UNIVERSIDADE FEEVALE

EDUARDO EISMANN SOUZA

DESCOBERTA DE CONHECIMENTO EM DADOS SOBRE COVID-19

Novo Hamburgo
2021

EDUARDO EISMANN SOUZA

DESCOBERTA DE CONHECIMENTO EM DADOS SOBRE COVID-19

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale.

Orientador: Dra. Marta Rosecler Bez

Novo Hamburgo
2021

EDUARDO EISMANN SOUZA

Trabalho de conclusão do Curso Ciência da Computação, com título DESCOBERTA DE CONHECIMENTO EM DADOS SOBRE COVID-19, submetido ao corpo docente da Universidade Feevale, como requisito necessário para obtenção do Grau de Bacharel em Ciência da Computação.

Inclui Termo de Consentimento da Empresa/Entrevistado: [ ] Sim; [x] Não

Aprovado por:

_____

Prof.ª Me. Marta Rosecler Bez

Orientadora

_____

Professor avaliador

_____

Professor avaliador

Agradecimentos

Gostaria de agradecer ao meu avô, Urbano Eismann, que, além de avô, sempre foi um grande amigo, e sem a ajuda dele, provavelmente, eu não teria chegado até aqui.

À minha mãe, Clarice Eismann, e ao, meu irmão, Gustavo Eismann, que sempre me deram apoio e motivação.

Agradecer, também, ao Professor César Almeida, quem me apresentou algoritmos pela primeira vez, em 2008, e me instigou ao interesse em programação.

Um agradecimento especial à minha orientadora, Marta Bez, que aceitou este desafio, sempre me motivando e direcionando nos momentos de indecisão.

À Jéssica Schäfer, que me ajudou e orientou em momentos difíceis.

À Bob, um grande amigo, que mudou a minha visão da vida em relação a muitos conceitos.

Ao Sr. K, que indiretamente me orientou com sua sabedoria.

E não menos importante, a todas as pessoas que passaram e àquelas que estão comigo até hoje, que contribuíram de forma positiva e negativa para o meu crescimento como pessoa.

A todos os supracitados, muito obrigado!

RESUMO

O ano de 2020 ficou marcado pela pandemia do vírus COVID-19. Países de todo o mundo tiveram que mobilizar profissionais da área de saúde para combater a doença que se alastrou e contaminou pessoas nos cinco continentes. Dentro deste cenário catastrófico, este estudo tem como objetivo apresentar o desenvolvimento de um *dashboard* que torne possível a visualização de dados sobre testes de COVID-19. Foi desenvolvida uma pesquisa aplicada, a fim de demonstrar as características dos dados recolhidos de quatro hospitais brasileiros, assim, utilizando da abordagem qualitativa para submeter os dados à avaliação de usuários para tornar possível a descoberta de conhecimento sobre as bases exploradas. Para tornar possível o desenvolvimento deste *dashboard*, foram aplicadas técnicas de *KDD* sobre os dados públicos dos hospitais e utilizada a linguagem de programação R. Desta forma, é possível demonstrar, em forma de gráficos, as características em comum de pacientes infectados, para que profissionais de saúde possam melhor compreender determinados cenários. O *link* com o *dashboard* e um questionário foram enviados aos participantes do Grupo de Pesquisa em Computação Aplicada da Universidade Feevale. Como resultado, os usuários apresentaram aprovação diante da ferramenta disponibilizada, sendo avaliada como de fácil uso e compreensão, além de proporcionar uma visualização clara dos cenários selecionados.

Palavras-chave: Visualização de dados. Descoberta de Conhecimento. Análise de Dados. Covid-19. *Dashboard.*

ABSTRACT

The year 2020 was marked by the pandemic of the COVID-19 virus, countries around the world had to mobilize health professionals to fight the disease that spread and contaminated people on the five continents. Within this catastrophic scenario, this study aims to present the development of a dashboard that makes it possible to view data on COVID-19 tests. An applied research was developed in order to demonstrate the characteristics of the data collected from four Brazilian hospitals, thus, using the qualitative approach to submit the data to the evaluation of users to make possible the discovery of knowledge about the explored bases. To make the development of this dashboard possible, KDD techniques were applied to the public data of hospitals and the programming language R was used. In this way, it is possible to demonstrate, in the form of graphs, the common characteristics of infected patients, so that professionals health professionals can better understand certain scenarios. The link with the dashboard and a questionnaire was sent to participants in the Applied Computing Research Group at Feevale University. As a result, users submitted approval for the tool provided, being assessed as easy to use and understand, in addition to providing a clear view of the selected scenarios.

Keywords: Data Visualization. Knowledge Discovery. Data Analysis. Covid-19. Dashboard.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ARFF	*Attribute-Relation File Format*
CPO	*Chief Product Officer*
DM	*Data Mining*
DPOC	Doença Pulmonar Obstrutiva Crônica
DRC	Doença Respiratória Crônica
DT	*Decision Tree*
EC2	*Amazon Elastic Compute Cloud*
FAPESP	Fundação de Apoio à Pesquisa do Estado de São Paulo
GF	Grupo Fleury
HAE	Hospital Albert Einstein
HSL	Hospital Sírio Libanês
HSP	Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
IDE	*Integrated Development Environment*
INCA	Instituto Nacional do Câncer José Alencar da Silva
KDD	*Knowledge Discovery Database*
MERS	*Middle East Respiratory Syndrome*
RAM	*Randomic Access Memory*
RF	*Random Forest*
S	*Spike Protein*
SARS	*Severe Acute Respiratory Syndrome*
SSD	*Solid State Drive*
SVM	*Support Vector Machine*
TI	Tecnologia da Informação
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Por meio de uma pesquisa divulgada pelo Ibope, foi mostrado que 44% dos brasileiros apresentam sintomas de doenças respiratórias, sendo, em sua maioria, asma e bronquite crônica. (VIDALE, 2015). O trabalho foi realizado com 2010 brasileiros com idade entre 18 e 65 anos, em todas as regiões do país. O estudo revelou que 65% da prevalência dos sintomas respiratórios são localizados nos estados do sul do país.

O pneumologista e professor na Universidade Federal de São Paulo (Unifesp), Clystenes Odyr Soares, atribui essa característica ao clima da região, que é caracterizado por invernos mais rigorosos e secos, em comparação às demais regiões do Brasil. Soares explica que a incidência de temperaturas mais baixas e a pouca umidade relativa do ar são vistos como um grande risco para o sistema respiratório. (VIDALE, 2015).

Em estudo publicado por Iuliano et al. (2017, tradução nossa) para o site The Lancet, segundo estimativas realizadas pelo Centro de Controle e Prevenção de Doenças Respiratórias dos Estados Unidos (US-CDC) e da Organização Mundial de Saúde (OMS), é possível reforçar o argumento de Soares, no qual associa-se mais de 450 mil mortes por ano no mundo ligadas à gripe sazonal.

Hein (2020) comenta que outra doença, que também ataca o sistema respiratório, deixou marcado o ano de 2020 em todo o mundo, o Coronavírus (Covid-19). Registrado inicialmente como uma grave pneumonia, a doença que teve início na China se alastrou e contaminou pessoas nos cinco continentes. (GONÇALVES, 2020). No dia 11 de março de 2020, a Organização Mundial de Saúde declarou, em uma comitiva de imprensa, que classifica os casos de Covid-19 como uma pandemia. O vírus Covid-19 é o primeiro do tipo coronavírus a receber essa classificação.

Em notícia publicada por Farge (2020) para o site Agência Brasil, diferente de doenças respiratórias que apresentam tendências maiores em determinadas épocas do ano, para a Dra. Margaret Harris, porta-voz da Organização Mundial de Saúde, o Covid-19 não é caracterizado como uma doença respiratória sazonal, diferente da característica observada por Iuliano *et al*. (2017, tradução nossa) em outras doenças respiratórias.

Segundo estimativa feita pela British Broadcasting Corporation (BBC), entre 4 de janeiro e 15 de maio de 2020, houve cerca de 440 mil óbitos em todo o mundo por infecção de Covid-19, estimou-se também que, além dos dados oficiais, ainda há aproximadamente 130 mil mortes relacionadas ao vírus, visto que muitos países divulgam apenas os dados de óbitos ocorridos dentro de hospitais. (DALE; STYLIANOU, 2020).

De acordo com dados exibidos por REVISTA ISTOÉ (2021), estimativas da OMS informam que, desde dezembro de 2019 até junho de 2021, o número de pessoas infectadas por Covid-19 no mundo todo chegou à marca de 173.537.280, e destes infectados, o número de óbitos identificados foi 3.739.777. A OMS ainda ressalta que como há grande quantidade de casos de infecção direta e indiretamente vinculada ao vírus, esta marca pode ser de duas a três vezes superior ao registrado.

Ao analisar o trabalho de Borges (2006), o autor comenta que, com o crescimento acelerado do volume de dados, se tornou inviável a análise humana sem o auxílio da tecnologia, e ressalta que, ao utilizar métodos de redução de dimensionalidade, é possível obter uma melhor compreensão sobre os resultados gerados. Ao encontro deste estudo, Heckler (2018) apresenta o uso de técnicas de análise preditiva com a finalidade de obter informações e identificar tendências de abandono do tratamento. As técnicas aplicadas pelo autor visam acelerar o processo de descoberta de dados.

Complementar aos dois estudos supracitados, Costa (2019) desenvolveu uma biblioteca, com o uso da linguagem de programação R, para facilitar o processo de obtenção e manuseio de bases de dados, facilitando o trabalho de profissionais da área.

A base de dados utilizada neste trabalho é a FAPESP COVID-19 DataSharing/BR. A FAPESP disponibiliza os dados de pacientes e exames em quatro hospitais brasileiros. Os dados são referentes a pacientes e seus respectivos exames. Todos os pacientes disponibilizados na base são apresentados de forma anônima. Dentre os exames dos pacientes há hemograma, PCR, urina, diferencial manual, entre outros. Este trabalho utilizará dados disponibilizados pelo repositório COVID-19 Data Sharing/BR, Disponível em: *<https://repositoriodatasharingfapesp.uspdigital.usp.br>*.

Como objetivo principal, o presente estudo visa o desenvolvimento de um *dashboard* que torne possível a visualização de dados sobre testes de COVID-19, para auxiliar profissionais da área de saúde a obter novos conhecimentos sobre o vírus, assim, colaborando para o processo de tomada de decisão.

No capítulo seguinte, é abordado o tema Descoberta de Conhecimento, onde é destacado o grande volume de dados existente na área da saúde hoje, e o crescimento constante destes, ressaltando a importância da aplicação de KDD sobre uma base de dados. Após a aplicação desta técnica é possível a obtenção de conhecimento novo, e assim auxiliar profissionais na tomada de decisão. Neste capítulo é descrito cada um dos passos do processo de descoberta de conhecimento.

Para o capítulo subsequente, é descrito sobre doenças respiratórias e o impacto delas para o ser humano e, em conjunto com este tema, são informados dados no que se refere à atual pandemia de COVID-19 no Brasil e no mundo, a fim de mostrar o grande impacto e periculosidade do vírus para os sistemas de saúde.

Com o intuito de mesclar as áreas de conhecimento, o próximo capítulo descreve três trabalhos correlatos à esta temática, abordando assuntos como mineração de dados, descoberta de conhecimento e a utilização da linguagem de programação R. Relacionando os trabalhos já realizados, se torna possível a análise da base de dados, a busca por conhecimento e, por fim, a visualização do conhecimento gerado através de gráficos.

Subsequente aos trabalhos correlatos é descrito o capítulo de desenvolvimento, onde são abordados os passos para seleção das bases utilizadas, juntamente com a limpeza e pré-processamento, além de detalhes sobre o uso da linguagem de programação R e a interface de desenvolvimento RStudio. São descritos, também, os passos para o desenvolvimento da aplicação, relatando acerca da interface de usuário (UI), a interface de servidor (Server) e o arquivo de dados utilizado para a importação das bases.

O capítulo seguinte é relacionado à validação que o *dashboard* foi submetido. De maneira que o *dashboard* foi exposto ao grupo Pesquisa em Computação Aplicada da Universidade Feevale e, na sequência, foi aplicado um questionário de 23 perguntas ao grupo. As perguntas abordavam a usabilidade do *dashboard,* além de questionamentos sobre o caráter técnico. A partir das respostas dos integrantes deste grupo, são exibidos os gráficos a fim de compreender cada uma das características questionadas. Na sequência são apresentadas as conclusões deste trabalho.

2 DESCOBERTA DE CONHECIMENTO

O volume de dados na área da saúde vem crescendo de maneira exponencial e, com a presente pandemia, foi possível verificar que há muitos dados, mas pouco se sabe sobre seus significados e as relações que podem ter entre si, assim, dificultando e tornando mais lenta a compreensão por parte dos profissionais da saúde. À medida que o tempo passa, mais dados surgem, tornando cada vez mais problemático o processo de identificação de características entre essas informações e, sem o devido manuseio e extração, esses dados podem ser utilizados de maneira errônea.

Os dados podem ser lidos superficialmente por profissionais da saúde, mas as correlações entre laudos de pacientes distintos, em relação a um mesmo diagnóstico, tendem a não ser explorados e aprofundados, visto que o grande volume de dados pode também estar disperso entre mais bases de dados. Desta forma, torna-se necessário a utilização de tecnologia para acurar essa gama de informações e extrair, da melhor forma possível, resultados que possam contribuir para o avanço dos profissionais da área e na descoberta de conhecimento.

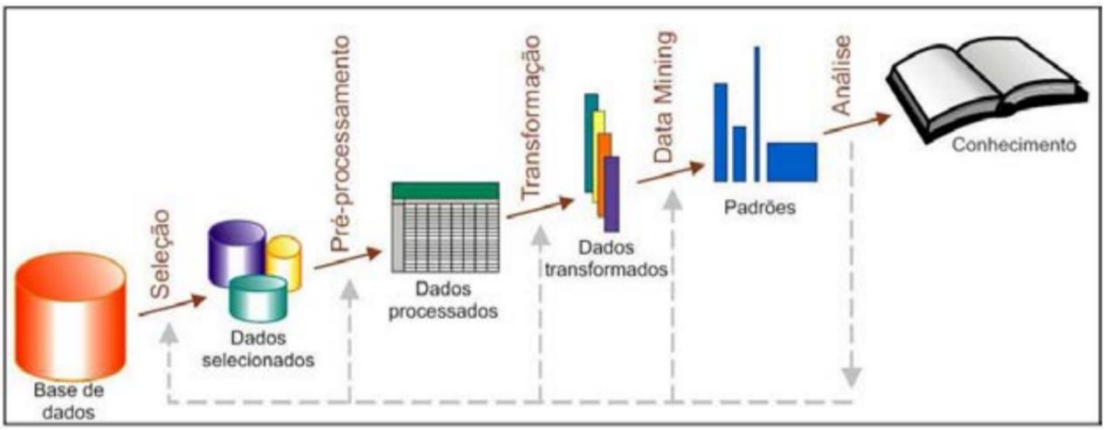
Visando encontrar informações em bases de dados de maneira automatizada, ao final da década de 80 surge o termo KDD (*Knowledge Discovery in Database*), Descoberta de Conhecimento em Bases de Dados. Este processo tem como um dos resultados identificar relações que, muitas vezes, não são percebidas por especialistas no assunto. Fayyad et al. (1996) definem o processo de KDD como "o processo não-trivial de identificação válida, em dados, novos, potencialmente úteis e finalmente com padrões compreensíveis".

2.1 ETAPAS DO PROCESSO DE KDD

De acordo com Borges (2006), a Mineração de Dados ou *Data Mining* (DM) é a etapa primordial do processo KDD, pois tem como objetivo a obtenção de padrões dos dados através de extração por meio de algoritmos.

FAYYAD et al. (1996) salientam que a denominação do processo de KDD era compreendida de forma errônea por muitos, assimilando-o apenas como DM, visto que o próprio *Data Mining* faz parte do processo de KDD, sendo a principal etapa. Na Figura 1, são ilustradas as etapas que compõem o processo de KDD, são elas: seleção dos dados; pré-processamento; transformação dos dados; Mineração de Dados e, por fim, a interpretação do conheciment.o. (FAYYAD et al., 1996).

Figura 1 - Etapas do processo KDD



Fonte: (GÓES; STEINER, 2012, p. 3741)

Ao longo deste processo, é possível encontrar informações explícitas, ou não, sobre o conhecimento, assim como outras informações inesperadas que, em um primeiro momento, não são identificadas nenhuma relação óbvia. Ainda assim, é possível que se façam presentes informações sem relevâncias significativas. (GÓES; STEINER, 2012).

Para que se tornasse possível o desenvolvimento deste trabalho, foi utilizado o processo de KDD definido por Fayyad et al. (1996), adentrando em cada uma das etapas supracitadas deste processo.

2.1.1 Seleção de dados

Góes e Steiner (2012) argumentam que o início do processo de KDD é dado pela escolha do conjunto de dados que será o foco principal da análise, desta maneira, definindo os atributos e registros que serão analisados.

Em essência, a seleção de dados compreende identificar quais informações existentes, em uma ou mais bases de dados, devem ser verdadeiramente consideradas no decorrer do processo de KDD. Informações como, por exemplo, o nome de uma pessoa, são completamente irrelevantes do ponto de vista de uma aplicação de KDD, principalmente quando o objetivo final da aplicação é prever comportamentos destas pessoas, ou mesmo como este determinado grupo de pessoas se comporta diante de um determinado cenário. Já informações, como data de nascimento, podem ser de total importância para a análise. É possível ainda separar a seleção dos dados em dois aspectos distintos, a escolha de atributos ou a escolha de registros, que são levados em consideração durante o processo de KDD. (GOLDSCHMIDT; PASSOS, 2005).

2.1.2 Pré-processamento

"Preparar insumos para uma investigação de mineração de dados geralmente consome a maior parte do esforço investido em todo o processo de mineração de dados." (WITTEN; FRANK, 2005, p. 52, tradução nossa).

Segundo Steiner et al. (2006), nesta fase, é determinada a qualidade dos dados, com o intuito de melhorar a eficácia dos algoritmos de classificação. Para tal, são eliminados dados redundantes, possíveis ruídos que venham a ser detectados e divergência nos dados. Além disso, outras técnicas podem ser aferidas, como reduzir o número de variáveis para que os métodos estatísticos possam ser melhor aplicados.

De um modo geral, o pré-processamento abrange as funções relacionadas à captação, organização, ao tratamento e à preparação dos dados para a etapa de Mineração de Dados, sendo esta a etapa de fundamental importância no que diz respeito à relevância durante o processo de descoberta de conhecimento. (GOLDSCHMIDT; PASSOS, 2005).

2.1.3 Transformação

Na sequência do pré-processamento, a etapa de transformação de dados visa tornar possível a criação de um modelo de dados analítico, em que a precisão e validade do resultado final será dependente de como a estruturação e apresentação das entradas foram concebidas. Esta fase incide em converter os valores para formatos padrões, assim, restringindo o número total de variáveis usadas, ou mesmo categorizá-las em concordância com o seu valor. (CABENA, 1998).

