ***
Idade                               0.031707   0.001945  16.301  < 2e-16 ***
`Número de empregos`               -0.271495   0.056979  -4.765 1.89e-06 ***
`Já foi diagnosticado com hipertensão`  0.530805   0.029493  17.998  < 2e-16 ***
`É diagnosticado com Diabetes?`     0.166794   0.035122   4.749 2.04e-06 ***
`Foi diagnosticado com colesterol alto`  0.722834   0.057554  12.559  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13615  on 39498  degrees of freedom
Residual deviance: 11797  on 39492  degrees of freedom
AIC: 11811

Number of Fisher Scoring iterations: 7
```

Figura4. Modelo de regressão logístico ajustado

As variáveis retidas no modelo foram o Sexo, que como visto anteriormente, ocorre em maior frequência em homens do que mulheres; Número de empregos que o entrevistado tinha no momento da entrevista; Já foi diagnosticado com Hipertensão, que se refere se o paciente já foi diagnosticado com pressão alta; Já foi diagnosticado com Diabetes, para verificar se o paciente já foi diagnosticado anteriormente com diabetes; e Foi diagnosticado com colesterol alto, que verifica se o entrevistado em algum exame já teve colesterol alto apontado. As variáveis com maior influência sobre a predição são hipertensão, idade e colesterol alto.
O R-quadrado= 86,6% indica bom ajuste do modelo. As estatísticas *odds ratio* indicam como cada perfil de paciente afeta a probabilidade de se ter um problema cardíaco. O Quadro 1 apresenta os valores de *odds ratio* e sua interpretação.

Variável	*Odds Ratio*	Interpretação
Sexo	0, 757068	Caso o paciente seja do sexo Masculino há mais chances de ser diagnosticado com malefício Cardíaco.
Idade	1,032214	Quanto maior a idade, maior a chance de ser diagnosticado com malefício cardíaco.
Número de empregos	0,762239	Quanto maior o número de empregos simultâneos, menor a chance de ser diagnosticado com malefício cardíaco.
Hipertensão	1,700300	Caso seja diagnosticado com Hipertensão previamente, maior as chances de também ser diagnosticado com algum malefício do coração.
Diabetes	1,181510	Caso seja diagnosticado com Diabetes previamente, maior as chances de também ser diagnosticado com algum malefício do coração.

Colesterol alto	2.0794263	Caso seja diagnosticado com Colesterol Alto previamente, maior as chances de também ser diagnosticado com algum malefício do coração.

Quadro 1.*Odds Ratio* do modelo logístico

No exemplo que estamos lidando, no caso de saúde, caso classifiquemos uma pessoa cardíaca como não cardíaca, ela pode demorar mais na fila do atendimento do SUS, e podendo falecer ou piorar seu estado de saúde nestes momentos de espera. Enquanto isso classificar uma pessoa sem malefício cardíaco como grupo de risco pode fazer com que cause maior estresse no paciente ou até mesmo sobrecarregar a fila de espera para este tipo de atendimento, porém não afeta diretamente a saúde física do paciente. Com estes parâmetros pode-se afirmar que classificar uma pessoa cardíaca como não cardíaca (falso negativo) é pior para a saúde pública, pois pode levar o paciente a óbito. O modelo classificou erroneamente 11,88% dos casos com algum problema cardíaco e 88,12% foram classificados corretamente. Com isso podemos fazer algumas considerações.

A base enviada pelo IBGE, apesar de ter um número satisfatório de casos, apresenta algumas limitações, ao excluir dados faltantes e casos sem informação, perderam-se 80% dos casos da amostra. Além disso, diferente da primeira proposta de variável dependente, quando tratava-se de mortes, o diagnóstico pode se passar como algo subjetivo, pois depende da última vez que o paciente foi a uma consulta médica e se alguma vez suspeitou sofrer de algum malefício do coração para que fosse investigado pelo médico.

Na base da PNS 2013 havia a pergunta se o entrevistado foi ao médico nos últimos 12 meses, porém menos de 35% de toda a base havia respondido essa pergunta, e, destas, menos de 20% haviam respondido que haviam visitado um médico no período. Isso demonstra que não necessariamente uma pessoa na base de dados, que está classificada como "não diagnosticada", realmente não sofre de algum malefício cardiovascular, apenas indica que, até a última vez que ele visitou ao médico, não havia sido diagnosticado. Além disso, existe a possibilidade do viés do erro no diagnóstico por parte do médico, mesmo fato que se torna muito mais raro quando a variável é morte.

Isso pode indicar que existe a possibilidade de a regressão logística funcionar satisfatoriamente para a previsão de risco em saúde pública, porém, no presente estudo tem-se como possíveis limitadores de performance a data da pesquisa (2013), o viés da variável dependente (o paciente pode não ter diagnóstico por falta de acompanhamento médico do paciente ou pelo diagnóstico errado) e o viés da amostra, que apesar de grande pode demonstrar algum viés da forma de pesquisa, e mesmo que aconteça em menor grau que em árvores de decisão, também afetam o algoritmo de Regressão Logística.

