Pré-Processamento de Dados Clínicos do Consórcio TCGA

Danielle Barbosa Escobar[1], Geraldo Henrique Neto[2], Daniel G. Tiezzi[3]

[1, 2]Faculdade de Tecnologia de FATEC Ribeirão Preto (FATEC)

Ribeirão Preto, SP – Brasil

[3]Faculdade de Medicina de Ribeirão Preto – Departamento de Ginecologia e Obstetrícia – Setor de Oncologia Ginecológica e Mastologia – Universidade de São Paulo – Ribeirão Preto, SP – Brasil

[1]danielleb.escobar@gmail.com,
[2]geraldo.henriqueneto@fatec.sp.gov.br, [3]dtiezzi@usp.br

Resumo. *O câncer é uma das doenças que mais causa mortes em todo o mundo, são realizadas todos os anos diversas pesquisas relacionadas à ela para encontrar respostas em relação do porque ela ocorre e como ela pode ser evitada. Em vista disso, o objetivo deste artigo é realizar o pré-processamento de dados clínicos do TCGA (The Cancer Genome Atlas), com o objetivo de ajudar a encontrar respostas para essas perguntas. O pré-processamento foi realizado utilizando técnicas, como data clean, fazendo uso da linguagem R como ferramenta. A partir disso foi obtido um data frame de mais de 6000 linhas e 60 colunas, sendo que cada linha representa um paciente e cada coluna uma característica que sua doença possa ter apresentado.*

Abstract. *Cancer is one of the diseases that causes the most deaths worldwide, several researches are carried out every year related to it to find answers regarding why it occurs and how it can be avoided. In view of this, the aim of this article is to pre-process clinical data from TCGA (The Cancer Genome Atlas), in order to help find answers to these questions. Pre-processing was performed using data pre-processing techniques, such as data clean, using the R language as a tool. From that, a data frame of more than 6000 rows and 42 columns was created, with each row representing a patient and each column a characteristic that their disease could present.*

1 Introdução

Câncer é o nome dado para o conjunto de mais de 100 doenças que tem como característica, o crescimento desordenado de células que invadem tecidos e órgãos. Essas células dividem-se rapidamente e tendem a ser agressivas e incontroláveis, formando tumores que podem se espalhar para outras regiões do corpo (INCA, 2019).

Além de ser uma enfermidade silenciosa – não apresenta sintomas preocupantes nos primeiros estágios –, também é difícil de ser explicada. As causas do câncer ainda são um mistério em algumas neoplasias, porém já é comprovado que em algumas situações, como o câncer de pulmão, a causa possa ser fatores externos, como o

tabagismo. Assim também é comprovado que em outras situações, como o câncer de mama ou de colo de útero, que os fatores causadores possam ser genéticos.

Segundo Fleury (2008), para alguns tipos de câncer ocorre um padrão de hereditariedade, o que sugere a participação de fatores genéticos herdados, isto é, se em uma família um membro apresentar um determinado tipo de câncer a probabilidade de outro membro apresentar o mesmo, ou um semelhante é maior do que o resto da população .

Ainda segundo Fleury (2008), isso acontece porque os genes supressores ou oncogenes, que são responsáveis por não permitir a existência de células mutantes, não funcionam do jeito que deveriam. Em casos em que o gene é herdado, não significa que é certeza que o indivíduo terá câncer, mas sim uma maior probabilidade, pois os genes citados estão localizados em cromossomos autossômicos, portanto estão presentes em duas cópias. Mutações herdadas em somente umas das cópias não são suficientes para determinar a mudança de comportamento da célula, é necessário que a segunda cópia sofra mutação para que assim o indivíduo possa apresentar o câncer. Todavia, a probabilidade dessa pessoa que já tem uma cópia do cromossomo mutante desenvolver o câncer é maior do que de uma pessoa que não apresenta essa mutação.

Todas essas descobertas foram realizadas por meio de pesquisas relacionadas ao assunto, e com a ajuda de diversos pacientes que apresentavam os mesmo casos clínicos. E afim de ajudar em pesquisas como essa que, em 2005 foi criado o *TCGA* (*The Cancer Genome Atlas*) *Research Network* (TCGA, 2005), que é um programa que tem como objetivo estabelecer a importância genômica do câncer e estudar mais a fundo como ele funciona e as causas que o fazem acontecer. Depois de mais de 12 anos, com contribuições de mais de 11 mil pacientes, o TCGA produziu um repositório de valor imensurável à pesquisa contra o câncer (NHI, 2019).