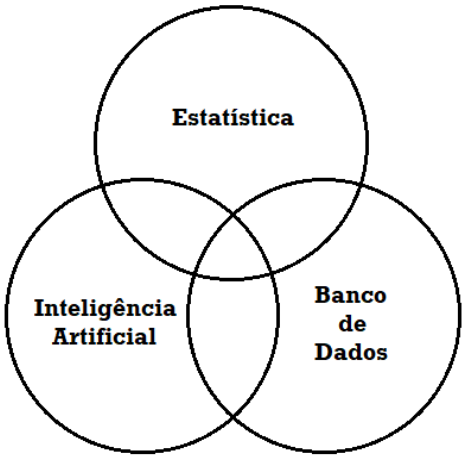
De acordo com Maimon e Rokach (2010), a transformação de dados tem como etapa inicial a preparação e geração de melhores dados para a mineração de dados. Dentre os métodos incluídos na redução de dimensão, estão a seleção e extração de recursos e amostragem de registro, já para a transformação de atributos, são utilizados a discretização de atributos numéricos e transformação funcional. Os autores reforçam a ideia de que esta é uma etapa crucial para o sucesso de qualquer projeto KDD, além de ser extremamente específica ao projeto. É possível utilizar, de exemplo, exames médicos, nos quais o quociente de atributos pode ser o fator mais importante, e não necessariamente cada atributo individualmente. Se utilizarmos como exemplo o ambiente de marketing, é preciso considerar efeitos além do nosso controle, como esforços e questões temporais, além de considerar efeitos do acúmulo de publicidade.

2.1.4 Mineração de dados

Vista como a etapa mais importante em todo o processo KDD, pois é nesse momento que são aplicadas técnicas e algoritmos, a fim de analisar os dados obtidos, utilizando-se de heurísticas ou meta-heurísticas, com o intuito de encontrar padrões. São diversos os possíveis métodos a serem aplicados, dentre eles, destacam-se: Árvore de Decisão, Redes Neurais e Algoritmos Genéticos. (GÓES; STEINER, 2012).

A Mineração de Dados, segundo Carvalho (2000), é uma área interdisciplinar que abrange, principalmente, estatística, inteligência artificial e banco de dados (Figura 2). Esta afirmativa é dada como coesa, pois, utilizando algoritmos de DM, torna-se possível realizar várias medidas estatísticas e, assim, por exemplo, classificar ou relacionar itens em uma base de dados.

Figura 2 - Ilustração da relação de interdisciplinaridade na fase de DM



Fonte: adaptada de Carvalho (2000, p. 25)

Para Rezende (2005), no que se trata do processo de mineração de dados, existem inúmeras classes de usuários que exercem relação, e o sucesso desta etapa está ligado à interação bem sucedida entre todas essas classes, destacando-se, principalmente, três: especialistas do domínio, analista e o usuário final, que são detalhados da seguinte forma:

- Especialistas do domínio: São responsáveis por oferecer o devido apoio para a execução do processo, pois é compreendido que estes possuem grande conhecimento e domínio sobre a aplicação;

- Analista: Tem conhecimento profundo sobre todas as etapas que abrangem o processo, além de ser o usuário especialista no procedimento de extração de conhecimento.
- Usuário final: Após a obtenção do conhecimento já extraído pelo analista, o usuário final utiliza estas informações para a tomada de decisão.

É citada por Borges (2006) a principal etapa do KDD e, em suma, sua finalidade é a extração de arquétipos a partir de dados obtidos, a fim de detectar padrões de comportamento, sendo, também, nesta etapa, a escolha dos algoritmos que realizarão a procura por conhecimento implícito e útil nas bases de dados e, à vista disso, qual é a representatividade destes elementos no mundo real.

2.1.5 Análise de dados

A análise de dados, também sendo referenciada como interpretação ou representação do conhecimento, por alguns autores, é a fase que sucede a etapa de mineração de dados, sendo vista, também, como o último ponto do processo KDD. Neste estágio, deve-se interpretar todo o conhecimento apresentado, verificando a proeminência de informações essenciais na busca de características que denotem a eficácia dos métodos aplicados na etapa anterior. Na hipótese de o analista julgar o conhecimento como inválido, o processo deve ser reiniciado, averiguando cada nível do KDD, em busca de melhorias e, quando necessário, refazer partes do processo, até que o conhecimento obtido seja julgado como válido. (GOLDSCHMIDT; PASSOS, 2005).

O foco principal da análise dos dados é facilitar a compreensão do conhecimento descoberto, utilizando medidas da qualidade da solução e da percepção de um analista de dados. Na sequência, será feita a consolidação deste conhecimento em forma de relatórios demonstrativos, juntamente com a documentação referente às informações relevantes transcorridas em cada nível do KDD. (BORGES, 2006).

Este capítulo apresentou os principais conceitos da descoberta de conhecimento que são utilizadas neste trabalho, com foco na análise de exames de doenças respiratórias, em especial o Covid-19. O tema doenças respiratórias será tratado no próximo capítulo.

3 DOENÇAS RESPIRATÓRIAS

Por meio de uma pesquisa divulgada pelo Ibope, foi mostrado que 44% dos brasileiros apresentam sintomas de doenças respiratórias, sendo, em sua maioria, asma e bronquite crônica. O trabalho foi realizado com 2010 brasileiros com idade entre 18 e 65 anos, em todas as regiões do país. O estudo revelou que 65% da prevalência dos sintomas respiratórios estão localizados nos estados do sul do país.

O pneumologista e professor na Universidade Federal de São Paulo (Unifesp), Clystenes Odyr Soares, atribui essa característica ao clima da região, que é caracterizado por invernos mais rigorosos e secos, em comparação às demais regiões do Brasil. Soares explica que a incidência de temperaturas mais baixas e a pouca umidade relativa do ar, são vistos como um grande risco para o sistema respiratório. (VIDALE, 2015).

Em estudo publicado por Iuliano et al. (2017, tradução nossa) para o site The Lancet, segundo estimativas realizadas pelo Centro de Controle e Prevenção de Doenças Respiratórias dos Estados Unidos (US-CDC) e da Organização Mundial de Saúde (OMS), é possível reforçar o argumento de Soares, no qual associa-se mais de 450 mil mortes por ano no mundo ligadas à gripe sazonal

3.1 DOENÇAS RESPIRATÓRIAS CRÔNICAS (DRCS)

De acordo com Goulart (2011), as DRCs são responsáveis por cerca de 7% de óbitos no mundo, representando aproximadamente 4,2 milhões de pessoas, já a DPOC afeta mais de 200 milhões em todo o mundo. Conforme dados divulgados pela Secretaria de Vigilância em Saúde (2016), entre os anos de 2003 e 2013, os registros de óbitos por DRCs no Brasil superaram 685 mil.

Segundo Rodrigues et al. (2002), as DRCs podem ser classificadas de duas formas: restritivas ou obstrutivas. Esta classificação é dada de acordo com o impacto causado nas funções pulmonares, sendo que ainda há doenças que unem essas duas características, que são chamadas de doenças respiratórias mistas.

3.1.1 Doenças restritivas

Barreto (2002) explica que as doenças restritivas se caracterizam por anomalidades hipodinâmicas, tanto neurais como musculares, gerando redução da capacidade pulmonar total. Complementar a essa informação, Sperandio et al. (2016) ressaltam que outro fator a se

observar sobre isso é a redução dos volumes pulmonares e, com o avanço da idade dos pacientes, a gravidade da doença tende a aumentar.

É sugerida a existência de transtornos restritivos sempre que for apresentada redução da capacidade vital que não seja explicada por distúrbio obstrutivo. A capacidade vital é compreendida pelo volume de ar corrente, isto é, o volume de ar inspirado e expirado de maneira espontânea pelo corpo ao longo de cada ciclo respiratório, além dos volumes inspirados e expirados de forma natural durante a respiração. De forma geral, para este grupo de doenças, a redução da capacidade pulmonar total pode ser a única característica correlacionada. (BARRETO, 2002).

3.1.2 Doenças obstrutivas

Como característica principal, as doenças obstrutivas apresentam redução do fluxo de ar máximo deslocado pelos pulmões. De acordo com Pellegrino et al. (2005), um dos fatores que proporciona esta característica é decorrido do estreitamento das vias aéreas no decorrer da expiração.

Costa e Jamami (2001) complementam este quadro salientando que, dependendo da gravidade do cenário, uma obstrução pulmonar pode ser irreversível, como exemplo, é citado a bronquite crônica e o enfisema pulmonar.

A Sociedade Paulista de Pneumologia e Tisiologia (2008) afirma que a medida do volume expiratório forçado no primeiro segundo, é uma das principais formas de diagnóstico da obstrução do fluxo aéreo. Este critério representa o volume máximo que um indivíduo é capaz de atingir no primeiro segundo de uma expiração máxima.

3.2 INFLUENZA

Conhecida popularmente por gripe, a influenza é uma doença infecciosa de origem viral que ataca o sistema respiratório. O vírus responsável pela influenza é o *Myxovirus Influenzae* e este subdivide-se nos tipos A, B e C.

Para Cox e Fukuda (1998), o aspecto para sofrer variações antigênicas frequentes e sem padrão determinado coloca a influenza em destaque no que se refere a doenças infecciosas emergentes.

Devido à natureza fragmentada do material genético da influenza, ela apresenta altas taxas de mutação durante a fase de replicação. Estas mutações acontecem de forma

independente e, devido a isso, ocasionalmente, ocorre o aparecimento de variantes para as quais a população ainda não desenvolveu imunidade. (FORLEO-NETO et al., 2003).

Com a chegada de estações mais frias como o inverno, regularmente, pessoas que têm doenças respiratórias sofrem com as vias aéreas ressecadas e apresentam maior dificuldade para respirar. Dentro deste cenário, as pessoas desenvolvem maior propensão a contrair gripes e resfriados, sendo um dos fatores a frequência de estarem em ambientes fechados e sem ventilação por longos períodos, em função do frio. Portanto, pacientes com Doenças Respiratórias Crônicas (DRCs) e Doença Pulmonar Obstrutiva Crônica (DPOC) apresentam maior chance de desenvolver crises, e em decorrência disso, serem hospitalizados. (REVISTA DA CIDADE, 2018).

3.2.1 Influenza A, B e C

Forleo-Neto et al. (2003) explicam que os vírus da influenza são subdivididos entre os tipos A, B e C. Dentre esses, o agente da Influenza A é o que apresenta maior variabilidade. O autor ainda esclarece que os vírus da influenza são os únicos com a habilidade de causar epidemias anuais recorrentes, porém, com chances menores de originar uma pandemia, atingindo quase todas as faixas etárias em um curto espaço de tempo. Esses fatores devem-se a alta variabilidade e capacidade de adaptação do vírus.

Segundo detalhamento fornecido pelo site Sinopse Pediátrica (2016), os aspectos que definem cada um dos subtipos de influenza são descritos a seguir:

- Influenza A: As maiores epidemias de influenza são causadas por este subtipo. Periodicamente, o vírus do tipo A sofre alterações em sua estrutura, caracterizando-se pela rápida variação antigênica e, assim, apresentando mais subtipos derivados da cepa A. As proteínas localizadas em sua superfície, hemaglutina (H) e neuraminidase (N), classificam os subtipos variados da influenza A. A proteína H está ligada à infecção das células do sistema respiratório, em que ocorre a proliferação do vírus, já a proteína N facilita a saída das partículas virais do interior das células infectadas. Tendo origem filogenética em aves aquáticas, o vírus do tipo A causa infeção em várias espécies de vertebrados.
- Influenza B: Dentre os três subtipos, é o que apresenta menor índice de variabilidade, além de sua transmissão ser exclusivamente entre humanos. Outra característica é a associação a epidemias localizadas, onde pode haver a predominância de cepas B.

- Influenza C: Com área de infecção restrita a humanos e suínos, a influenza C é considerada antigenicamente estável, provocando enfermidade, em grande parte das vezes, sem manifestação de sintomas, sendo que raramente causa uma doença grave.

O Quadro 1, a seguir, complementa as características de cada um dos subtipos supracitados.

Quadro 1 – Diferenças entre Influenza A, B e C

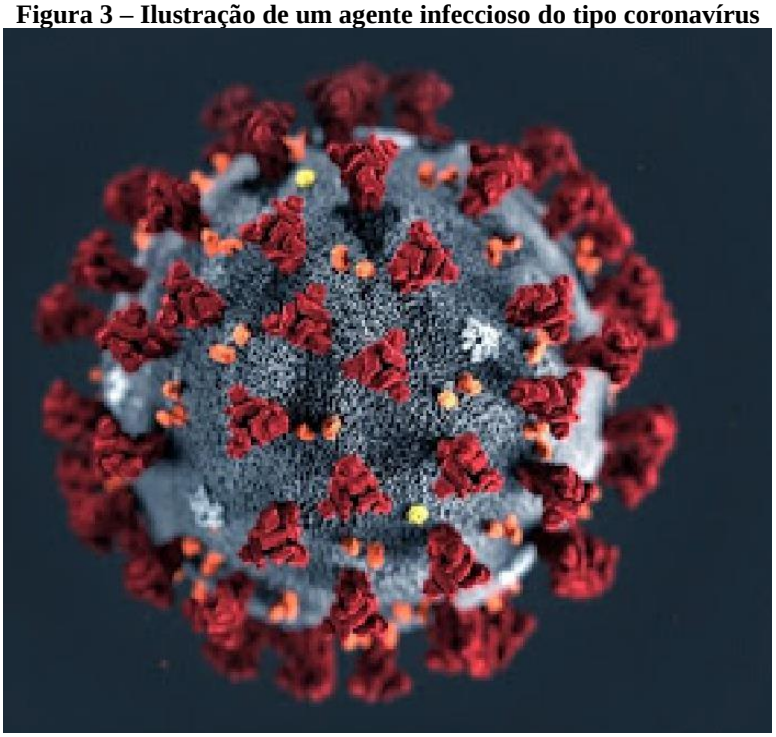
	Influenza A	Influenza B	Influenza C
Hospedeiro	Humanos, suínos, equinos, aves, outros mamíferos marinhos e terrestres	Humanos e mamíferos marinhos	Humanos e suínos
Características Epidemiológicas	Causa epidemias e pode causar pandemias	Causa epidemias e pode causar pandemias	Sem sazonalidade marcada

Fonte: adaptado de Sinopse Pediátrica (2016)

3.2.2 Coronavírus

De acordo com o Governo do Brasil (2020), o coronavírus pertence a uma família de vírus que causa infecções respiratórias. Em 1937, este agente infeccioso foi detectado pela primeira vez em humanos, porém teve sua definição e descrição como coronavírus apenas em 1965, quando uma análise por microscópio revelou sua aparência. O nome corona vem da língua espanhola. Assim, o vírus recebeu esse nome por causa de suas características visuais que se assemelham a uma coroa. A Figura 3 apresenta uma imagem de um agente infeccioso do tipo coronavírus.

Figura 3 – Ilustração de um agente infeccioso do tipo coronavírus



Fonte: GOVERNO DO BRASIL (2020)

O novo vírus, que tem atraído a atenção das autoridades de saúde em todo o mundo, foi descoberto em 31 de dezembro de 2019, e recebeu o nome técnico de Covid-19, apresentando as mesmas aparências relacionadas à coroa. Por esta razão, tem sido chamado de novo coronavírus.

O Governo do Brasil (2020) informa que, apesar das autoridades de saúde estarem em alerta, a maioria das pessoas se infecta com os coronavírus comuns ao longo da vida, dos quais as crianças pequenas são as mais propensas a se infectarem com essas variantes. Adicionalmente, é informado que os tipos regulares que infectam humanos são o alpha coronavírus 229E e NL63, e o beta coronavírus OC43, HKU1.

Dentre os tipos de coronavírus indicados pelo Governo do Brasil (2020), até o momento, os tipos de coronavírus existentes mais conhecidos são:

- Alpha coronavírus 229E e NL63
- Beta coronavírus OC43 e HKU1
- SARS-CoV
- Mers-CoV
- Sars-CoV-2 (Covid-19)

Gruber (2020), professor do Departamento de Parasitologia do Instituto de Ciências Biomédicas da USP, descreve algumas das características dos vírus da família *coronaviridae*,

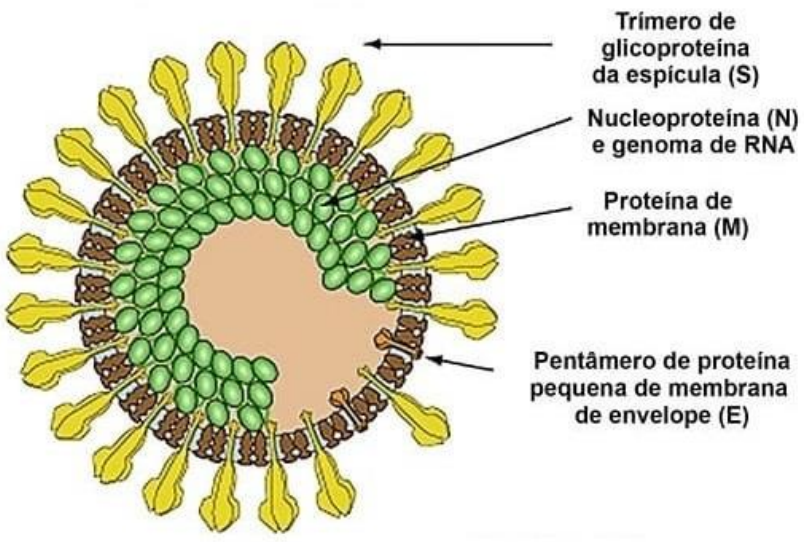
como a infecção tanto em seres humanos quanto em outros animais, o tamanho aproximado de 125 nanômetros de diâmetro, e a sua superfície, que apresenta projeções em forma de espícula são formadas pela proteína S (*Spike Protein*). Estas espículas são responsáveis pela adesão do vírus nas células do hospedeiro, para, assim, iniciar o processo de interiorização, onde ocorre a fusão entre a membrana viral e a célula proporcionando a entrada do vírus no citoplasma.

Têm-se noticiado, até o momento, a existência de sete espécies de vírus que infectam humanos. Neste grupo, três deles produzem doenças graves, são elas, respectivamente: Sars-Cov e Sars-Cov-2, que são agentes da pandemia de SARS, e o Mers-Cov, que é causador da MERS. Já os coronavírus causadores de doenças com sintomas mais brandos são relativos aos vírus HKU1, NL63, OC43 e 229E. (GRUBER, 2020).

Dentre os coronavírus comuns, que infectam seres humanos, estão incluídos 229E, NL63, OC43 e HKU1. Estes causam doenças em níveis que variam de leves a moderados no trato respiratório superior, como o resfriado comum. Ao longo da vida, as pessoas tendem a ser infectadas com um ou mais desses vírus. (CENTERS FOR DISEASE CONTROL AND PREVENTION, 2020, p. 1, tradução nossa).