CONCLUSÕES

Algoritmos de *Machine Learning* são muito suscetíveis à vieses de bases de dados. Em um país como o Brasil, onde estão apenas iniciando o processo de unificação de bases de saúde, com tanta diversidade populacional, assim como sua densidade, a falta de consistência de dados torna muito difícil recomendar árvores de decisão para esse tipo de aplicação.

A Regressão Logística, por outro lado, apesar de errar um percentual considerável de casos, quando consideradas as classificações de pacientes com algum problema cardíaco, o modelo conseguiu classificar corretamente 88% dos casos, mesmo com um número de dados que poderia ser maior,

e que podem estar atualmente desatualizados. Assim, é possível checar grande potencial para este tipo de algoritmo em tomadas de decisões na área pública da Saúde.
Os principais meios onde pode-se observar o uso deste tipo de ferramenta são: Um processo de triagem que considera mais fatores que apenas os biológicos para a gestão de filas, assim, o grupo classificado como ameno, não deixaria de ser atendido em hospitais especializados em doenças cardíacas, mas seriam atendidos após o grupo classificado como de maior risco pelo algoritmo. Desta maneira os 88% dos pacientes com algum malefício cardíaco (taxa que foi consistente de acerto na classificação na base de teste) estariam nos 50% mais rápidos a serem atendidos, gerando maior chances de sobrevivência nos médio e longo prazos, em caso de não urgência.
O outro uso deste algoritmo poderia ser em um processo de encerramento de uma ida a uma UPA ou hospital público, desta maneira, quando o paciente saísse de sua consulta, com poucas perguntas, poderia ser classificado como grupo de risco e incentivado pelo médico, ou enfermeiro, que o atendeu a marcar uma consulta pela rede pública em um cardiologista.
Com estes dois usos pode-se auxiliar em duas limitações do sistema público de saúde, a mudança de paradigma de um sistema reativo aos malefícios do coração para um sistema com mais chances de ser preventivo, principalmente nas classes com menor conhecimento e acesso à informação; e na gestão de filas, que não reduzem de tamanho, mas que, em hospitais especializados e em consultas em cardiologistas podem aumentar a chance de sobrevivência, escolhendo com maiores chances atender primeiro quem necessita mais do serviço antes de quem pode esperar um pouco mais por atendimento, considerando também mais variáveis que o processo atual de agendamento.
Assim pode-se afirmar que há grandes expectativas e possibilidades para a utilização de *Machine Learning* na saúde pública, mas para que o processo de utilização possa ser implementado corretamente, se faz totalmente necessário a criação de uma gestão de dados consistente, pública, e transparente com o cidadão, tornando mais fácil e possibilitando uma maior confiança e credibilidade nos algoritmos criados.

REFERÊNCIAS

DEPARTAMENTO DE ANÁLISE DE SAÚDE E VIGILÂNCIA DE DOENÇAS NÃO TRANSMISSÍVEIS, 2019. Disponível em:< http://svs.aids.gov.br/dantps/centrais-de-conteudos/paineis-de-monitoramento/mortalidade/gbd-brasil/principais-causas/>. Acesso em 07 de jun. de 2020.

GEREMIAS DOS SANTOS, HELLEN. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. 2019. Disponível em:< http://cadernos.ensp.fiocruz.br/csp/artigo/792/machine-learning-para-anlises-preditivas-em-sade-exemplo-de-aplicao-para-predizer-bito-em-idosos-de-so-paulo-brasilAcesso em 07 de jun. de 2020.

HAIR JR., J.F.; WILLIAM, B.; BABIN, B.; ANDERSON, R.E. Análise multivariada de dados. 6.ed. Porto Alegre: Bookman, 2009.

LANTZ, BRETT. Machine Learning with R, 2013.

MINISTÉRIO DA SAÚDE. Relatório de Gestão 2018., 2018. Disponível em:< https://www.saude.gov.br/relatorio-de-gestao>. Acesso em 07 de jun. de 2020.

ORGANIZAÇÃO PAN – AMERICANA DA SAÚDE, 2017. Doenças cardiovasculares, Disponível em:<https://www.paho.org/bra/index.php?option=com_content&view=article&id=5253:doencas-cardiovasculares&Itemid=1096>. Acesso em 07 de jun de 2020.

REHEM DE SOUZA, RENILSON. O SISTEMA PÚBLICO DE SAÚDE BRASILEIRO. Biblioteca Virtual em Saúde, 2002. Disponível em:< http://bvsms.saude.gov.br/bvs/publicacoes/sistema_saude.pdf>. Acesso em 07 de jun. de 2020.

SEKARAN, UMA; BOUGIE, ROGER. Research Methods for Business, Seventh Edition. 2016.

SILVA DE OLIVEIRA, KELLI. Fatores de Risco em Pacientes com infarto agudo no miocárdio em um hospital privado de Ribeirão Preto – SP. Ribeirão Preto, 2004.

WIENS, JENNA; S SHENOY, ERICA. Clinical Infectious Diseases, Volume 66. 2017. Disponível em:< https://academic.oup.com/cid/article/66/1/149/4085880 >. Acesso em 07 de jun de 2020.