E a partir de dados coletados pelo programa TCGA que foi realizado este trabalho, com o objetivo de realizar novas descobertas em relação ao câncer e suas causas. Utilizando o *data frame* coletado para encontrar padrões entre diversos tipos de câncer e realizar análises de dados futuras que possam ser a resposta para muitas perguntas.

Este artigo tem como finalidade documentar atividades relacionadas ao projeto que será apresentado como trabalho de finalização do Curso de Análise e

Desenvolvimento de Sistemas, assim como fazer considerações sobre o mesmo, apresentar os resultados e as fontes consultadas para o assunto.

2 Ferramentas

2.1 Linguagem R

O R é uma linguagem para computação estatística e gráficos. Foi criado em 1993, por Ross Ihaka e Robert Gentleman do departamento de Estatística da Universidade Auckland, Nova Zelândia (WIKIPEDIA, 2020). Ele é muito similar ao ambiente e a linguagem S, tanto que muitos códigos escritos em R podem ser executados em S e vice-versa (R Foundation).

É um Software Livre, isso é, "os usuários possuem a liberdade de executar, copiar, distribuir, estudar, mudar e melhorar o software"(Free Software Foundation, 2019). E é através disso que atualmente o R é mantido por uma comunidade de colaboradores voluntários que contribuem com o código fonte e com a criação de funcionalidades através de bibliotecas.

O R fornece uma grande variedade de técnicas estatísticas, como: modelagem linear e não linear, testes estatísticos clássicos, análise de série temporal, classificação e agrupamento. Uma das melhores vantagens de se usar R é a grande capacidade do software de realizar gráficos de boa qualidade, incluindo os símbolos e fórmulas matemáticas (R Foundation).

2.2 RStudio.

O RStudio é uma IDE (*Integrated Development Environment*) ou ambiente de desenvolvimento para o R. Ele tem 4 ambientes, o console, o editor de texto, histórico de variáveis usadas e um espaço para o gerenciamento das pastas de trabalho (RStudio).

Ele é disponibilizado tanto em edições comerciais como de código aberto, podendo ser usado em qualquer sistema operacional ou até mesmo via browser através do RStudio Server (RStudio).

3 Pré-Processamento de Dados

3.1 Procedimentos de Pré-Processamento de Dados.

Muitos fatores podem comprometer a qualidade dos dados e acabarem comprometendo um projeto inteiro. Por mais que se tenha um banco de dados com diversas variáveis e milhares de tuplas isso não significa que o trabalho será um sucesso. Segundo Han et al. (2016, p. 84), "Os dados têm qualidade se satisfizerem os requisitos do uso pretendido." e, afim de deixar os dados prontos para satisfazerem seus requisitos é necessário o pré-processamento.

Os problemas que podem ser resolvidos na fase de pré-processamento são diversos, porém os mais comuns são ausência de valores, dados ruidosos (*outliers*), atributos de naturezas distintas e redundância de dados. Com o objetivo de resolver essas imperfeições foram criados procedimentos de pré-processamento de dados, como por exemplo limpeza de dados (*data clean)*, integração de dados (*data integration*) e transformação de dados (*data transformation*) (SILVA, 2016).

Na limpeza de dados são corrigidos dois grandes problemas: ausência de valores e os ruídos (*outliers*). Esses problemas podem ser causados por várias razões, como descuido do usuário na hora de digitar os dados, usuário não quis inserir um valor ou não havia um valor para ser inserido, e algum erro em uma manipulação dos dados prévia (HAN, 2016) .

Nesses casos há várias formas de se corrigir, é claro que se deve ver se o problema é relacionado a *outliers* ou a ausência de valores. Se o problema for relacionado ao primeiro, pode-se realizar uma atenuação dos dados ou uma regressão, a partir dos dados que estão corretos, para substitui-los. Agora se for relacionado a ausência de valores, pode-se preencher os campos manualmente ou excluir as tuplas, se isso não for atrapalhar no resultado na análise futura (HAN, 2016).