Na ilustração a seguir (Figura 4), adaptada de Gruber (2020), é demonstrada a estrutura viral da SARS, na qual é possível observar algumas das características supracitadas, como a presença de espículas S. O genoma de RNA e a Nucleoproteína estão localizados ao centro da estrutura, estes são envelopados por uma membrana de proteína.

Figura 4 – Estrutura do coronavírus da SARS



Fonte: adaptada de Gruber (2020)

3.2.3 Covid-19

Com o constante crescimento de casos de COVID-19 em todo o mundo, causados pelo vírus SARS-CoV-2, foram feitas muitas comparações com o vírus Influenza e suas variantes, causadores da gripe. A semelhança existente é que ambos causam doenças respiratórias, porém existem importantes diferenças entre os dois vírus e a forma como eles se propagam. São essas nuances que precisam ser identificadas a fim de que a saúde pública consiga melhor entender e mensurar as implicações, bem como as medidas que devem ser tomadas em resposta a cada vírus. (BIOEMFOCO, 2020).

Azevedo (2020) explica que zoonose são doenças infecciosas que podem ser transmitidas de forma natural de animais para seres humanos e vice-versa. Em concordância à essa afirmação, Lima (2020) informa que, em maio de 2020, o especialista em zoonoses e doenças alimentares da OMS, Peter Ben Embarek, comunicou que a Covid-19 foi originada em morcegos. Esta afirmação é dada pela semelhança metagenômica da Sars-cov-2 com o coronavírus encontrado nestes mamíferos voadores, lembrando que morcegos são conhecidos por serem reservatórios de coronavírus.

As chamadas doenças zoonóticas ou zoonoses, contemplam diversos tipos de enfermidades transmitidas de animais para humanos. Esta transmissão pode ocorrer de várias maneiras: pelas picadas de insetos e animais, ao passar a mão ou manipular animais doentes, ao consumir carne malpassada, leite não pasteurizado ou água contaminada. Alguns tipos de patógenos que podem ser transmitidos pelos animais aos humanos são bactérias, parasitas, fungos e vírus. (CUTHBERT, 2020).

De acordo com Slayer (2017), doenças zoonóticas emergentes representam uma ameaça não apenas à saúde dos animais e humanos, mas também à segurança sanitária global. Estima-se que 60% das doenças infecciosas conhecidas e até mesmo 75% das doenças infecciosas emergentes são de origem zoonótica. Em todo o mundo, as doenças infecciosas são responsáveis por 15,8% de todas as mortes e 43,7% das mortes em países com baixos recursos.

Também, é estimado que as zoonoses sejam responsáveis por 2,5 bilhões de casos de doenças humanas e 2,7 milhões de mortes humanas em todo o mundo a cada ano, sendo responsáveis por algumas das epidemias mais devastadoras, em comparação à epidemia de Ebola, de 2014, que foi responsável por 11.316 mortes e US$ 2,2 bilhões em perdas econômicas, enquanto, a cada ano, a raiva, *Lyssavirus*, é responsável por aproximadamente 59.000 mortes humanas e cerca de US$ 8,6 bilhões em perdas econômicas em todo o mundo. (SLAYER, 2017, p.1, tradução nossa).

Os primeiros relatos de Covid-19 foram datados em 12 de dezembro de 2019, na cidade de Wuhan na China, contudo, houve um caso clínico com sintomas similares em 1 de dezembro de 2019. Após análise do fluído broncoalveolar, foi identificado um vírus com genoma que se mostrou semelhante aos coronavírus causadores da SARS e MERS. Inicialmente, o vírus foi denominado WHCV e mostrou alta similaridade genômica com o vírus Bat SL-CoVZC45, coletado de um morcego na China. Partindo deste resultado, acreditou-se que esta poderia ser uma possível origem do novo coronavírus. Posteriormente à denominação de WHCV, o vírus passou a ser classificado como Sars-Cov-2. (GRUBER, 2020).

Segundo estimativa feita pela British Broadcasting Corporation (BBC), entre 4 de janeiro e 15 de maio de 2020, houve cerca de 440 mil óbitos em todo o mundo por infecção de Covid-19. Estimou-se também que, além dos dados oficiais, ainda há aproximadamente 130 mil mortes relacionadas, visto que muitos países divulgam apenas os dados de óbitos ocorridos dentro de hospitais. (DALE; STYLIANOU, 2020).

De acordo com a notícia publicada em 31 de agosto de 2020, pelo jornal português Público, em relatório da OMS, é descrito que o impacto do Covid-19 abalou os sistemas de saúde em todo o mundo com severas interrupções em serviços essenciais de saúde em quase todos os países. A OMS classifica como serviços essenciais de saúde condições crônicas, saúde mental, reprodutiva, materna e infantil. (PÚBLICO, 2020, tradução nossa).

A mortalidade por COVID-19 aparenta ser mais alta quando comparado com a Influenza, especialmente quando se fale da influenza sazonal. Mesmo que a verdadeira mortalidade por COVID-19 leve algum tempo para ser completamente compreendida, os dados que se têm até o momento indicam que a taxa de mortalidade bruta está entre 3% e 4%, já para a gripe sazonal, esta taxa geralmente se mantem abaixo de 0,1%. (BIOEMFOCO, 2020).

O site Coronavírus Brasil (2020) foi desenvolvido com o intuito de ser o veículo oficial, no Brasil, para a comunicação no que se trata da situação epidemiológica de COVID-19 no país. São realizadas atualizações diárias sobre os casos confirmados e óbitos relacionados ao coronavírus. Os dados são atualizados pelo Ministério da Saúde através do repasse de informações fornecidas pelas Secretarias Estaduais das 27 unidades federativas.

4 TRABALHOS CORRELATOS

Com o intuito de aprofundar os estudos no que se relaciona a DM, KDD e ao uso da linguagem de programação R, foram selecionados três trabalhos que se assemelham no que tange à utilização destas tecnologias O objetivo é auxiliar o trabalho humano, para que empecilhos, principalmente, no tempo e na precisão dos dados, não aconteçam.

Os trabalhos foram selecionados de maneira a encontrar dentre os temas a descoberta de conhecimento, análise de dados e desenvolvimento de aplicações utilizando a linguagem R. Os trabalhos dos autores HECKLER (2018) e COSTA (2019) foram encontrados no site do TC-*Online*, da Universidade Feevale. Já o trabalho de BORGES (2006) foi encontrado por meio de pesquisas em artigos relacionados ao tema descoberta de conhecimento.

4.1 REDUÇÃO DE DIMENSIONALIDADE EM BASES DE DADOS DE EXPRESSÃO GÊNICA (BORGES, 2006)

Segundo análise desenvolvida por Borges (2006), em bases de dados em pesquisas na área de expressão gênica, foi observado que, com o crescimento acelerado do volume de dados, a análise humana se tornou inviável sem o auxílio da tecnologia. Os dados avaliados apresentavam grande número de atributos e um pequeno volume de amostras, desta maneira, comprometendo o desempenho do algoritmo de mineração de dados.

Utilizando métodos de redução de dimensionalidade, foi possível a remoção de dados redundantes e irrelevantes, assim, melhorando a compreensão sobre os resultados gerados e melhor identificando cada atributo selecionado e seu respectivo nível de expressão.

Os métodos aplicados no presente estudo são: a seleção de atributos e a projeção aleatória. Ambos serão utilizados como uma etapa de pré-processamento em DM.

O objetivo de se utilizar a seleção de atributos é descobrir um subconjunto de dados relevantes para uma tarefa alvo, assim, tornando mais eficiente o processo de aprendizagem, além da redução da dimensionalidade. De acordo com Borges (2006), este processo mostrou resultados promissores nas bases em questão.

Outro fator observado foi referente ao uso do método de projeção aleatória, em que foi percebida a diminuição do custo computacional, produzindo resultados significativos quando relacionado principalmente a atributos.

No estudo de Borges (2006), durante o desenvolvimento do trabalho, foram seguidas etapas do processo de descoberta de conhecimento, baseando-se em três etapas principais: pré-

processamento, DM e pós-processamento. Em conjunto, foi utilizada a versão 3.4 do software *Weka* (*Waikato Environment for Knowledge Analysis*) para a execução de uma parcela dos experimentos. Foi utilizado o *Weka*, pois conta com implementações de algoritmos com diversas técnicas de mineração de dados, como [WIT05], [SCU04] e [SCU06].

A concentração de esforços inicial foi direcionada para a classificação dos conjuntos de dados utilizando todos os atributos e, na sequência, foi feita a seleção dos atributos e subsequente a esta etapa, a classificação dos subconjuntos. Já na terceira fase, após ser feita a execução do método de projeção aleatória, foram aplicados os algoritmos de classificação. Na quarta e última parte desse processo, foram utilizados dois métodos em conjunto: a projeção aleatória e a seleção de atributos, para, posteriormente, os subconjuntos serem submetidos aos algoritmos de classificação.

Inicialmente foram selecionados cinco conjuntos de dados que seriam estudados, estes extraídos de repositórios já formatados como arquivos "*arff*", do software *Weka*. Após análise detalhada, verificou-se a quantidade de atributos, número de classes e número de amostras, totalizando 47 exemplos.

Contemplando a etapa de pré-processamento, foi feita a redução da dimensionalidade dos dados, utilizando dois métodos: seleção de atributos e projeção aleatória, na qual, em um primeiro momento, essas duas técnicas foram executadas separadamente, para, somente depois, serem processadas em sequência, sendo a projeção aleatória a primeira e a seleção de atributos em subsequência.

Em prosseguimento da fase de seleção de atributos, foram utilizados dois tipos de buscas em conjunto com duas abordagens de medidas de avaliação. Para as buscas, foram usadas as técnicas sequencial e aleatória, já para a avaliação, o filtro e o *wrapper* tiveram papel nesta fase. Para o uso do filtro, foram utilizadas duas medidas conhecidas, diferente do *wrapper* que utilizou algoritmos de DM para a classificação.

Um dos principais objetivos do estudo teve o foco na redução da quantidade de atributos em bases de micro arranjos, de modo que este redimensionamento não pode ser excessivo, pois, desta forma, afetará o poder de discriminação do classificador. Sendo assim, é essencial verificar a variação do comportamento do classificador com base na quantidade de atributos, para que seja possível estimar o dimensionamento ideal.

Borges (2006) aplicou dois métodos de redução de dimensionalidade: a seleção de atributos e o método de projeção aleatória. Inicialmente, cada método foi executado individualmente e, na sequência, os dois métodos foram executados em conjunto. Subsequente

ao método de projeção aleatória, foram aplicados os algoritmos de seleção de atributos e, para tal, foram usadas duas abordagens: filtro e *wrapper*.

Foi observada uma melhora significativa quanto ao uso desses métodos. Isto deve-se ao fato de que, mesmo nos piores casos, a taxa de acerto do classificador foi superior se comparado ao aplicado sobre as bases de dados com todos os atributos, resultando em uma grande redução na quantidade destes. Ao comparar os resultados obtidos em ambas, foi observado que a abordagem *wrapper,* em conjunto com a busca em sequência, produziram melhores resultados. Embora a diferença, no que diz respeito à taxa de acerto dos classificadores, tenha sido pequena, foi percebido maior uso dos recursos computacionais ao utilizar o *wrapper*.

De modo geral, o uso de algoritmos com a abordagem de filtro teve um processamento dentro da ordem de segundos a minutos, em contrapartida, o uso do *wrapper* elevou esta ordem para horas e dias de processamento, assim, tornando-se inviável a sua aplicação em alguns casos.

Sendo assim, foi percebido que a seleção de atributos é um método de redução de dimensionalidade que fornece bons resultados quando aplicados nas bases do presente estudo. Em paralelo, o método de projeção aleatória foi visto como alternativo, pois também está relacionado à redução do custo computacional, e o uso de ambos em conjunto proporciona bons resultados.

Por fim, os resultados demonstram que as aplicações desses métodos produzem uma taxa de acerto maior do que quando aplicado somente o algoritmo de mineração sobre as bases com todos os atributos.

4.2 ANÁLISE PREDITIVA SOBRE PACIENTES DO “PROJETO DE EXTENSÃO REABILITAÇÃO PULMONAR” DA UNIVERSIDADE FEEVALE (HECKLER, 2018)

O presente trabalho aborda o assunto de doenças respiratórias, principalmente no que se trata de doenças pulmonares. Este estudo propõe o desenvolvimento de uma ferramenta que auxilie no processo de *machine learning* dentro do "Projeto de Extensão Reabilitação Pulmonar", desenvolvido pela Universidade Feevale.

Um dos principais obstáculos percebidos e se destacam no presente estudo é a escassez de informações completas nas bases de dados, fruto do abandono do tratamento por parte de pacientes. Sendo assim, o foco principal deste trabalho é identificar as tendências de abandono

dos pacientes que estão ingressando neste tratamento e, assim, extrair informações sobre a base de dados a fim de contribuir no tratamento de reabilitação pulmonar.

A ferramenta em questão disponibiliza visualizações a fim de auxiliar na leitura e compreensão dos dados e resultados por parte de profissionais na área da saúde.

O "Projeto de Extensão Reabilitação Pulmonar" da Universidade Feevale é classificado como um Programa de Reabilitação Pulmonar (PRP) e compreende o atendimento da comunidade do Vale dos Sinos, de ambos os sexos e idade superior a 40 anos, visando o tratamento de pessoas com DRCs, DPOC, dentre outras doenças que atingem o sistema respiratório.

Comprovando a eficácia do uso de PRPs no tratamento dessas doenças, Nici et al. (2006) comentam que, no que se refere à reabilitação pulmonar, é uma intervenção baseada em evidências, multidiciplinar e abrangente em pacientes com doenças crônicas, que, muitas vezes, diminuem as atividades da vida diária. Desta forma, a reabilitação pulmonar é projetada para reduzir os sintomas, otimizar o estado funcional e reduzir os custos dos cuidados de saúde, promovendo a estabilização ou mesmo a reversão das manifestações da doença.

Com o intuito de auxiliar na coleta de dados, são realizadas entrevistas pré-tratamento e pós-tratamento por meio de formulários para registro de informações. As informações coletadas são, por exemplo, gênero, idade, altura, peso, doença respiratória e período do tratamento.

Além dos dados supracitados, também são coletados indicadores relacionados à saúde do paciente. Estas informações podem ser número de vezes que o paciente tossiu em um determinado período, a existência de secreção, pressão no peito e ocorrência de dispneia. Juntamente a estes dados, também são coletadas informações que tratam da qualidade de vida que o paciente leva, como o grau de limitação de atividades diárias comuns (caminhar, conversar, arrumar a cama, lavar a louça, entre outras). A qualidade do sono e o nível de disposição também são dados armazenados para esta pesquisa.

Tomando por base a revisão sistemática e o estudo aprofundado sobre as técnicas de *machine learning* vistas no decorrer do estudo de Heckler (2018), as técnicas selecionadas para fim de comparação de desempenho foram SVM (*Support Vector Machine*), DT (*Decision Tree*) e RF (*Random Forest*).

Por serem técnicas comumente identificadas na revisão sistemática, a SVM e DT foram inicialmente selecionadas, e mesmo RF também sendo identificado com frequência, esta foi utilizada pela sua grande similaridade com DT, o que facilita a aplicação. Como ponto foco

da análise, o abandono foi escolhido para a análise preditiva, visto que este é o maior problema do projeto.

Pelo fato de o projeto de extensão ter seus dados armazenados em planilhas eletrônicas, foi vista a necessidade de uma etapa para o ajuste dos dados, dentro da fase de pré-processamento, a fim de torná-los aptos para a análise. Este tipo de armazenamento é caracterizado por vários tipos de problemas, dentre eles: as diferentes formas de escrita de uma mesma palavra e valores não informados pela falta de obrigatoriedade de preenchimento. Também existem dados que não são coletados pelo fato de que o paciente pode não ter realizado um determinado exame, ou mesmo tenha abandonado o tratamento ainda na etapa de coleta de dados.

Ao decorrer da etapa de transformação dos dados, alguns dos atributos utilizados na análise são preenchidos com valores numéricos, a fim de sinalizar que se refere a um valor descritivo, por exemplo, quando o atributo relacionado a gênero há como valores possíveis "1" e "2" para indicar respectivamente na descrição "MASCULINO" e "FEMININO".

Subsequente à etapa de pré-processamento e de transformação dos dados, foi gerado, a partir de um modelo, um novo conjunto de dados ajustados que compara o desempenho entre as técnicas de *machine learning*, assim, gerando um modelo preditivo para cada técnica, considerando todos os atributos desse novo conjunto gerado. É importante salientar que, para que as comparações possam ser igualmente aplicadas, o mesmo conjunto foi aplicado para todas as técnicas.

De acordo com Alpaydin (2010), *Decision Tree* (DT) é uma estrutura de dados hierárquica que pode ser utilizada para classificação e regressão. De maneira geral, sistemas baseados em DT apresentam maior acurácia, além de representar de forma intuitiva os resultados, melhorando a compreensão por parte dos seres humanos. A vista disso, o uso de DT é muito popular na área de *machine learning.* (HAN; KAMBER; PEI, 2012).

Han et al. (2012) explicam que sistemas de DT geram um modelo preditivo com a estrutura de árvore semelhante a um fluxograma, no qual o nodo mais acima é denominado nodo raiz, e cada nodo interno representa a validação de um dos atributos das instâncias de dados. Deste modo, cada ramificação gerada pelo nodo interno representa um resultado do teste e, cada nodo folha, também chamado de nodo terminal, representa um rótulo de uma classe que será designada à instância analisada.

Liu (2011) complementa que mais de uma árvore pode ser gerada a partir de um mesmo conjunto de dados e, ressalta, também, que quanto menor for a árvore gerada, melhor será seu desempenho, visto que, dessa maneira, a árvore pode ser compreendida como um

modelo mais genérico. Outro fator que se mostra importante, quanto a árvores com tamanho diminuto, é que sua compreensão se torna mais fácil para humanos, e este último fator recebe importância porque, em alguns casos, o entendimento por parte do usuário é fundamental. Um exemplo a este caso de compreendimento, é o diagnóstico de uma determinada doença, em que os médicos necessitam saber quais os fatores que classificam se uma pessoa é portadora ou não.

Outra técnica de aprendizado de máquina, considerada no trabalho de Heckler (2018), é *Random Forest* (RF). Consiste em um conjunto de árvores de decisão que são geradas através da seleção aleatória de atributos em cada nodo, por esse motivo é denominada "floresta". É considerada uma técnica robusta e eficiente em bancos de dados com grande volume de dados. A fim de alcançar um bom desempenho no decorrer da utilização de RF, é extremamente indicado diminuir a correlação entre os classificadores individuais. (HAN; KAMBER; PEI, 2012).

Para Breiman (2001), RF é uma coleção de classificadores estruturados em árvore, onde cada classificador depende de um vetor aleatório, que é distribuído de forma equilibrada entre os classificadores.

O processo de classificação de RF é subdividido em três etapas. A primeira fase consiste na geração dos vetores aleatórios após o treinamento no conjunto de dados. O próximo passo refere-se à criação de árvores de decisão aleatórias, na qual cada árvore é criada a partir de um vetor aleatório, que, por sua vez, são gerados partindo de uma distribuição de probabilidade fixa. Como última etapa, a classificação da instância baseada em uma estrutura de votação por maioria, que atribui para a instância da classe que obtiver mais votos. (TAN; STEINBACH; KUMAR, 2009).

Considerado um dos algoritmos mais populares, no que diz respeito a *machine learning*, o *Support Vector Machine* (SVM) é marcado pela precisão de classificação, visto que ela é maior comparando com outros algoritmos em aplicações que envolvem grande volume de dados.

Tomando por base um sistema linear, o SVM constrói classificadores de duas classes (positiva e negativa). Partindo do vetor de entrada, o algoritmo identifica uma função linear para a construção de um modelo para classificação. O uso da função linear tem intuito de gerar um limite entre as classes, facilitando a classificação das instâncias. Desta forma, são atribuídas à classe positiva as instâncias que ficarem acima deste limite, já o que for apresentado como abaixo do limite será atribuído à classe negativa. (LIU, 2011).

Alpaydin (2010) complementa que o vetor de pesos é um parâmetro da função linear que é utilizado para formar o subconjunto de vetores de suporte. Estas são as instâncias mais

próximas do limite que separam as duas classes. Sendo assim, são denominadas casos incertos que, por estarem próximas ao limite, permitem a extração do conhecimento.

O trabalho de Heckler (2018), no "Projeto de Extensão Reabilitação Pulmonar", identificou oportunidades de utilização de *machine learning* nesse ambiente, bem como, alguns benefícios que o uso dessa tecnologia pode gerar. Com o intuito de prever a tendência de abandono dos pacientes, foram utilizadas técnicas de análise preditiva como SVM, DT e RF.

Percebeu-se que o modelo preditivo que utilizou a aplicação da técnica de RF, obteve melhor resultado, comparado às demais técnicas analisadas. Com isso, foi observado que o uso de análise preditiva pode auxiliar na previsão da tendência de abandono dos pacientes. Partindo dessa visão, é possível traçar estratégias para reduzir o número de abandono do tratamento.

A ferramenta desenvolvida no estudo torna possível a especialistas da área fazer a devida análise no que se trata das tendências de abandono, mesmo sem conhecimento na área de informática. Para facilitar a interpretação dos dados, a ferramenta conta com um painel interativo que apresenta, de forma amigável, o resultado da análise.

Visto que o armazenamento dos dados foi obtido em uma planilha, este fator também ocasionou alguns problemas na coleta dos dados. Um dos principais identificados foi a falta de padronização de doenças, tornando necessários ajustes antes de iniciar a análise dos dados. Sobre este acontecimento, Heckler (2018) comenta que um dos impactos relacionados ao abandono do tratamento, por parte dos pacientes, prejudica a coleta de dados e atributos e, por consequência, o desempenho dos modelos acaba sendo prejudicado, sendo que as técnicas SVM e RF exigem um conjunto de dados em que não existam valores ausentes. No trabalho do autor, foram adotadas estratégias para preenchimento destes valores ausentes.

A grande ausência de valores ocasionou criação de modelos preditivos desconsiderando esse atributo e, em contrapartida, se mostraram menos eficientes. Por fim, foi percebido que a utilização de múltiplos atributos tornou inviável a visualização do modelo graficamente, pois não é possível gerar visualizações com mais de três dimensões. Juntamente a este fator, em contraposto ao que ocorreu com o modelo criado com a técnica RF, não foi identificada nenhuma técnica para visualização dos atributos mais importantes para o modelo.

4.3 INCADATABR: UMA BIBLIOTECA EM R PARA MANIPULAÇÃO DE DATASETS DO INCA (COSTA, 2019)

Para o presente estudo, são utilizados dados públicos do Instituto Nacional de Câncer José Alencar da Silva (INCA). Neste trabalho, é destacada a dificuldade em ter acesso aos dados públicos, visto que há etapas no processo que causam esta dificuldade.

À vista disso, no decorrer do trabalho, foi desenvolvida uma biblioteca com o objetivo de facilitar a interação dos profissionais da área da saúde e TI, que tenham conhecimento básico em programação. A biblioteca desenvolvida facilita diversos aspectos no que se trata da obtenção dos dados, como, por exemplo, importação, exportação e interação dos dados, auxiliando, assim, na análise dos dados.

O trabalho de Costa (2019) aborda a dificuldade por parte dos pesquisadores no que se refere a análise e processamento de dados do Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), onde são destacadas dificuldades desde o processo de *download* até a necessidade de se utilizar um software que seja exclusivo para o sistema operacional Windows.

Devido aos empecilhos apresentados, o estudo propôs o desenvolvimento de uma biblioteca, a fim de facilitar o percurso de importação, exportação e interação para, assim, auxiliar na análise dos dados. A biblioteca também conta com a geração de gráficos de forma dinâmica, de maneira a facilitar a interpretação.

Como é explicado no estudo, a redução no custo de armazenamento colaborou para que o armazenamento de dados se tornasse um dos principais objetivos das instituições e, aliado a isso, a utilização de ferramentas que automatizam a coleta de dados tornou possível às instituições o acúmulo de dados de diferentes tipos, dentre eles, arquivos de texto, planilhas, bancos de dados, dentre outros.

Foi percebido pelas organizações, que a velocidade de coleta de dados é superior à análise, bem como, no que se trata do processamento destes dados coletados. Este atraso na análise gera um problema que ao mesmo tempo torna-se uma contradição. Por observar que há grande quantidade de dados, pode se concluir de forma errônea que a instituição está bem informada. Porém, esta informação não coaduna, visto que o atraso na análise dos dados acarreta que a informação e o conhecimento existente já não se adéquem à realidade da instituição e, em um cenário ainda mais desvantajoso, não há descoberta de informação útil nos dados coletados.

Mesmo que a quantidade de dados existente para análise seja grande, pesquisadores encontram dificuldades para dar seguimento em projetos, cujo foco seja realizar a análise de dados. Dentre os problemas identificados, é possível salientar a complexidade envolvida com o processo de obtenção de dados que, em boa parte dos casos, está associado à burocracia, má estruturação, inexistência de estruturas de armazenamento, ou mesmo à omissão de dados por parte da instituição.

Outra barreira encontrada para a pesquisa de dados relaciona-se com a dificuldade de encontrar ferramentas que auxiliem nesta etapa, além do fato de ser necessário que o

pesquisador tenha ao menos conhecimento intermediário em informática. Esse conhecimento tem-se requerido em vista da necessidade de realizar ajustes nos dados que serão utilizados na pesquisa, para fins de remover ou corrigir dados incorretos ou duplicados e, desta forma, evitar que inconsistências ocorram e prejudiquem o resultado da análise.

No presente trabalho, a biblioteca apresentada foi pensada com base no problema descrito por Pujari (2001), que é considerado o desenvolvimento de ferramentas que auxiliem no processo de aquisição de conhecimento sobre os dados fornecidos pelo INCA. Após a análise destes dados, é possível contribuir para o planejamento de ações no que se trata de educação e detecção de doenças. (COSTA, 2019, p. 24).

O objetivo da construção deste pacote é permitir que pesquisadores possam realizar análises utilizando a linguagem R, ao mesmo tempo que evitem as etapas de pré-processamento necessárias, como acontece, por exemplo, com outras ferramentas. Para facilitar o uso por parte do pesquisador, a biblioteca criada permite que usuários com conhecimentos básicos na linguagem R possam utilizá-lo, deixando a análise mais simples. Assim, o pacote nomeado INCADATABR torna-se uma nova alternativa para análises.

Para o desenvolvimento do pacote mencionado neste estudo, foi utilizada a linguagem R em virtude da sua relevância no que se trata da área de análise de dados, bem como, o crescente número de usuários.

Costa (2019) comenta que o pacote desenvolvido é baseado no modelo de arquitetura monolítica, em que a arquitetura denota que a aplicação seja projetada sem modularidade externa, de forma que a construção da aplicação seja um módulo a ser utilizado por outra aplicação.

No decorrer da execução do trabalho de Costa (2019), foram feitas pesquisas a fim de obter uma melhor forma para o desenvolvimento e estruturação da biblioteca. Para o desenvolvimento desta ferramenta, foi utilizado a linguagem R.

Com a utilização desta biblioteca, é possível ao usuário realizar a geração de gráficos de forma dinâmica, utilizando a IDE. O pacote está à disposição para profissionais da saúde, sendo, desta forma, uma ferramenta extra para complementar as já existentes. Foram realizadas visualizações para muitas variáveis existentes em *datasets* disponibilizados pelo INCA, sem que o usuário precise fazer ajustes.

O pacote desenvolvido tem potencial de tornar-se uma ferramenta de extrema relevância para profissionais da área da saúde, principalmente pela sua simplicidade para utilização e pelo bom desempenho, no que se trata da geração de gráficos.

5 DESENVOLVIMENTO

Neste capítulo, será abordado todo o processo de desenvolvimento do *dashboard*, visando a criação de uma ferramenta que proporcione, à profissionais da área da saúde, uma fácil análise referente a dados de COVID-19.

Para tornar possível a construção desta ferramenta, será utilizada a linguagem de programação R juntamente com a interface de desenvolvimento RStudio, e, desta forma, apresentar os dados para o usuário utilizando gráficos e resumos de valores para agilizar a identificação de possíveis grupos de pacientes, assim como ter uma rápida percepção no que se trata de números de casos de uma maneira geral.

A seguir, serão detalhados os passos que foram galgados desde a seleção e limpeza dos dados até o *dashboard* em funcionamento e disponível para consulta pública.

5.1 SELEÇÃO DAS BASES DE DADOS

Conforme descrito em site oficial da FAPESP (2020), as bases de dados utilizadas são fornecidas pelo repositório COVID-19 Data Sharing/BR, uma iniciativa da FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo, que, inicialmente, conta com a participação do Instituto Fleury, Hospital Sírio Libanês e Hospital Israelita Albert Einstein. O objetivo principal é trazer contribuições para pesquisas relacionadas nesta temática. Os dados disponibilizados pelo repositório FAPESP COVID-19 DataSharing/BR estão disponíveis em: <*https://repositoriodatasharingfapesp.uspdigital.usp.br/*>.

Existem duas categorias de informações dentre os dados disponibilizados: dados demográficos e dados de exames clínicos e/ou laboratoriais, além de movimentações do paciente, quando houver, como internações e desfecho dos casos, como recuperação e óbito.

No que se trata de disponibilidade de dados e regras de citação, os dados são abertos ao público, sem qualquer custo para baixar. O repositório de dados da FAPESP é periodicamente atualizado, e os dados fornecidos pelas instituições de saúde já são pseudonimizados, de maneira que as instituições não se responsabilizam pelo uso indevido dos dados.

Todos os dados disponibilizados pelo repositório da FAPESP utilizam a licença CC-BY, que indica que toda e qualquer publicação ou apresentação que utilizar total ou parcialmente os dados deste repositório, deve citar a página web do FAPESP COVID-19 Data Sharing/BR.

Referente à licença CC-BY, Grossi (2017) explica que *Creative Commons* é um tipo de licença de atribuição, em que os autores especificam de que maneira seu trabalho poderá ser utilizado. Conteúdo do tipo CC, em sua maioria, não exige que seja solicitada permissão antecipada, ou seja efetuado pagamento de qualquer tipo ao autor.

Existem níveis de licença CC reconhecidas internacionalmente, projetadas para estarem em conformidade com os direitos autorais em âmbito internacional. Dentre as licenças existentes, a CC BY é a mais permissiva, pois torna possível que o conteúdo seja distribuído, alterado e reinventado, sendo a única obrigatoriedade que a fonte original seja citada. (GROSSI, 2017).

No que relaciona às bases disponibilizadas no repositório da FAPESP (2020), este conta com a participação do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, o Instituto Fleury, Hospital Sírio Libanês e o Hospital Israelita Albert Einstein. Os arquivos fornecidos pelas instituições seguem um modelo de tabelas e colunas semelhantes, proporcionando integração entre as bases. Todas as instituições supracitadas fornecem um arquivo no formato .xlsx, intitulado Dicionário de Dados, e dois arquivos no formato .csv, que representam duas tabelas, Pacientes e Exames. Além destes arquivos, o Hospital Sírio-Libanês fornece um arquivo extra, em formato .csv, que representa a tabela de Desfecho, na qual constam resultados referentes à situação final de cada paciente, como alta ou óbito. (FAPESP, 2020).

Para fim de armazenamento dos dados referente a Paciente, Exames e Desfecho, são apresentados os Quadros 2, 3 e 4, com o arquivo Dicionário de Dados e seus campos.

Quadro 2 – Dicionário de dados do Paciente

Variáveis	Descrição	Formato	Conteúdo
ID_PACIENTE	Identificação única	Caracteres alfanuméricos	Conjunto de letras e números que é usado como chave identificadora do paciente.
IC_SEXO	Sexo	1 caractere alfanumérico	F - Feminino; M - Masculino

Variáveis	Descrição	Formato	Conteúdo
AA_NASCIMENTO	Ano de nascimento	4 números	Exemplo: 1959 AAAA, quando anterior a 1930.
CD_PAIS	País de residência	Alfanumérico	Exemplo: BR
CD_UF	Unidade federativa de residência	2 caracteres alfanuméricos	Sigla do estado ou unidade federativa.
CD_MUNICIPIO	Município de residência	Alfanumérico	Exemplo: São Paulo MMMM: Quando houver poucas ocorrências
CD_CEP	CEP de residência	5 números	5 primeiros dígitos do CEP (**) CCCC quando houver poucas ocorrências

Fonte: Adaptado de FAPESP (2020)

Quadro 3 – Dicionário de dados de Resultados dos Exames

Nome Variável	Descrição	Formato	Conteúdo	Observação
ID_PACIENTE	Identificação única do paciente que correlaciona com ID_PACIENTE da tabela de PACIENTES	Caracteres alfanuméricos	Conjunto de letras e números que é usado como chave identificadora do paciente.	

Nome Variável	Descrição	Formato	Conteúdo	Observação
DT_COLETA	Data que o material foi coletado do paciente	Data (aaaa/MM/dd)	Data no formato ano/mês/dia	
DE_ORIGEM	Origem do paciente	4 caracteres alfanuméricos	Fixo em HOSP, que representa exame realizado dentro de Unidade Hospitalar	
DE_EXAME	Descrição do exame realizado	Alfanumérico	*String*, Exemplo: (HEMOGRAMA, SÓDIO, POTÁSSIO)	Um exame é composto por 1 ou mais analitos.
DE_ANALITO	Descrição do analito	Alfanumérico	*String*, Exemplo: (Eritrócitos, Leucócitos, Glicose, Ureia, Creatinina)	Para Hemograma, tem-se o resultado de vários analitos: Eritrócitos, Hemoglobina, Leucócitos, Linfócitos, etc. A maioria dos exames tem somente 1 analito,

Nome Variável	Descrição	Formato	Conteúdo	Observação
				como exemplo a Glicose, Colesterol Total, Ureia e Creatinina.
DE_RESULTADO	Resultado do exame, está associado com DE_ANALITO	Alfanumérico	Se DE_ANALITO exigir valor numérico, então Inteiro ou Decimal. Se DE_ANALITO exigir qualitativo, *String* com domínio restrito.	Exemplo de domínio restrito: Positivo, Detectado, Reagente, não reagente, etc.
CD_UNIDADE	Unidade de Medida utilizada na Metodologia do Laboratório para analisar o exame	Alfanumérico	*String*, exemplo: g/dL (gramas por decilitro)	
DE_VALOR_REFERENCIA	Valores de referência para	Alfanumérico	*String*: Faixa de resultados em que é considerado	

Nome Variável	Descrição	Formato	Conteúdo	Observação
	DE_RESULTADO		normal para este analito, na população. 'Valor Mínimo' a 'Valor Máximo'; (Não Detectado/Detectado); Exemplo: Glicose: 75 a 99 Progesterona: Até 89	

Fonte: Adaptado de FAPESP (2020)

Quadro 4 – Tabela de dados de Desfecho

Nome Variável	Descrição	Formato	Conteúdo
ID_PACIENTE	Identificação único do paciente que correlaciona com ID_PACIENTE das tabelas de PACIENTES e EXAMES.	32 caracteres alfanuméricos	*String* anonimizado
ID_ATENDIMENTO	Identificação única do atendimento de	Alfanumérico	*String* anonimizado

Nome Variável	Descrição	Formato	Conteúdo
	forma anonimizada que se relaciona com os atendimentos na tabela de resultados de exames. Cada atendimento gera um desfecho a não ser que o paciente seja internado.		
DT_ATENDIMENTO	Data da realização do atendimento.	Data (DD/MM/AAAA)	DD = Dia MM = Mês AAAA = Ano Exemplo: 24/06/2020
DE_TIPO_ATENDIMENTO	Descrição do tipo de atendimento realizado.	Texto livre	*String*, exemplo: Pronto Atendimento
ID_CLINICA	Identificação da clínica onde o evento aconteceu.	Numérico	Exemplo: 1013
DE_CLINICA	Descrição da clínica onde o evento aconteceu.	Texto livre	Exemplo: Retorno Digital Adulto
DT_DESFECHO	Data do desfecho.	Data (DD/MM/AAAA)	DD = Dia MM = Mês AAAA = Ano

Nome Variável	Descrição	Formato	Conteúdo
			Exemplo: 24/06/2020
DE_DESFECHO	Descrição do desfecho.	Texto livre	Exemplo: Alta médica melhorado

Fonte: Adaptado de FAPESP (2020)

É importante ressaltar que, na página inicial do site da FAPESP (2020), é informado que a base do Hospital das Clínicas da Universidade de São Paulo foi publicada em 17 de fevereiro de 2021, e as demais bases tiveram sua publicação em 30 de junho de 2020.

5.2 LIMPEZA E PRÉ-PROCESSAMENTO

A etapa de limpeza e pré-processamento das bases, foi iniciado com a base de dados do Hospital Sírio Libanês (HSL). Nesta base, o primeiro arquivo analisado foi o correspondente à tabela de pacientes (hsl_patient_1.csv), que conta com 2731 registros de pacientes que fizeram um ou mais exames laboratoriais.

Em alguns preenchimentos do campo AA_NASCIMENTO, que corresponde ao ano de nascimento do paciente, foi encontrado o valor (AAAA), e, seguindo o descrito no Dicionário de Dados do HSL, isso se refere aos dados de ano de nascimento iguais ou anteriores a 1930. Para fins de normalização e leitura por parte do software, todos os locais com ocorrência do valor (AAAA) foram reescritos com o valor (1930).

Dentre os países presentes nesta base de dados, o Afeganistão apresentou uma quebra de formatação na representação da letra (ã). Desta forma, o conjunto de caracteres que estavam em seu lugar foi representado na base de dados original por (ĂŁ). Assim, todos os locais onde apresentavam estes caracteres foram substituídas por (a), sem a acentuação, justamente para prevenir possíveis conflitos de caracteres.

Com estas alterações, o arquivo hsl_patient_1.csv, que representa a tabela de Pacientes na base de dados do HSL, pôde ser importado para a ferramenta R Studio sem impedimentos.