Além disso, é necessário ressaltar uma característica muito importante em relação a estatística descritiva nessa fase do pré-processamento. Segundo Batista (2003), o processo de limpeza de dados é semiautomático, isto é, ele depende de uma pessoa que seja capaz de identificar os problemas nos dados, a natureza deles e, é claro, resolver esses problemas. É quase impossível encontrar os erros exemplificados anteriormente sem uma

pessoa para representar os dados em gráficos, e interpretá-los a fim de encontrar os problemas.

O procedimento de integração de dados serve para integrar dados de duas ou mais fontes diferentes. Quando se integra dados assim podem ser encontrados dois problemas: valores inconsistentes e redundância de dados. Um exemplo de valores inconsistentes seria quando se tem a variável "data de nascimento" e em um *dataset* e os dados estão armazenados com dia, mês e ano por números, enquanto no outro eles estão escritos por extenso. Em relação a redundância de dados os três principais fatores seriam, o uso de nomenclaturas diferentes para atributos equivalentes, a inserção de dados repetidos no conjunto por consequência de um erro de aquisição e o armazenamento de dados derivados de outros atributos. (SILVA, 2016)

As formas de se corrigir os erros exemplificados seriam; para valores inconsistentes pode-se remover o valor inconsistente, corrigir manualmente ou realizar a análise do esquema das fontes e dos conjuntos de dados, a fim de construir procedimentos de correção automática; para a redundância de dados é interessante realizar a redução dos dados, que pode ocorrer de forma horizontal, removendo exemplares, ou vertical, removendo atributos repetidos. Nessas situações, é interessante o procedimento procurar o que é útil ao invés de procurar os valores redundantes, e também é importante enfatizar que a redução não deve mudar a capacidade analítica do conjunto original. (SILVA, 2016)

O objetivo da transformação de dados é normalizar dados categóricos, sejam eles quantitativos ou qualitativos. Características que são modificadas durante essa fase são, por exemplo, o caso de dois sensores que mensuram com grandezas diferentes terem coletado dados de um mesmo fenômeno, ou quando se tem categorias avaliativas como (ruim, regular, bom, muito bom) e também quando se tem dois atributos distintos como "idade" e "salário", ambos com amplitudes diferentes, o que pode causar dificuldades em algoritmos que utilizam todos esses valores para compor um único valor de comparação entre esses exemplares (SILVA, 2016).

As características citadas devem ser mudadas a fim de melhorar os resultados da análise que será realizada. No primeiro caso citado, deve-se escolher uma grandeza e converter todas as outras para a escolhida. No segundo, é indicado converter os valores para bits, como exemplo o caso das categorias (ruim, regular, bom, muito bom) se

tornarem, respectivamente, (1000, 0100, 0010, 0001); e no último caso existem técnicas de normalizações que são apresentadas a partir de fórmulas, como por exemplo, *normalização z-score*:

$$z = \frac{x - \mu}{\sigma}$$

Equação 1. Normalização z-score

Fonte: (SILVA, 2016)

em que x é um valor do conjunto de valores **v**, z é o novo valor de x, μ é a média do conjunto **v** e σ é o desvio padrão. A partir da aplicação dessa fórmula os dados ficam normalizados em uma amplitude equivalente. (SILVA, 2016)

Ainda existem muitos outros procedimentos relacionados ao pré-processamento de dados, porém neste capítulo foram apenas retratados esses três pois são os que serão implementados no trabalho.

3.3 Representação de Dados.

A visualização de dados é uma das formas de apresentar informações em formato de imagens e gráficos. Ela permite que analistas e tomadores de decisão vejam resultados de análises visualmente, a fim de compreender conceitos difíceis e identificar novos padrões. É importante salientar também que a representação de dados não influencia somente no resultado da análise, mas também durante ela, novos padrões, *insights,* problemas e informações sobre aquele conjunto de dados são encontrados (SILVA, 2016).

Durante a realização do pré-processamento, e até mesmo depois dele, são utilizados vários tipos de gráficos para demonstrar dados que estão escondidos entre as milhares de linhas em uma tabela, e até mesmo em diversas tabelas. Um exemplo de gráfico muito utilizado é o *Boxplot*, ora representado na Figura 1.