A próxima tabela do HSL a ser limpa e processada é referente aos exames dos pacientes, tendo seus dados salvos no arquivo (hsl_lab_result_1.csv). Esta tabela possui 371357 registros e, devido a este grande volume de dados, tornou-se necessário criar um quadro dos passos a serem seguidos para fazer a limpeza de seus dados. A seguir, foi criada uma tabela de

correspondência (Quadro 5) em ordem de execução das substituições a serem feitas. Ao finalizar estas alterações, o arquivo pode ser importado no *software* sem erros.

Quadro 5 – Limpeza de dados da tabela de Exames do HSL

Ordem de execução	Valor Inicial	Próximo Valor
1	,	$
2	|	,
3	“	‘
4	$	,
5	ĂlՋ	o
6	Ălł	o
7	ĂĄ	a
8	ĂŠ	e
9	Ă§oo	cao
10	Ă-	i
11	Ă§o	co
12	ĂŞ	e
13	Ăcido	Acido
14	Ărico	Urico
15	Ă´	o
16	NĂO	NAO
17	Ăndice	Indice
18	TĂVEL	TAVEL

Ordem de execução	Valor Inicial	Próximo Valor
19	noo	nao
20	Ă§a	ca
21	soo	sao
22	Ă˜	a
23	Ăc	Ac

Fonte: Elaborado pelo autor

A última tabela da base de dados do HSL a ser limpa foi a de Desfecho. Esta tabela conta com 9633 registros que se relacionam com a tabela de Pacientes, utilizando como chave de referência entre as tabelas o campo ID_PACIENTE. A seguir, foi criada uma tabela de correspondência em ordem de execução das substituições a serem feitas (Quadro 6). Depois destas alterações, o último arquivo da base de dados do HSL pode ser importado no *software* sem erros.

Quadro 6 – Limpeza de dados da tabela de Desfecho do HSL

Ordem de execução	Valor inicial	Valor final
1	“	‘
2	|	,
3	lĂ-	li
4	ĂŠ	e
5	SĂ-	Si
6	ĂŞ	e
7	sĂŁo	sao
8	iĂĄt	iat

Ordem de execução	Valor inicial	Valor final
9	apĂls	apos
10	Ăbito	Obito
11	aĂ§ĂŁo	acao
12	rĂlp	rop
13	aĂ§Ăl’es	acoes
14	lĂ˜n	lan
15	Ă§	c
16	tĂlr	tor
17	ĂĄ	a

Fonte: Elaborado pelo autor

Com todas as tabelas do HSL já limpas e processadas, o conjunto de dados que foi pré-processado na sequência corresponde às tabelas do Hospital Albert Einstein (HAE). Diferente do HSL, esta base não contém uma tabela de desfecho, dessa forma, foi analisada primeiro a tabela de pacientes, que contêm 43562 registros e, subsequentemente, foi analisada a tabela de exames, com 1853695 registros. Para a tabela de pacientes não foi necessário fazer alterações nos dados, pois todos já estavam em um formato adequado para a importação desta tabela no *software*. Já para a tabela de exames, foi necessário apenas remover os caracteres de acentuação, substituindo cada um pela respectiva letra sem a acentuação, conforme exemplificado no Quadro 7.

Quadro 7 – Limpeza de dados da tabela de Exames do HAE

Valor Original	Novo Valor
ã	a
á	a
é	e

Valor Original	Novo Valor
ê	e
í	i
ó	o
õ	o
ô	o
ú	u
ç	c

Fonte: Elaborado pelo autor

Com as tabelas do HAE finalizadas, as próximas tabelas a serem limpas foram do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HSP). Iniciando pela tabela de pacientes, assim como para a tabela de pacientes do HAE, não houve mudanças necessárias, esta tabela conta com 3750 registros. Desta forma, a única tabela que foi modificada para o HSP foi referente aos exames, com 2498650 registros encontrados, seguindo as alterações demonstradas no Quadro 8.

Quadro 8 – Limpeza de dados da tabela de Exames do HSP

Valor Inicial	Novo Valor
á	a
à	a
ã	a
â	a
é	e
ê	e

í	**i**
ó	**o**
õ	**o**
ô	**o**
ö	**o**
ú	**u**
ç	**c**

Fonte: Elaborado pelo autor

Com estas alterações, o último conjunto de dados a ser limpo é referente ao Grupo Fleury (GF), que, da mesma maneira que o HAE e HSP, não foram necessárias alterações para a tabela de pacientes, que contém com 129595 registros armazenados. Para a tabela de exames, com 2496591 registros, as alterações foram na acentuação, conforme apresenta o Quadro 9.

Quadro 9 – Limpeza de dados da tabela de Exames do GF

Valor Inicial	Novo Valor
á	**a**
à	**a**
ã	**a**
â	**a**
é	**e**
ê	**e**
í	**i**
ó	**o**
ô	**o**

Valor Inicial	Novo Valor
õ	**o**
ú	**u**
ç	**c**

Fonte: Elaborado pelo autor

Após este conjunto de alterações, todas as tabelas já estão devidamente preparadas para serem importadas pelo *software* RStudio.

5.3 LINGUAGEM R E RSTUDIO

Visto que a linguagem de programação R pode ser utilizada para a representação de gráficos sobre dados coletados, esta foi selecionada para ser a linguagem principal desta ferramenta.

De acordo com Fellows (2012), a linguagem de programação R, mesmo possuindo os conceitos e paradigmas de linguagens de programação tradicionais, destaca-se por ser uma poderosa ferramenta estatística que, por meio de inúmeros pacotes e bibliotecas disponíveis, faz uso de técnicas computacionais com foco em solucionar problemas que apresentam grande nível de complexidade.

Para efetuar o *download* e instalação da linguagem R, foi acessado o site do fabricante, que direciona o usuário para uma página de *downloads*, no qual é possível selecionar o instalador da linguagem que melhor se encaixa ao sistema operacional do usuário.

Após baixar o arquivo de instalação, foi feito duplo clique sobre o arquivo, e seguiu-se os passos de instalação sugeridos pelo fabricante.

Como etapa subsequente, é feito o *download* e instalação da ferramenta de interface de desenvolvimento RStudio, selecionada para facilitar o manuseio e utilização da linguagem R, de forma que o projeto poderia ser desenvolvido sem a sua utilização.

O RStudio pode ser baixado do site do fabricante, no qual são oferecidas as opções de *download* para diversos sistemas operacionais. Tendo selecionada a melhor opção para o sistema operacional e baixado o arquivo, ao executar um duplo clique sobre este, a instalação é iniciada, e deve ser seguido o processo de instalação sugerido pelo fabricante.

5.4 DESENVOLVIMENTO DA APLICAÇÃO

O desenvolvimento do *dashboard* foi concentrado na criação de uma ferramenta que tenha um lado cliente e outro servidor, de maneira que o servidor acessa um arquivo de dados em que é feito o direcionamento para as bases selecionadas, além de mesclar tabelas e organizar dados quantitativos e qualitativos.

Será mostrado a seguir, como foi a criação de cada um destes arquivos, bem como, a comunicação entre eles, para que, desta forma, seja possível a replicação deste *dashboard* por qualquer desenvolvedor.

Dentre os passos abordados a seguir, estão o desenvolvimento da interface de usuário (UI), a área de servidor e um arquivo de dados responsável pela comunicação e manuseio das bases.

5.4.1 Interface do usuário (ui)

Na construção da UI, foram utilizadas as bibliotecas *shinydashboard* e *ggplot2*. Para que estas bibliotecas sejam utilizadas pelo projeto elas foram incluídas no início do código.

Subsequente ao uso destas bibliotecas, é informado onde está localizado o arquivo de dados, como é ilustrado na Figura 5.

Figura 5 – Ilustração de importação de bibliotecas e chamada de arquivo de dados

```
1  library(shinydashboard)
2  library(ggplot2)
3  source("data.R")
4
5  ui <- dashboardPage(
6
7      dashboardHeader(
8          title = "Corona Viewer"
9      ),
10
11     dashboardSidebar(
12         sidebarMenu(
```

Fonte: Elaborada pelo autor

A etapa seguinte consiste em iniciar a criação de uma página utilizando *shinydashboard*, onde é nomeada uma variável como ui, e esta recebe um componente para a criação da página, neste componente são informados três subcomponentes:

- *DashboardHeader*: Receberá o título da aplicação, que neste caso foi chamada de Corona Viewer.
- *DashboardSidebar*: Que controla o menu lateral esquerdo, onde é possível selecionar o tipo de dado que o usuário deseja consultar.

- *DashboardBody*: Onde são construídos os demais componentes visuais da tela, como as caixas coloridas que informam dados quantitativos, os componentes seletores para que o usuário possa filtrar e refinar sua busca, os gráficos interativos e os dados presentes no menu sobre.

Na Figura 5 é possível identificar que entre as linhas 7 e 9, é feita a construção do componente de cabeçalho da aplicação, da mesma forma que na Figura 6, é utilizado o trecho de código entre as linhas 11 e 48 para fazer a criação do menu lateral, que conta com sete opções.

Figura 6 – Construção de menu lateral

```
11      dashboardSidebar(
12          sidebarMenu(
13              menuItem(
14                  "Visão Geral",
15                  tabName = "visaoGeral",
16                  icon = icon("chart-bar")
17              ),
18              menuItem(
19                  "Pacientes com Exames",
20                  tabName = "pacientesComExames",
21                  icon = icon("chart-bar")
22              ),
23              menuItem(
24                  "Resultados Negativos",
25                  tabName = "positivosNegativos",
26                  icon = icon("chart-bar")
27              ),
28              menuItem(
29                  "Resultados Positivos",
30                  tabName = "positivosNegativosPOSITIVOS",
31                  icon = icon("chart-bar")
32              ),
33              menuItem(
34                  "Óbitos",
35                  tabName = "obitosRecuperacoes",
36                  icon = icon("chart-bar")
37              ),
38              menuItem(
39                  "Recuperações",
40                  tabName = "obitosRecuperacoesRECUPERACOES",
41                  icon = icon("chart-bar")
42              ),
43              menuItem(
44                  "Sobre",
45                  tabName = "sobre",
46                  icon = icon("chart-bar")
47              )
48          )
```

Fonte: Elaborada pelo autor

O próximo e último componente a ser criado para a interface de usuário, foi o componente do corpo da aplicação. Este componente recebeu todos os subcomponentes referentes à visualização dos gráficos e dos mostradores de dados quantitativos.

Nas Figuras 7 e 8, são ilustradas, na ordem, a construção das caixas de informações quantitativas, da linha 58 até a linha 73, e a construção do primeiro conjunto de quatro gráficos que mostram informações gerais para ano de nascimento, sexo, cidade e estado.

Figura 7 – Construção de caixas coloridas com informações quantitativas

```
dashboardBody(
    tabItems(
        tabItem(
            # ------------------------------------------------------------
            # -- VISAO GERAL -----------------------------------------------
            # ------------------------------------------------------------
            tabName = "visaoGeral",
            fluidRow(
                valueBoxOutput(width = 2, outputId = "pacientes_analisados"),
                valueBoxOutput(width = 2, outputId = ""),
                valueBoxOutput(width = 2, outputId = ""),
                valueBoxOutput(width = 2, outputId = "exames_coletados"),
                valueBoxOutput(width = 2, outputId = "quantidade_negativos"),
                valueBoxOutput(width = 2, outputId = "quantidade_positivos")
            ),
            fluidRow(
                valueBoxOutput(width = 2, outputId = "quantidade_homens"),
                valueBoxOutput(width = 2, outputId = "quantidade_mulheres"),
                valueBoxOutput(width = 2, outputId = ""),
                valueBoxOutput(width = 2, outputId = "desistencia_tratamento"),
                valueBoxOutput(width = 2, outputId = "quantidade_altas"),
                valueBoxOutput(width = 2, outputId = "quantidade_falecidos")
            ),
```

Fonte: Elaborada pelo autor

Figura 8 – Área dos gráficos da opção inicial, Visão Geral

```
            ),
            fluidRow(
                box(
                    title = "Pacientes por faixa de Ano de Nascimento",
                    plotOutput( outputId = "visaoGeralAnoNascimento" )
                ),
                box(
                    title = "Pacientes por Sexo",
                    plotOutput( outputId = "visaoGeralSexo" )
                ),
                box(
                    title = "Pacientes por Estado",
                    width = 12,
                    plotOutput( outputId = "visaoGeralPacientesPorEstado" )
                ),
                box(
                    title = "Pacientes por Cidade",
                    width = 12,
                    plotOutput( outputId = "visaoGeralPacientesPorCidade" )
                )
            )
        )
```

Fonte: Elaborada pelo autor

Foi utilizado o subcomponente *fluidRow* para receber os 4 gráficos, colocados dentro de um subcomponente box, que recebe como parâmetros o título e a chamada *plotOutput* para que o gráfico seja exibido.

É importante salientar que, ao utilizar o *shiny*, o comprimento da linha padrão pode receber o valor máximo 12, sendo que, quando não são informados valores, por padrão, o *shiny* interpreta como 6. Sabendo disso, as duas primeiras saídas de gráfico serão exibidas na mesma linha, pois não possuem o parâmetro *width*, diferente das duas últimas chamadas de gráficos, em que cada um recebe explicitamente o valor 12. Assim cada um deles é exibido em uma linha diferente.

Outra área da construção da UI importante de ser apresentada é a dos trechos que utilizam os filtros para refinar os resultados. Com exceção das opções do menu “Visão Geral”

e “Sobre”, todas as outras opções do menu utilizam filtros que refinam a busca pelos dados de ano de nascimento, sexo, estado e cidade. No que se relaciona ao desenvolvimento de cada um desses filtros, é importante destacar:

- Ano de Nascimento: Para facilitar a busca por parte do usuário, bem como contribuir para uma leitura mais clara, foi utilizado para a exibição do ano de nascimento, o componente *sliderInput*. Este componente recebe como atributos, além do identificador (ID), o menor e maior ano de nascimento detectado e um intervalo inicial e final que deve ser definido para que o *sliderInput* coloque os primeiros pontos de intervalo do filtro. Para este caso, foi definido como intervalo inicial 1950 e final 2000. Assim, logo que o usuário acessa a tela, é possível detectar um grupo inicial de pacientes e utilizar o cursor do *mouse* para alterar o intervalo de ano de nascimento conforme o interesse da sua pesquisa.
- Sexo: Para filtrar os dados por sexo, foi feito o uso de *checkbox*, onde inicialmente vêm selecionados os dois sexos, masculino e feminino, e o usuário pode desmarcar um ou outro para fins de sua pesquisa. Para prevenir erros por parte do usuário, se nenhum dos dois sexos estiverem marcados, serão exibidos todos os sexos.
- Estado: No caso dos estados brasileiros, foi utilizado o *selectInput*, onde o usuário pode selecionar uma dentre todas as opções de estados brasileiros presentes na base de dados. É importante salientar que, como existem registros de pacientes onde não foi identificado o estado, estes receberam por definição das entidades que fornecem os dados, o valor (UU).
- Cidade: Semelhante ao que foi feito para estados, o campo de cidade também faz uso de um *selectInput* que, da mesma forma, busca as cidades dentre as disponíveis na base de dados de pacientes.

Com estes quatro filtros, é possível focar a busca por características em grupos específicos de pacientes e, após a seleção dos devidos filtros, os quatro gráficos presentes em tela exibirão informações que podem ser interpretadas por profissionais da área da saúde.

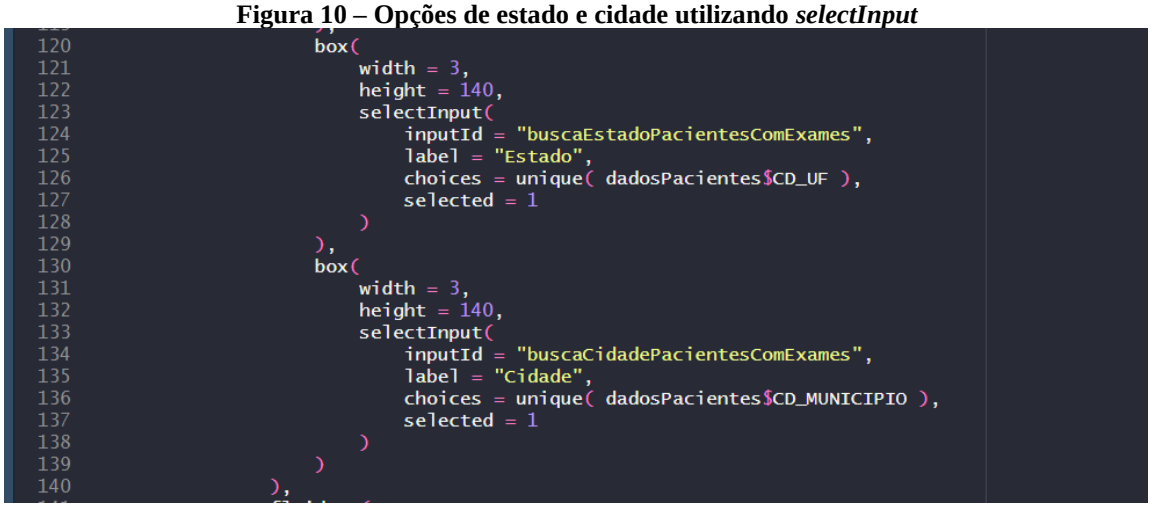
Sobre a construção base destes, quatro componentes de seleção, as Figuras 9 e 10, ilustram como foi feita a criação de cada um, ano de nascimento e sexo, e em seguida estado e cidade.

Figura 9 – Seletor de intervalo de ano de nascimento e *checkbox* para gênero

```
95            tabItem(
96                # ------------------------------------------------------------
97                # -- PACIENTES COM EXAMES ----------------------------------
98                # ------------------------------------------------------------
99                tabName = "pacientesComExames",
100               fluidRow(
101                   box(
102                       width = 3,
103                       height = 140,
104                       sliderInput(inputId = "rangeAnoNascimentoPacientesComExames", strong("Ano Nascimento"),
105                                   min = 1930, max = 2020,
106                                   value = c( 1950, 2000 )
107                       )
108                   ),
109                   box(
110                       width = 3,
111                       height = 140,
112                       title = strong("Sexo"),
113                       checkboxInput(inputId = "sexoMasculino",
114                                     strong("M - Masculino"),
115                                     value = TRUE),
116                       checkboxInput(inputId = "sexoFeminino",
117                                     strong("F - Feminino"),
118                                     value = TRUE)
119                   ),
```

Fonte: Elaborada pelo autor

Figura 10 – Opções de estado e cidade utilizando *selectInput*



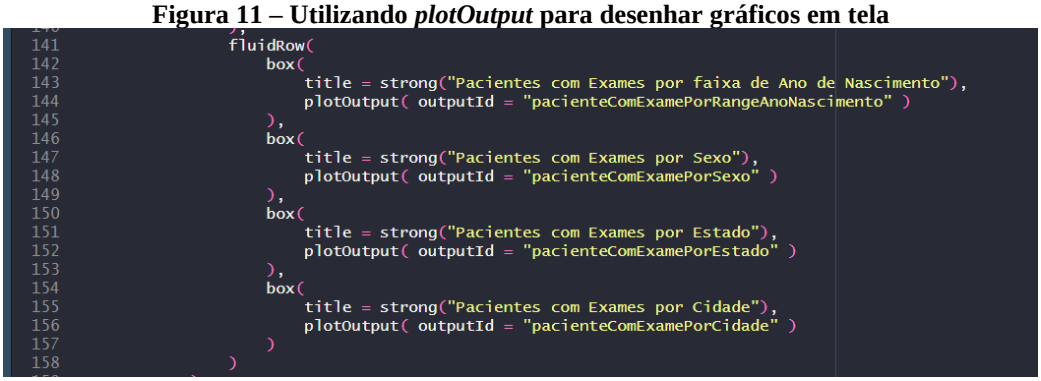
```
120                   box(
121                       width = 3,
122                       height = 140,
123                       selectInput(
124                           inputId = "buscaEstadoPacientesComExames",
125                           label = "Estado",
126                           choices = unique( dadosPacientes$CD_UF ),
127                           selected = 1
128                       )
129                   ),
130                   box(
131                       width = 3,
132                       height = 140,
133                       selectInput(
134                           inputId = "buscaCidadePacientesComExames",
135                           label = "Cidade",
136                           choices = unique( dadosPacientes$CD_MUNICIPIO ),
137                           selected = 1
138                       )
139                   )
140               ),
```

Fonte: Elaborada pelo autor

Após a construção dos filtros, o próximo passo foi exibir os gráficos resultantes dos atributos que foram selecionados pelo usuário. Para tornar a exibição possível, foi feita a chamada do *plotOutput*, responsável por "desenhar" o gráfico em tela.

Na Figura 11, é possível identificar, entre as linhas 141 e 158, o uso do *plotOutput* para finalizar esta etapa.

Figura 11 – Utilizando *plotOutput* para desenhar gráficos em tela



```
141               fluidRow(
142                   box(
143                       title = strong("Pacientes com Exames por faixa de Ano de Nascimento"),
144                       plotOutput( outputId = "pacienteComExamePorRangeAnoNascimento" )
145                   ),
146                   box(
147                       title = strong("Pacientes com Exames por Sexo"),
148                       plotOutput( outputId = "pacienteComExamePorSexo" )
149                   ),
150                   box(
151                       title = strong("Pacientes com Exames por Estado"),
152                       plotOutput( outputId = "pacienteComExamePorEstado" )
153                   ),
154                   box(
155                       title = strong("Pacientes com Exames por Cidade"),
156                       plotOutput( outputId = "pacienteComExamePorCidade" )
157                   )
158               )
```

Fonte: Elaborada pelo autor

Esta tarefa de criação de filtros e exibição de gráficos foi feita para cada uma das opções laterais do menu, pois a única alteração na UI para os outros casos foi o identificador (ID) de cada opção do menu.

Finalizando a etapa de construção da UI, foi adicionada, ainda, uma opção no menu para informar dados pessoais para contato. A fim de melhorar a exibição destas informações, foi feita a construção utilizando o componente que recebe entrada de HTML, como é ilustrado na Figura 12.

Figura 12 – Área do menu com dados para contato

```
        tabItem(
            tabName = "sobre",
            # ------------------------------------------------------------
            # -- SOBRE -------------------------------------------------
            # ------------------------------------------------------------
            box(
                width = 12,
                HTML("
                    <html>
                        <h3>Este trabalho utilizou dados disponibilizados por (FAPESP, 2020)</h3>
                        <h4><strong>Bases de Dados utilizadas:</strong></h4>
                        <h4> &#8226 Hospital das Clinicas da Faculdade de Medicina da Universidade de Sao Paulo</h4>
                        <h4> &#8226 Grupo Fleury</h4>
                        <h4> &#8226 Hospital Israelita Albert Einstein</h4>
                        <h4> &#8226 Hospital Sírio-Libanês</h4>
                        <hr>
                        <h4><strong>Nome Autor: </strong>Eduardo Eismann</h4>
                        <h4><strong>Orientadora: </strong>Dra. Marta Rosecler Bez</h4>
                        <hr>
                        <h4>Contato:</h4>
                        <h4><strong>E-mail: </strong>eduardo.eismann@gmail.com</h4>
                        <h4><strong>Instagram: </strong><a href=\"https://www.instagram.com/eduardoeismann/\">@eduardoeismann</a></h4>
                        <h4><strong>Twitter: </strong><a href=\"https://twitter.com/EduardoEismann/\">@eduardoeismann</a></h4>
                        <hr>
                        <h4>FAPESP. FAPESP COVID-19 Data Sharing/BR, Available from https://repositoriodatasharingfapesp.uspdigital.usp.br/.
                            Accessed on July 2nd 2020</h4>
                    </html>
                ")
            )
        )
```

Fonte: Elaborada pelo autor

5.4.2 Organização dos dados (data)

Antes de explicar sobre o desenvolvimento do lado servidor, é necessário compreender o que foi construído no arquivo de dados “data.R”. Este arquivo é responsável por, além de direcionar para onde estão salvas as bases de dados, mesclar tabela de pacientes com exames e fazer alguns filtros que serão utilizados pelo servidor.

Inicialmente, foi removida a coluna ID_ATENDIMENTO de duas das bases que foram usadas, pois este campo não tem qualquer influência nas análises e acabaria por deixar o processo mais lento. As colunas em questão estão presentes em duas das quatro bases de dados, base do HSL e HSP. Para remover estas colunas, foram utilizados apenas comandos da própria linguagem R, como apresentado na Figura 13.

Figura 13 – Remoção da coluna ID_ATENDIMENTO

```
1  # -- RM COLUMN ID_ATENDIMENTO ------------------------------------------------
2  excluirColunaIdAtendimento <- c("ID_ATENDIMENTO")
3  # ------------------------------------------------------------------------------

29 dadosExamesSirLib <- dadosExamesSirLib[,!(names(dadosExamesSirLib) %in% excluirColunaIdAtendimento)]
30 dadosExamesHCliSP <- dadosExamesHCliSP[,!(names(dadosExamesHCliSP) %in% excluirColunaIdAtendimento)]
31 head(dadosExamesSirLib)
32 head(dadosExamesHCliSP)
33
```

Fonte: Elaborada pelo autor

Definida a coluna a ser removida nas devidas bases, o próximo passo foi fazer a importação de todos os arquivos de dados que definem as tabelas, ao todo são 9 arquivos, quatro tabelas de pacientes, quatro para os exames, e uma tabela de desfecho.

Para que a importação seja possível, foi utilizado o comando “read.csv” da linguagem R. Este comando possibilita que seja informado, além do caminho do arquivo, alguns parâmetros para a configuração de leitura do arquivo. Neste caso, para tornar possível a leitura, os parâmetros informados foram referentes ao cabeçalho, separador de colunas, delimitador de texto, identificador de decimais e o atributo *fill* para completar, caso exista campos nulos.

Na Figura 14, o trecho de código da linha 7 até 12, é responsável por fazer a importação da tabela de pacientes do HSP. Estes mesmos atributos utilizados como ilustrado na imagem, foram seguidos para todas as outras importações dos arquivos de dados.

Figura 14 – Importação de arquivos de dados e concatenação de dados

```
7  dadosPacientesHCliSP <- read.csv("C:\\Users\\Eduardo\\Desktop\\Bases de Dados Covid\\Hosp Clinicas SP\\HC_PACIENTES_1.csv",
8                                   header = TRUE,
9                                   sep = "|",
10                                  quote = "",
11                                  dec = ".",
12                                  fill = TRUE)
13
14 dadosPacientesAlbEin <- read.csv("C:\\Users\\Eduardo\\Desktop\\Bases de Dados Covid\\Hosp Albert Einstein\\einstein_full_datase
15 dadosPacientesGruFle <- read.csv("C:\\Users\\Eduardo\\Desktop\\Bases de Dados Covid\\Hosp Grupo Fleury\\Grupo_Fleury_Dataset_Co
16 dadosPacientesSirLib <- read.csv("C:\\Users\\Eduardo\\Desktop\\Bases de Dados Covid\\Hosp Sirio Libanes\\hsl_patient_1.csv", he
17
18 dadosPacientes <- rbind( dadosPacientesHCliSP, dadosPacientesAlbEin, dadosPacientesGruFle, dadosPacientesSirLib )
```

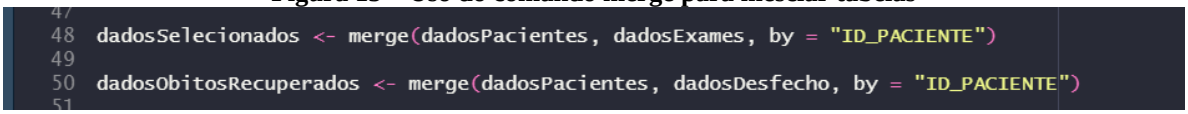
Fonte: Elaborada pelo autor

Além de fazer a importação, ainda na Figura 14, linha 18, é utilizado o comando “rbind” para concatenar todas as tabelas que são correspondentes à mesma informação, assim agrupando os pacientes de todas as entidades em uma única variável da mesma maneira que são agrupados todos os exames em uma mesma variável. A única tabela que não necessita concatenar a outras é referente ao desfecho, pois é exclusiva do HSL.

Para conectar os dados de pacientes e exames, foi utilizado o comando “merge”, informando três parâmetros: a primeira tabela, a segunda, e por fim o campo que é comum entre as duas tabelas para que possa ser feita a conexão, no caso das tabelas usadas, o campo para conectar as duas foi ID_PACIENTE.

Além da relação entre pacientes e exames, também foi necessário criar uma relação entre pacientes e desfecho, para este caso o comando merge também foi utilizado, como ilustrado na Figura 15.

Figura 15 – Uso do comando merge para mesclar tabelas



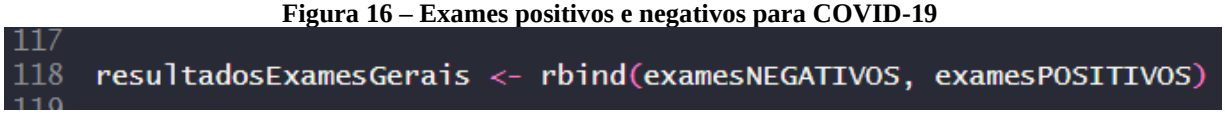
Fonte: Elaborada pelo autor

Na etapa seguinte, foram buscados dois conjuntos de dados, o primeiro referente à exames, no qual os resultados para a presença de coronavírus foram positivos, e o segundo em que a presença de coronavírus não foi identificada.

Para detectar o tipo de exame aplicado, foi lido o campo DE_ANALITO, no qual existem dezessete possíveis nomenclaturas para os exames de detecção de coronavírus. Já o resultado está disponível na coluna DE_RESULTADO.

O resultado dos exames pode apresentar quatro opções, sendo eles: “Detectado” e “Reagente” para casos positivos de COVID-19, “Não detectado” e “Não reagente” para casos negativos. Para fins de popular os campos quantitativos do *dashboard*, os resultados foram agrupados em três variáveis, “examesNEGATIVOS”, “examesPOSITIVOS” e “resultadosExamesGerais”. A criação destas variáveis pode ser identificada na Figura 16, linha 118.

Figura 16 – Exames positivos e negativos para COVID-19



Fonte: Elaborada pelo autor

Com as alterações desenvolvidas para o arquivo de dados, foi possível iniciar o desenvolvimento do lado servidor, em que as tabelas já mescladas e as demais variáveis criadas aqui, são utilizadas.

5.4.3 Servidor

Dentre as funções desenvolvidas no lado servidor, as principais concentram-se em popular as caixas coloridas da UI, que apresentam dados quantitativos para rápida leitura por parte do usuário, fazer a construção dos gráficos a partir dos dados fornecidos pelo arquivo “data.R” e interpretar as ações dos filtros.

No que se trata de quantificação, para detectar o valor total de exames de COVID-19 realizados em todos os hospitais selecionados, foi necessário identificar quais os exames foram aplicados para a detecção de COVID-19.

Para cada nomenclatura de exame, foi criada uma variável que recebe a soma daquele tipo de exame, buscando sempre pelo campo DE_ANALITO. Para a quantificação total, foi utilizada uma variável chamada “totalExamesCovidRealizados”.

Em alguns casos, o mesmo tipo de exame apresentou nomenclatura diferente de um hospital para o outro. Assim, foi necessário checar, em todas as bases, como estava sendo referenciado o nome do exame. A Figura 17 ilustra a variedade de nomes encontrados.

Figura 17 – Diferentes nomenclaturas dos hospitais para exames de COVID-19

```
5
6  server <- function(input, output) {
7
8      exameColetadoNomenclatura01 <- sum(dadosExames$DE_ANALITO == "CoronavirusNL63")
9      exameColetadoNomenclatura02 <- sum(dadosExames$DE_ANALITO == "Resultado COVID-19:")
10     exameColetadoNomenclatura03 <- sum(dadosExames$DE_ANALITO == "CoronavirusOC43")
11     exameColetadoNomenclatura04 <- sum(dadosExames$DE_ANALITO == "Coronavirus229E")
12     exameColetadoNomenclatura05 <- sum(dadosExames$DE_ANALITO == "CoronavirusHKU1")
13     exameColetadoNomenclatura06 <- sum(dadosExames$DE_ANALITO == "COVID IgG Interp aqui")
14     exameColetadoNomenclatura07 <- sum(dadosExames$DE_ANALITO == "Covid 19, Deteccao por PCR")
15     exameColetadoNomenclatura08 <- sum(dadosExames$DE_ANALITO == "Covid 19, Anticorpos IgM, Quimioluminescencia")
16     exameColetadoNomenclatura09 <- sum(dadosExames$DE_ANALITO == "Covid 19, Anticorpos IgA, Elisa")
17     exameColetadoNomenclatura10 <- sum(dadosExames$DE_ANALITO == "Covid 19, Anticorpos IgG, Elisa")
18     exameColetadoNomenclatura11 <- sum(dadosExames$DE_ANALITO == "Teste Rapido para SARS-CoV-2- Pesquisa de anticor")
19     exameColetadoNomenclatura12 <- sum(dadosExames$DE_ANALITO == "Coronavirus 2019-nCoV")
20     exameColetadoNomenclatura13 <- sum(dadosExames$DE_ANALITO == "Coronavirus humano NL63 (Cor63)")
21     exameColetadoNomenclatura14 <- sum(dadosExames$DE_ANALITO == "Coronavirus humano OC43 (Cor43)")
22     exameColetadoNomenclatura15 <- sum(dadosExames$DE_ANALITO == "Coronavirus humano HKU1 (HKU)")
23     exameColetadoNomenclatura16 <- sum(dadosExames$DE_ANALITO == "Coronavirus humano 229E (Cor229)")
24     exameColetadoNomenclatura17 <- sum(dadosExames$DE_ANALITO == "Coronavirus (2019-nCoV)")
25
```

Fonte: Elaborada pelo autor

Como tarefa seguinte, foram construídas as caixas coloridas que exibem informações quantitativas na UI. Estas caixas foram feitas de maneira simples e clara para possibilitar uma fácil e rápida leitura por parte do usuário.

Na Figura 18, é ilustrado, da linha 84 até 89, como é feita a construção de cada uma das caixas, que recebem quatro parâmetros, sendo o valor numérico o principal atributo e, na sequência, é informado o título, um ícone e uma cor. Ao todo, são nove caixas quantitativas, porém como o desenvolvimento é similar, na Figura 18, são apresentadas apenas três destas caixas.

Figura 18 – Criação das caixas com dados quantitativos

```
83
84  output$pacientes_analisados <- renderValueBox({
85      valueBox(
86          nrow(dadosPacientes), "Pacientes Analisados", icon = icon("sort-amount-up"),
87          color = "blue"
88      )
89  })
90
91  output$quantidade_homens <- renderValueBox({
92      valueBox(
93          sum(dadosPacientes$IC_SEXO == "M"), "Quantidade Homens", icon = icon("mars"),
94          color = "purple"
95      )
96  })
97
98  output$quantidade_mulheres <- renderValueBox({
99      valueBox(
100         sum(dadosPacientes$IC_SEXO == "F"), "Quantidade Mulheres", icon = icon("venus"),
101         color = "purple"
102     )
103 })
104
```

Fonte: Elaborada pelo autor

O trecho desenvolvido na sequência também é exibido juntamente com as caixas coloridas, este é referente aos gráficos iniciais apresentados no *dashboard*, onde inicialmente não possível a interação por parte do usuário, pois é apenas a sessão de “Visão Geral”. A interação do usuário acontecerá para todas as opções seguintes do menu.

Para o desenvolvimento de cada um dos gráficos da área de visão geral, utilizou-se o “ggplot”, que transforma os dados vindos do arquivo “data.R”, juntamente com os parâmetros informados, no gráfico que será criado e enviado para a UI, para ser exibido para o usuário.

Dentre os parâmetros necessários para a criação do gráfico por parte do “ggplot” estão a tabela que está sendo analisada, o tipo de dado que estará localizado no eixo X, os nomes a serem exibidos para os eixos X e Y, e mais atributos para cor e contorno. Na Figura 19, é possível perceber como foi feito o desenvolvimento destes gráficos iniciais.

Figura 19 – Usando ggplot para criar gráficos

```
49
50  output$visaoGeralAnoNascimento <- renderPlot({
51      ggplot(
52          dadosPacientes,
53          aes( x = AA_NASCIMENTO )
54      ) + geom_bar( fill = "#3d3d3d", color = "#ffffff", alpha = 0.7, show.legend = TRUE ) +
55      labs( x = "Ano de Nascimento", y = "Quantidade" )
56  })
57
58  output$visaoGeralSexo <- renderPlot({
59      ggplot(
60          dadosPacientes,
61          aes( x = IC_SEXO )
62      ) + geom_bar( fill = "#6070b2", color = "#00ffff", alpha = 0.7, show.legend = TRUE ) +
63      labs( x = "Sexo do Paciente", y = "Quantidade" )
64  })
65
66  output$visaoGeralPacientesPorEstado <- renderPlot({
67      ggplot(
68          dadosPacientes,
69          aes( x = CD_UF )
70      ) + geom_bar( fill = "#126e36", color = "#c9c900", alpha = 0.7, show.legend = TRUE ) +
71      labs( x = "UF dos Pacientes", y = "Quantidade" )
72  })
73
74  output$visaoGeralPacientesPorCidade <- renderPlot({
75      ggplot(
76          dadosPacientes,
77          aes( x = CD_MUNICIPIO )
78      ) + geom_bar( fill = "#126e36", color = "#c9c900", alpha = 0.7, show.legend = TRUE ) +
79      labs( x = "Cidade dos Pacientes", y = "Quantidade" )
80  })
81
```

Fonte: Elaborada pelo autor

A maioria das opções do menu lateral contém os filtros que estão disponíveis para interação do usuário. Para o desenvolvimento destes filtros, também foi usado o “ggplot” e, para cada opção do menu, foi seguido um padrão de desenvolvimento, tornando possível a interação dos quatro filtros sobre cada um dos gráficos a serem exibidos em tela.