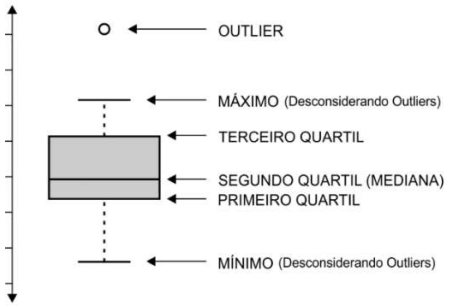

Figura 1. *Box Plot* com Outlier

Fonte: (OLIVEIRA, 2019)

O diagrama de caixa ou *Boxplot* é uma forma popular de se visualizar a dispersão dos dados. A partir desse gráfico é possível tirar algumas conclusões (SILVA, 2016):

- O segundo quartil é marcado pela mediana.
- *Outliers* podem ser representados como pontos fora das extremidades do gráfico, como o apresentado na Figura anterior.
- É possível a partir somente da Figura distinguir onde os dados estão mais concentrados e onde estão mais dispersos.

Este gráfico é extremamente importante quando se deseja entender o comportamento dos valores relatados por um atributo, e daí em diante entender como essas novas informações impactam no que se sabe sobre os dados.

Outro exemplo de gráfico muito importante é o Histograma. O Histograma é a representação gráfica de uma distribuição de frequências, como o exemplo apresentado na Figura 2.

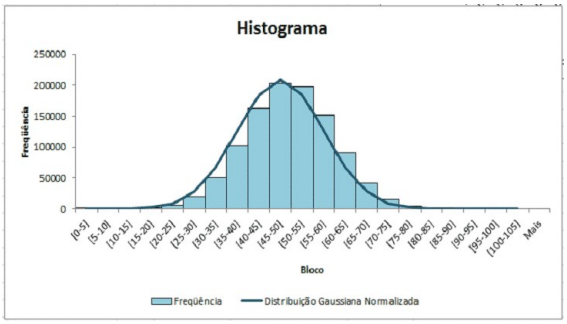

Figura 2. Histograma

Fonte: (ERICA, 2018)

Para cada faixa de valores é atribuída uma barra na qual a sua altura representa quantas vezes um valor apareceu, e é baseando-se nas barras que é criada a curva de dispersão. Essa curva pode representar particularidades sobre os dados, como por exemplo, a curva simétrica ou em forma de sino, que está presente na Figura 2, ela simboliza que valores equidistantes do valor modal tem a mesma frequência (SILVA, 2016).

Existem vários outros tipos de gráficos além do *Boxplot* e do Histograma, todos eles com o objetivo de mostrar padrões e informações escondidos dentro de um conjunto de dados. A principal escolha que se deve fazer é qual gráfico dará resultados mais satisfatórios para o *dataset* que se tem em mãos.

3.4 Sumarização de Dados

A sumarização de dados é um modo resumido de contextualizar informações acerca do seu conjunto de dados. Ela é extremamente utilizada para obter informações primárias e importantes sobre grandes conjuntos de dados, como: média, mediana, quartis e frequência. A sumarização é muito importante durante o pré-processamento, pois demonstra padrões e características dos dados sem realizar análises mais complexas.

Um exemplo de como obter a sumarização de dados é a função “summary(x)” existente na linguagem R. Ela realiza a condensação do objeto “x” selecionando-o e realizando a sumarização de cada atributo presente nele. Realizando essa função no IDE (*Integrated Development Environment* ou Ambiente de Desenvolvimento) RStudio, o retorno dessa função apresenta o menor valor, o primeiro e o terceiro quartil, a mediana, a média e o maior valor armazenado, assim como mostrado na Figura 3.

```
summary (x)
  crime_rate
Min.   : 0.00632
1st Qu.: 0.08204          airport
Median : 0.25651          NO :227
Mean   : 3.61352          YES:279
3rd Qu.: 3.67708
Max.   :88.97620
```

Figura 3 - summary(x)

Fonte: Os autores

Na Figura 3 podemos observar “crime_rate” e “airport” são os atributos e nas linhas abaixo estão a sumarização deles. Observando os números do primeiro, podemos perceber a presença de *outliers*, pois o valor máximo está muito distante dos outros valores, indicando a presença de ruídos, ainda que seja necessário realizar outras análises para se ter certeza, já é o indício do problema.

Em valores categóricos ou qualitativos, como os da variável “airport’, é mostrada a frequência. Ainda que os únicos valores presentes sejam “yes” e “no” é possível observar que a frequência dos dois é similar, algo que pode ou não influenciar no ambiente geral de onde foi retirado esse dado (SUMMARY STATISTICS).

Além disso a sumarização também pode ser retratada por meio de gráficos e medidas de dispersão. Esses métodos são mecanismos rápidos de se obter informações consideráveis da amostra de dados.

4 Metodologia

O primeiro passo para a realização do projeto foi a escolha de quais dados seriam utilizados para o pré-processamento. Tendo em base que o data frame final deve ser coerente, foram escolhidas 16 das 33 tabelas disponibilizadas pelo *TCGA Research Network* (TCGA, 2005), sobretudo se fossemos utilizar todo o conjunto de dados, o processo de análise exploratória futura poderia ser comprometida, já que entre essas 33 tabelas existem diferentes tipos de neoplasias, algumas que não tem nenhum tipo de relação com as outras.

As 16 tabelas selecionadas tem como padrão o fato de que seus pacientes sofreram de um subtipo de câncer chamado Adenocarcinoma, denomidado como uma neoplasia mais comum no humano depois dos tumores de pele, sendo um tipo de câncer que afeta as glândulas e o tecido epitelial dos órgãos excretores, os principais órgãos afetados são as mamas, a próstata, o útero, o estomago e o colón (ONCOMARKERS, 2018).

Após uma minuciosa seleção foi necessário identificar os atributos de interesse que as 16 tabelas tinham em comum, e também alguns que representavam importância para a análise genômica que esses dados irão passar. Como no caso do último, era possível de que esse atributo não existisse em alguma tabela, foi determinado que fosse adicionado o termo NA em seu lugar.

Tendo as tabelas e as colunas de interesse foi possível então realizar a junção, obtendo assim um *dataframe* de 6831 linhas e 60 colunas, sendo cada linha um paciente e cada coluna uma das variáveis de interesse explicadas anteriormente.

Depois da união foi realizado o *data clean* ou limpeza de dados, em que foram removidos ruídos e ausência de dados. Nas linhas, a ausência de dados era representada por "[Not Available]", ["Not Applicable]", "[Not Evalueted]" ou "[Unknwon]". Para realizar a padronização dessa ausência e deixá-la mais fácil de ser contabilizada, esses valores foram trocados para o termo NA, e também foram removidas linhas que poderiam se tornar ruídos, pois elas não continham nenhum dado, apenas o nome e código da coluna.

Depois de padronizar as linhas ausentes, foi contabilizado a porcentagem de dados ausentes em cada coluna, todas as colunas que tinham mais de 90% e não representassem mudança na capacidade analítica do conjunto original dos dados foram excluídas.

Para a realização do pré-processamento explicado anteriormente, foi utilizado a IDE RStudio. Como principais pacotes utilizados temos o'dplyr' e o 'data.table', além de funções nativas da IDE, o 'dplyr' é um pacote próprio para manipulação de dados, que fornece um conjunto consistente de funções para a realização do mesmo (WICKHAM, 2018), quanto ao 'data.table' ele fornece uma versão de alto desempenho de Rbase , com aperfeiçoamentos para facilidade de uso (DOWLEY, 2020).

5 Resultados

Dos problemas enfrentados podem ser descritos principalmente a ausência de dados e redundância no nome das colunas. Como explicado na Seção 3.1, isso ocorre por diversos fatores, no caso desse conjunto de dados, isso pode ter ocorrido pelo fato deles possuírem diferentes fontes, o que ocasiona a falta de padronização dos dados e dos nomes das colunas. Para a resolução desse problema também foi utilizado os exemplos mostrados na Seção 3.1.

Seguido da resolução desses problemas e dos passos evidenciados na Metodologia, obteve-se um *data frame* de mais de 6000 linhas e 60 colunas, cada linha representando um paciente e cada coluna uma característica da patologia acometida pelo mesmo, canalizando para um *data frame* considerado consistente e factível para uma análise genômica.

Os arquivos e códigos utilizados para a realização do pré-processamento e o *data frame* resultante estão disponíveis na íntegra através do GitHub (ESCOBAR, 2020).