Na Figura 20, é utilizado como exemplo o filtro de sexo, presente na opção “Pacientes com Exames”, na qual é informada a tabela “dadosSelecionados” que contêm a tabela de pacientes e exames mescladas, linha 175. Entre as linhas 177 e 183 é possível identificar o controle dos *checkbox*.

Figura 20 – Construção de gráfico interativo com os quatro filtros possíveis

```
171
172     # PACIENTES COM EXAMES POR SEXO
173     output$pacienteComExamePorSexo <- renderPlot({
174         novo <- subset(
175             dadosSelecionados,
176             (
177                 if( input$sexoMasculino == TRUE && input$sexoFeminino == FALSE ) {
178                     IC_SEXO == "M"
179                 } else if( input$sexoMasculino == FALSE && input$sexoFeminino == TRUE ) {
180                     IC_SEXO == "F"
181                 } else {
182                     ( IC_SEXO == "M" | IC_SEXO == "F")
183                 }
184             )
185             & (AA_NASCIMENTO >= input$rangeAnoNascimentoPacientesComExames[1] & AA_NASCIMENTO <= input$range
186             & (
187                 DE_ANALITO == "CoronavirusNL63" | DE_ANALITO == "Resultado COVID-19:" | DE_ANALITO == "Coron
188                 DE_ANALITO == "Coronavirus229E" | DE_ANALITO == "CoronavirusHKU1" | DE_ANALITO == "COVID IgG
189                 DE_ANALITO == "Covid 19, Deteccao por PCR" | DE_ANALITO == "Covid 19, Anticorpos IgM, Quimio
190                 DE_ANALITO == "Covid 19, Anticorpos IgG, Elisa" | DE_ANALITO == "Teste Rapido para SARS-CoV-
191                 DE_ANALITO == "Coronavirus humano NL63 (Cor63)" | DE_ANALITO == "Coronavirus humano OC43 (Co
192                 DE_ANALITO == "Coronavirus humano 229E (Cor229)" | DE_ANALITO == "Coronavirus (2019-nCoV)"
193             )
194             &( CD_UF == input$buscaEstadoPacientesComExames )
195             &( CD_MUNICIPIO == input$buscaCidadePacientesComExames )
196         )
197
198         ggplot(
199             novo,
200             aes( x = IC_SEXO )
201         ) + geom_bar( fill = "#6070b2", color = "#00ffff", alpha = 0.7, show.legend = TRUE ) +
202         labs( x = "Sexo", y = "Quantidade" )
203     })
```

Fonte: Elaborada pelo autor

A linha 195, representada na Figura 20, faz o controle do intervalo selecionado para o ano de nascimento, da mesma forma que as linhas 194 e 195 são responsáveis pelo controle de seleção de estado e cidade.

No caso da seleção de pacientes com exames, foi adicionado, ainda, um filtro extra para que sejam buscados apenas os pacientes com exames de COVID-19. Esta seleção pode ser observada entre as linhas 187 e 192 da Figura 20.

Seguindo este modelo de desenvolvimento para cada opção do menu, foi possível a construção dos demais filtros, alterando apenas o tipo de busca desejada e mantendo todas as demais buscas do filtro.

Da mesma maneira, como foi feito para a visão geral, entre as linhas 198 e 202 da Figura 20, é utilizado o “ggplot” para renderizar o gráfico que será chamado pela UI e em seguida exibido ao usuário.

5.5 HOSPEDAGEM DA APLICAÇÃO

Com a aplicação finalizada e em funcionamento, foi necessário encontrar um serviço de hospedagem para disponibilizar a ferramenta de *dashboard* para o público. Por fornecer um período de um ano gratuito e apresentar simplicidade no que se trata de criação de conta e contratação de serviço, a plataforma escolhida para hospedar a aplicação foi a Amazon AWS, e, dentre os serviços disponibilizados, foi selecionado o Amazon EC2 (*Amazon Elastic Compute Cloud*).

A escolha por este serviço deu-se por conta dos elementos fornecidos, como possibilidade de escolha de sistema operacional, quantidade de memória RAM, unidade de armazenamento com SSD, opções de tráfego de rede e quantidade de processadores.

Para criar uma conta no Amazon EC2, são necessários um e-mail e um cartão de crédito para a verificação de saldo e possíveis cobranças futuras. Todo o processo de criação de conta é extremamente rápido e simples, de forma que qualquer usuário pode criar uma conta apenas seguindo os passos informados pelo fornecedor do serviço.

Para este projeto, foi selecionada a opção de servidor “t2.xlarge”. Este servidor conta com a seguinte configuração:

- Memória RAM: 16 Gigabytes
- Processadores: 4
- Desempenho de rede: Até 1 Gigabits por segundo
- Sistema operacional: Ubuntu

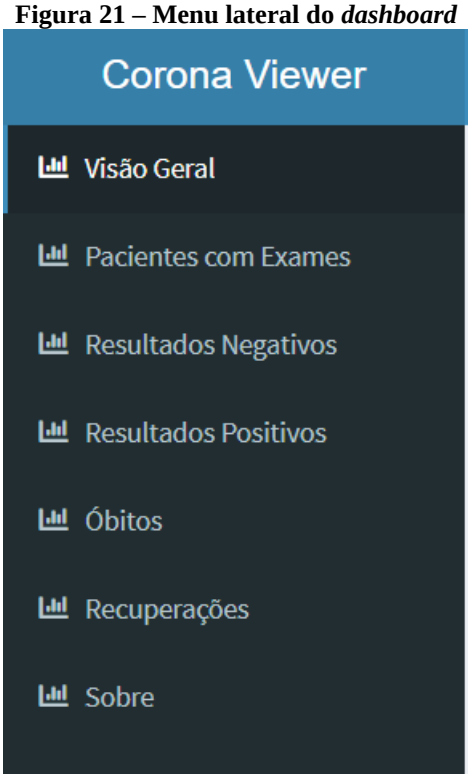
Os passos para configurar o ambiente, na Amazon, e subir a aplicação são apresentados no Apêndice B deste trabalho.

5.6 A APLICAÇÃO FINALIZADA

Ao acessar a aplicação, o usuário se depara com a página inicial do *dashboard*. Esta página contém um menu na lateral esquerda que mostra sete opções, sendo a primeira delas a “Visão Geral”, que é a página atual. Além do menu, o usuário pode ver também. na ordem de cima para baixo, 9 caixas coloridas, contendo dados quantitativos sobre informações gerais das bases analisadas, por exemplo, a quantidade total de pacientes que realizaram exames, além de

outras informações, como quantidades de casos positivos e negativos. Nas Figuras 21 e 22, é possível identificar, na ordem, o menu na lateral esquerda e as caixas coloridas com os dados gerais das bases.

Figura 21 – Menu lateral do *dashboard*



Fonte: Elaboraao pelo autor

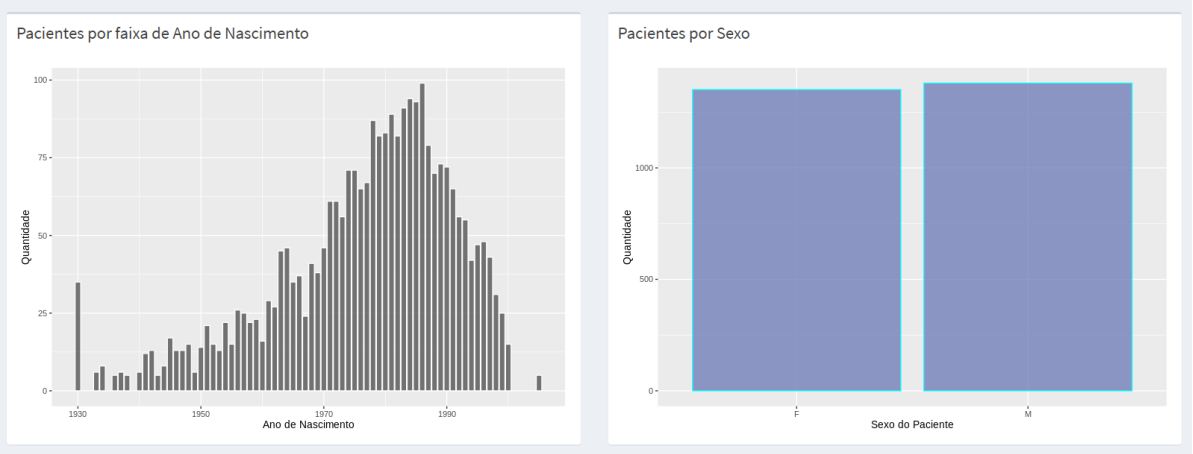
Figura 22 – Caixas coloridas com informações gerais sobre as bases

Fonte: Elaborada pelo autor

Logo abaixo das caixas coloridas, são apresentados os primeiros gráficos do *dashboard*. Estes gráficos da página de “Visão Geral” são estáticos, ou seja, não é possível a interação por meio de filtros, pois têm o intuito de mostrar como está a situação de todos os dados na base.

Os dois primeiros gráficos exibidos ocupam o espaço de uma linha e correspondem, em ordem, a quantidade de pacientes analisados pelo ano de nascimento. O segundo, a quantidade de pacientes por sexo, como é ilustrado na Figura 23.

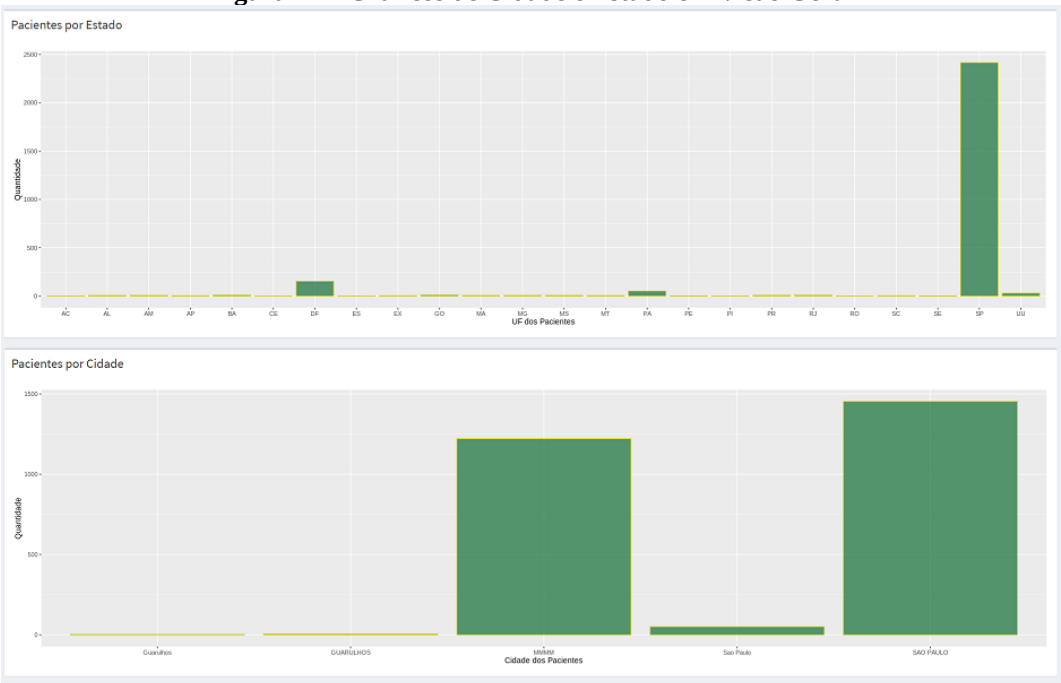
Figura 23 – Gráficos de Ano de Nascimento e Sexo em Visão Geral



Fonte: Elaborada pelo autor

Diferente dos gráficos anteriores e apenas na tela de “Visão Geral”, para cidade e estado, os gráficos foram desenvolvidos de maneira a ocupar a linha inteira, assim ficando um abaixo do outro e proporcionando uma melhor visualização no caso de haverem mais cidades e estados disponíveis nas bases, como é ilustrado na Figura 24.

Figura 24 – Gráficos de Cidade e Estado em Visão Geral



Fonte: Elaborada pelo autor

Ao seguir para qualquer outra das cinco opções do menu, o usuário pode encontrar informações visuais diferentes, além de interagir diretamente com os gráficos, pois o *dashboard* dispõe de quatro filtros que se repetem nas cinco opções subsequentes à “Visão geral”. Estes filtros permitem ao usuário selecionar um intervalo de ano de nascimento, o sexo dos pacientes,

estado e cidade. Após a seleção dos filtros, os gráficos se ajustarão em tempo real, permitindo ao usuário observar a alteração do padrão anterior para o atual.

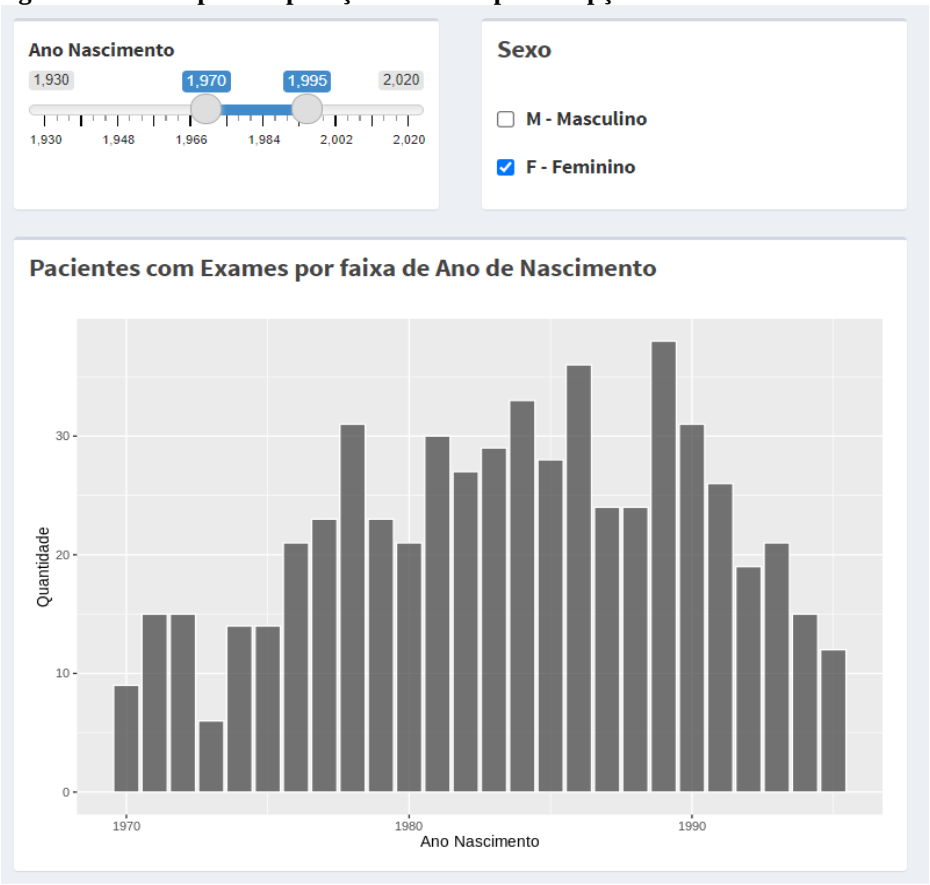
Figura 25 – Filtros padrão para refinar exibição dos gráficos

Fonte: Elaborada pelo autor

Caso o usuário queira verificar duas opções diferentes do menu e aplicar filtros diferentes, a seleção feita em uma das opções não sobrescreve ou apaga a seleção da opção anterior, dessa forma, mantendo a pesquisa do usuário intacta.

A Figura 26 ilustra um recorte da tela do *dashboard* que analisa a opção de "Pacientes com Exames". No exemplo, é selecionado um intervalo de ano de nascimento e apenas um gênero. Como resultado, o gráfico é alterado imediatamente após a seleção. Na Figura 26, o recorte apresenta apenas um gráfico, porém, a modificação se aplica a todos os quatro gráficos visíveis em tela.

Figura 26 – Exemplo de aplicação de filtros para a opção de Pacientes com Exames

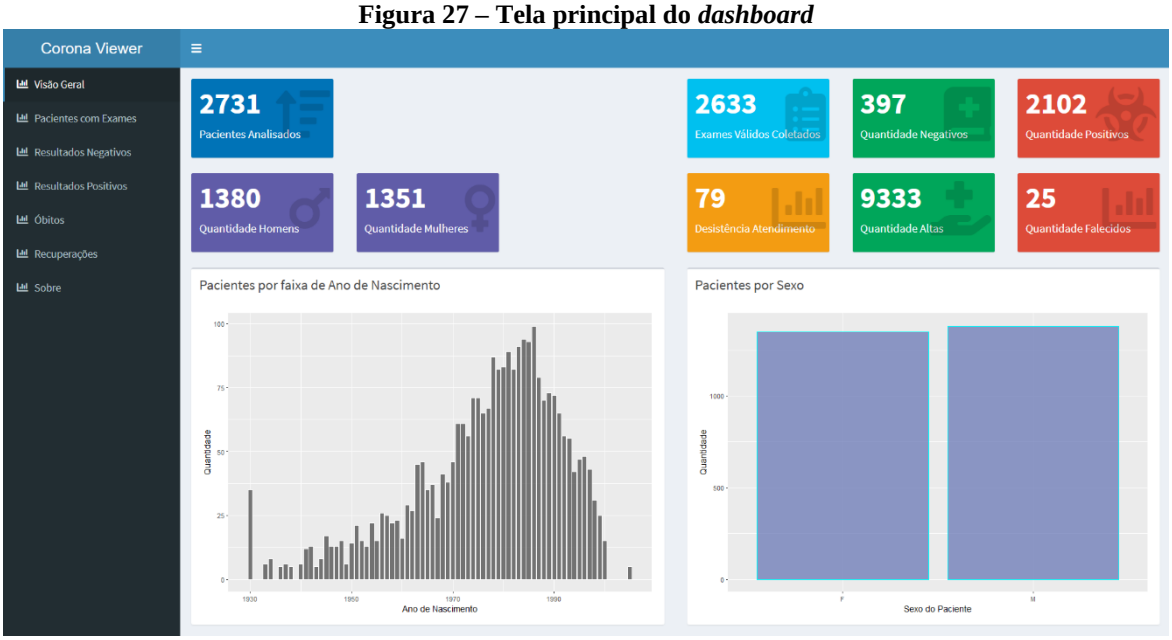


Fonte: Elaborada pelo autor

Além das opções com filtros, ao final do menu, há uma opção “Sobre”, que informa dados referentes à FAPESP, que é a fundação que mantêm os repositórios de dados, o autor do projeto, a professora orientadora, e informações para contato.

As Figuras 27 ilustra a tela principal do *dashboard*, nomeada “Visão Geral”, que é a primeira visualização que o usuário terá ao acessar o *dashboard*, já na Figura 28 é ilustrado a tela de “Pacientes com Exames”, onde é possível ao usuário a interação com os gráficos por meio dos filtros disponíveis na parte superior do *dashboard*.