6 Conclusão

A partir dos dados obtidos como resultado podemos concluir que os mesmos estão aptos para análise genômica. Infelizmente ainda não se pode afirmar que esses dados poderão responder, posteriormente ao processo de análise de dados, algumas das milhares de perguntas existentes em relação ao câncer, somente a análise genômica nos dará essa resposta, e mesmo tendo ela não significa que será fidedigna. O que é possível no momento é acreditar que, com a realização deste e de muitos outros trabalhos oriundos a análise de dados relacionados ao assunto, um dia chegaremos a obter mecanismos para alcançarmos a cura do câncer.

Referências

BATISTA, Gustavo Enrique A. P. A. Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. 2003. Tese (Doutorado em Ciencias de Computação e Matemática Computacional. – Universidade de São Paulo, São Carlos, 2003.

ESCOBAR, Danielle Barbosa. Repósitório contendo documentação, código-fonte e arquivos relacionados ao projeto. Disponivel em: <https://github.com/thedunny/TCGA> . Acesso em: 25 Nov 2020.

ERICA. Como Fazer um Histograma no Excel e Todos os Detalhes. Engenheira do Excel. Disponível em: <https://engenheiradoexcel.com.br/histograma-no-excel/>. Acesso em: 30 mai. 2020

DOWLE, Matt. Data.table. RDocumentation, 2020. Disponível em: < https://www.rdocumentation.org/packages/data.table/versions/1.13.21>. Acesso em: 08 Nov 2020.

HAN, J. et al. *Data Mining: Concepts and Techniques*. 3ª ed. Waltham: Elsevier, 2016.

INCA. O que é câncer?. Instituto Nacional do Cancer (INCA), 2019. Disponível em: < https://www.inca.gov.br/o-que-e-cancer>. Acesso em: 24 de Jun 2020.

FLEURY: medicina e saúde. Marcadores Genéticos de Predisposição ao Câncer de Mama, 2008. Disponível em: < https://www.fleury.com.br/medico/artigos-cientificos/marcadores-geneticos-de-predisposicao-ao-cancer-de-mama>. Acesso em: 24 de Jun 2020.

FREE SOFTWARE FOUNDATION. O que é software livre? O Sistema Operacional GNU, 2019. Disponivel em: < https://www.gnu.org/philosophy/free-sw.pt-br.html>. Acesso em: 22 Nov 2019.

NHI - National Cancer Institute. Outcomes & Impact of The Cancer Genome Atlas, 2019. Disponível em: < https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history>. Acesso em: 24 de Jun 2020.

OLIVEIRA, Bruno. *Boxplot*: como interpretar? . Oper Acelere a evolução através do significado de dados. Disponivel em: <https://operdata.com.br/blog/como-interpretar-um-boxplot/>. Acesso em: 30 mai. 2020.

ONCOMARKERS. Adenocarcinoma: tratamento e desafio no diagnóstico, 2018. Disponivel em: https://www.oncomarkers.com.br/adenocarcinoma-tratamento/. Acesso em: 22 Nov 2020.

RFOUNDATION. *What is R?. The R Project for Statistical Computing.* Disponivel em: < https://www.r-project.org/about.html>. Acesso em: 22 Nov 2020.

RSTUDIO. *Take control of your R code*. *RStudio*. Disponivel em: < https://rstudio.com/products/rstudio/>. Acesso em: 22 Nov 2020.

SILVA, L. et al. Introdução à Mineração de Dados: Com aplicações em R. 1ª ed. Rio de Janeiro: Editora Elsevier, 2016

SUMMARY STATISTICS: *Definition and Examples. Statistics How To*. Disponivel em: < https://www.statisticshowto.com/summary-statistics/>. Acesso em: 30 mai. 2020.

TCGA: *The Cancer Genome Atlas. NHI – National Cancer Institute.* Disponivel em: < https://www.cancer.gov/tcga>. Acesso em: 26 Nov 2020.

WICKHAM, Hadley. *A Grammar of Data Manipulation,* 2018. *RDocumentation.* Disponível em: <https://www.rdocumentation.org/packages/dplyr/versions/0.7.8>. Acesso em: 21 Jun 2020.

WIKIPEDIA. R (Linguagem de programação), 2020. Dísponivel em: <https://pt.wikipedia.org/wiki/R_(linguagem_de_programa%C3%A7%C3%A3o)>. Acesso em: 22 Nov 2020.