Figura 27 – Tela principal do *dashboard*



Fonte: Elaborada pelo autor

Figura 28 – Tela com filtros para refinar exibição dos gráficos

Fonte: Elaborada pelo autor

Visto que o foco deste projeto é auxiliar no trabalho de profissionais da área da saúde, este *dashboard* foi exposto a participantes do Grupo de Pesquisa em Computação Aplicada da Universidade Feevale, dentre eles, professores e alunos. Foi solicitado que utilizassem o *dashboard* e respondessem a um questionário para avaliar a ferramenta. O próximo capítulo apresenta os resultados da avaliação realizada.

O *link* para acessar o código fonte do *dashboard* desenvolvido está disponível em: < https://github.com/eduardoeismann/CoronaViewer>.

6 AVALIAÇÃO DO *DASHBOARD*

Para a avaliação do *dashboard* desenvolvido, foi elaborado um questionário no Google Docs (Apêndice A). Foram convidados os participantes do grupo de pesquisa de Computação Aplicada da Universidade Feevale para usar o *dashboard* e responder ao questionário. No total, 12 pessoas contribuíram com suas respostas e, com base nas informações coletadas, cada pergunta será apresentada com suas respectivas respostas.

Nas três primeiras perguntas do questionário, foi solicitado ao avaliado que informasse dados gerais, como profissão, área de atuação e tempo de experiência. No que se refere à área de atuação dos participantes, o grupo contém 3 profissionais da área de saúde, 5 da área de computação, 1 da área de educação, 1 da área de segurança do trabalho, 1 do setor calçadista e 1 que informou ser estudante.

Quanto a profissão dos entrevistados, as respostas foram divididas em diversas profissões: Professor, arquiteto, programador, suporte de TI, técnico de segurança do trabalho, CPO, trabalhador do setor polivalente, técnico administrativo e estudante.

Em relação ao tempo de experiência em sua área de trabalho, o grupo apresentou valores entre um mês e vinte anos, de maneira que o tempo de experiência de cada um foi apresentado conforme o Quadro 11.

Quadro 10 – Tempo de experiência dos profissionais

Tempo de Experiência	Respondente
1 mês	1
2 anos	2
3 anos	4
4 anos	1
5 anos	1
16 anos	1
19 anos	1
20 anos	1

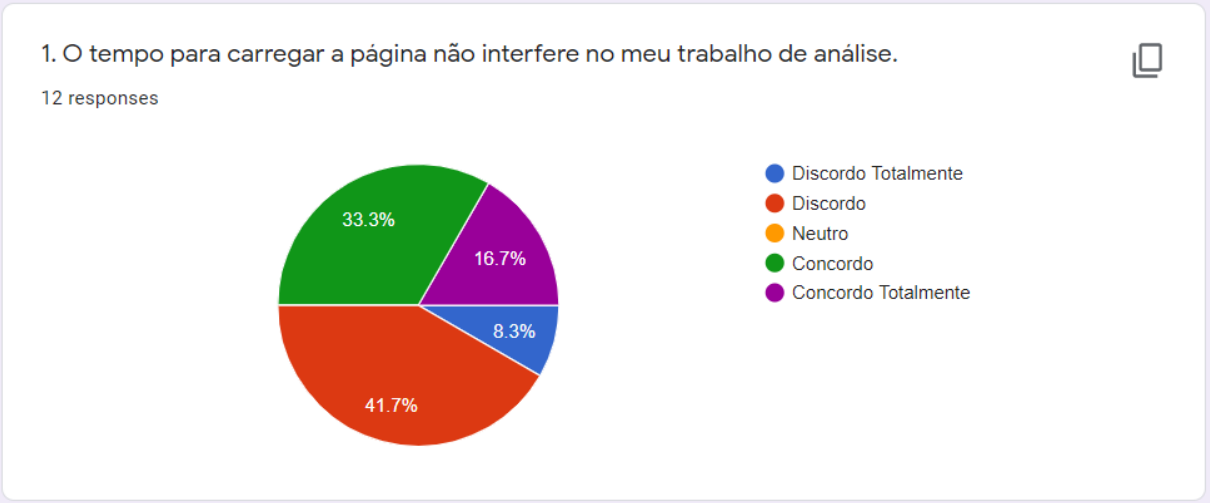
Fonte: Elaborado pelo autor

Após a apresentação dos dados pessoais supracitados, os dados seguintes são referentes à experiência do usuário no que se refere à utilização do *dashboard*, tomando como base as “regras de ouro” descritas por Agni (2015). O formulário de avaliação utiliza um conjunto de afirmações desenvolvidas na Escala Likert, que buscam gerar dados quantitativos para a análise do artefato. A Escala Likert se trata de uma metodologia comum na análise de atitudes, que busca permitir a mensuração de itens intangíveis, como pensamento, ação e sentimento, de uma maneira que possa ser agrupada em um intervalo. (BOONE; BOONE,

2012). Likert propôs o uso de uma escala que vai de 1 a 5, permitindo o desenvolvimento de pesquisas em que o respondente possa inferir a sua percepção, através de cinco opções: (1) Discordo fortemente; (2) Discordo; (3) Neutro; (4) Concordo; (5) Concordo fortemente. Na sequência, são apresentadas as respostas recebidas nos questionários.

A Figura 29 apresenta as respostas sobre a interferência no tempo de carregamento do *dashboard* para desempenhar o trabalho do profissional. De acordo com os dados coletados, 50% dos avaliados inclinaram sua opinião a discordar, enquanto os outros 50% a concordar. Em relação a este número dividido de respostas, é possível inferir que a velocidade da internet de cada um dos participantes teve influência nos resultados, dessa maneira, a experiência de carregamento da página pode ter apresentado tempos diferentes para cada pessoa.

Figura 29 – Tempo para carregar página



Fonte: Elaborada pelo autor

A segunda pergunta, apresentada na Figura 30, é complementar à primeira, em que é questionado se, ao apresentar demora no carregamento da página, é exibido algum indicativo de carregamento. Embora a página não implemente nenhum indicador de carregamento, 41.7% dos entrevistados informaram sua opinião como neutro e 16.7% concordaram, enquanto 33.3% discordaram e 8.3 discordaram totalmente, assim obtendo um nível de discordância de 41.6%. É possível observar que a soma da discordância ainda é menor que os informados como neutro. Desta forma, é possível levar em consideração novamente a rede de internet utilizada, em que, para alguns usuários, as informações são carregadas mais rapidamente do que para outros.

Figura 30 – Informar tempo restante em caso de demora ao carregar página

Fonte: Elaborada pelo autor

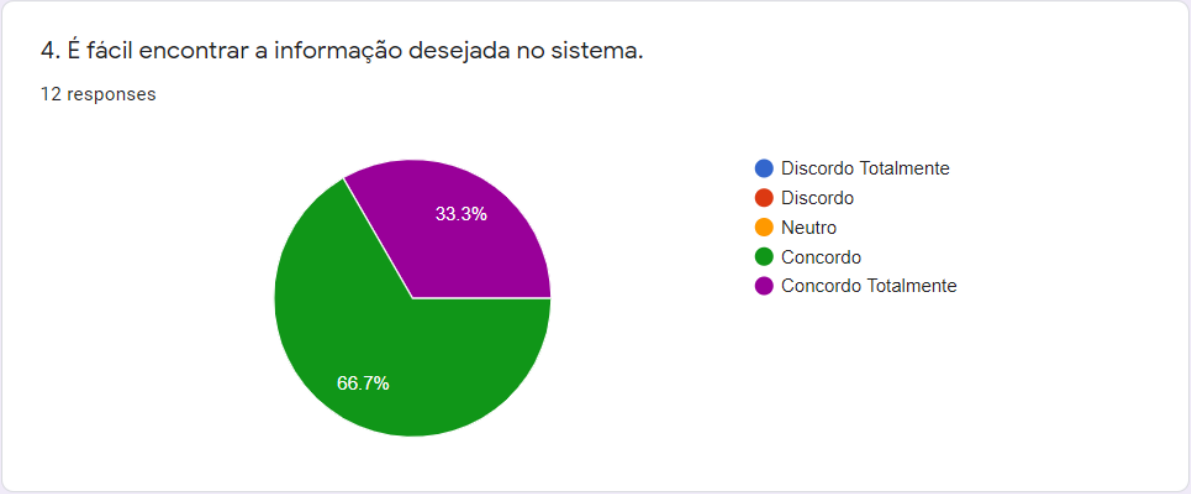
Para a situação do gráfico da Figura 31, os participantes tiveram suas opiniões direcionadas a concordância geral, onde 50% concordaram e 50% concordaram totalmente, no que se refere a facilidade de localizar links e menus no *dashboard*. Esses dados têm grande relevância para o uso do *dashboard*, visto que se refere à facilidade que os usuários demonstram para o utilizar a ferramenta.

Figura 31 – Links e menus facilmente encontrados

Fonte: Elaborada pelo autor

Quanto a facilidade de encontrar informações desejadas, foram exibidos também bons resultados, visto que conforme ilustra a Figura 32, 66.7% do participantes demonstraram concordar e 33.3% concordar totalmente. Percebe-se que nenhum dos usuários demonstrou qualquer dificuldade ou dúvida quanto a localização de informações.

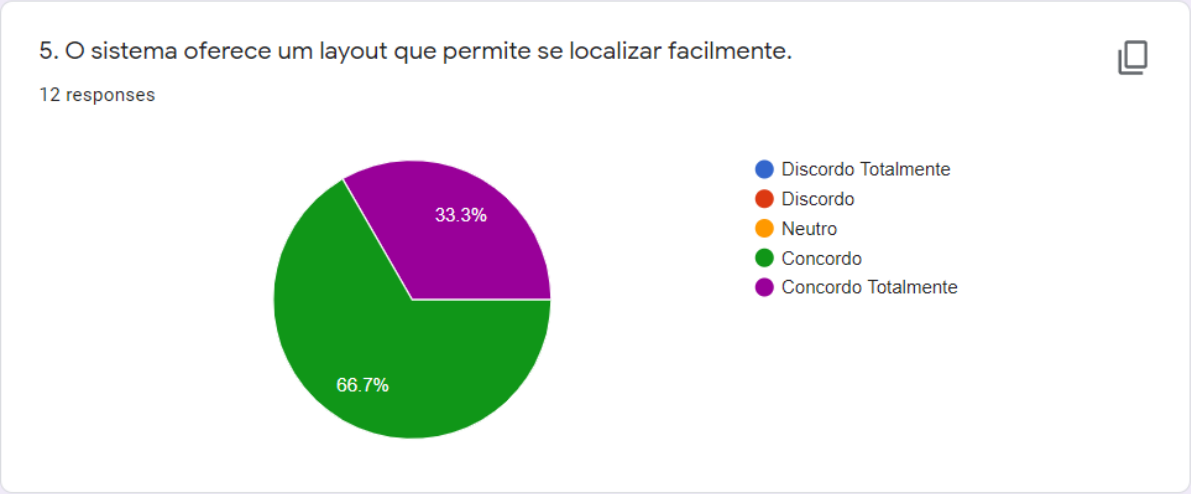
Figura 32 – Facilidade em encontrar informações desejadas no sistema



Fonte: Elaborada pelo autor

A próxima pergunta é referente a facilidade de localização e navegação entre os menus e links por parte do usuário no *dashboard*. Como ilustra a Figura 33, os usuários tiveram o mesmo quadro de respostas da Figura 32, tendo suas opiniões inclinadas a concordância geral. Partindo dos dados avaliados para esta pergunta, é possível observar que não houve qualquer dificuldade de navegação e localização por parte do usuário no decorrer da utilização do *dashboard*.

Figura 33 – Layout que permite fácil localização



Fonte: Elaborada pelo autor

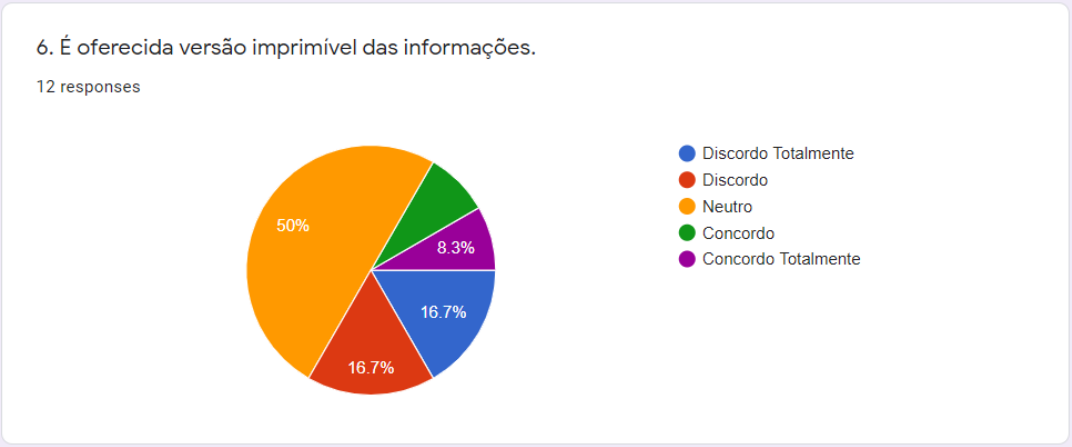
Como parte do que se refere a usabilidade do *dashboard*, foi incluída uma pergunta referente a impressão da página, mesmo que esta opção não tenha sido oferecida pelo *dashboard* diretamente ao usuário, pois muitos navegadores já a oferecem.

Porém, o motivo principal de não haver uma opção de impressão, é relacionado ao impacto no meio ambiente que a utilização de papel proporciona. Segundo o Instituto Information Management (2020), estima-se que cada brasileiro utiliza aproximadamente 44

quilos de papel por ano, gerando impacto não somente pelo consumo de árvores, mas também pelo consumo de água para a irrigação que, de acordo com a Embrapa (Empresa Brasileira de Pesquisa Agropecuária) para cada uma tonelada de papel produzido são utilizados 540 mil litros de água e entre duas e três toneladas de madeira.

No que se refere à percepção dos avaliados quanto à pergunta ilustrada na Figura 34, 50% se mostraram neutros quando há existência de uma opção de impressão, enquanto os outros 50% ficaram divididos entre as demais opções. Os respondentes que concordaram ou concordaram totalmente podem estar se referindo à opção fornecida pelo navegador de sua utilização, enquanto os que discordaram e discordaram totalmente, avaliaram o fato de a opção não estar presente dentro do *dashboard*.

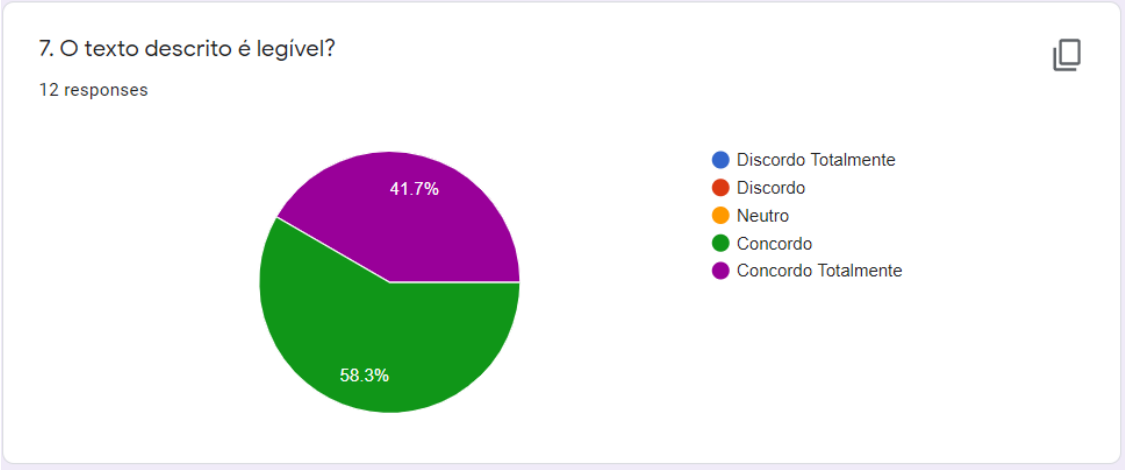
Figura 34 – Versão imprimível da página



Fonte: Elaborada pelo autor

A pergunta seguinte refere-se ao quão legível são os textos descritos ao longo do *dashboard*. As respostas dos avaliados foram 100% positivas, onde 58.3% concordaram e 41.7% concordaram totalmente, conforme ilustra a Figura 35.

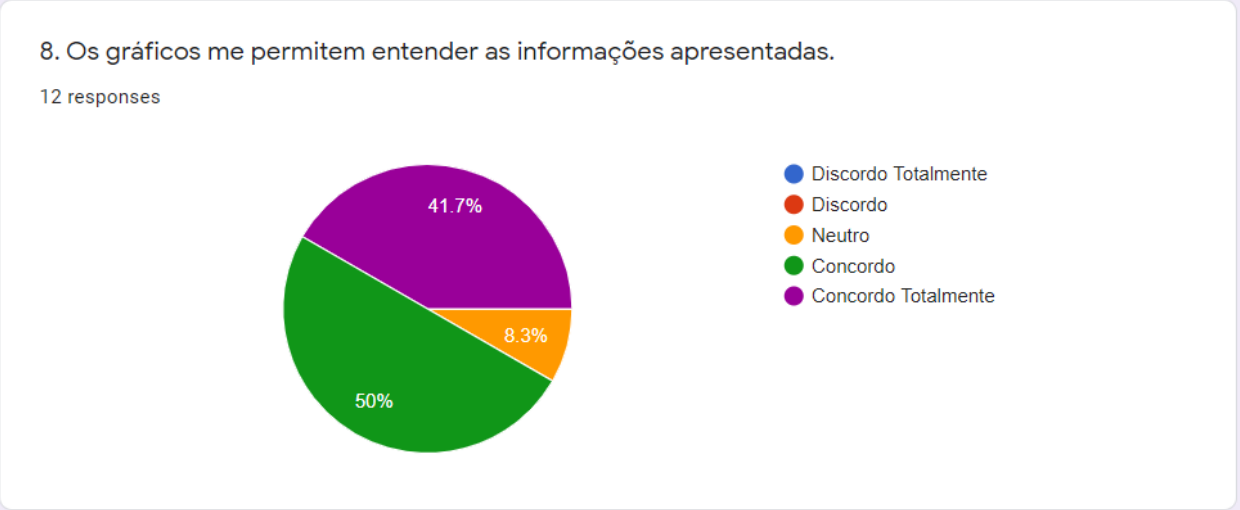
Figura 35 – Texto descrito na página é legível



Fonte: Elaborada pelo autor

Sobre a compreensão das informações apresentadas pelo *dashboard*, de acordo com a ilustração do gráfico na Figura 36, é possível perceber que 50% dos respondentes concordaram, 41.7% concordaram totalmente e apenas 8.3% se mostraram neutros. Assim, tem-se um entendimento das informações apresentadas de 91.7% por parte dos participantes.

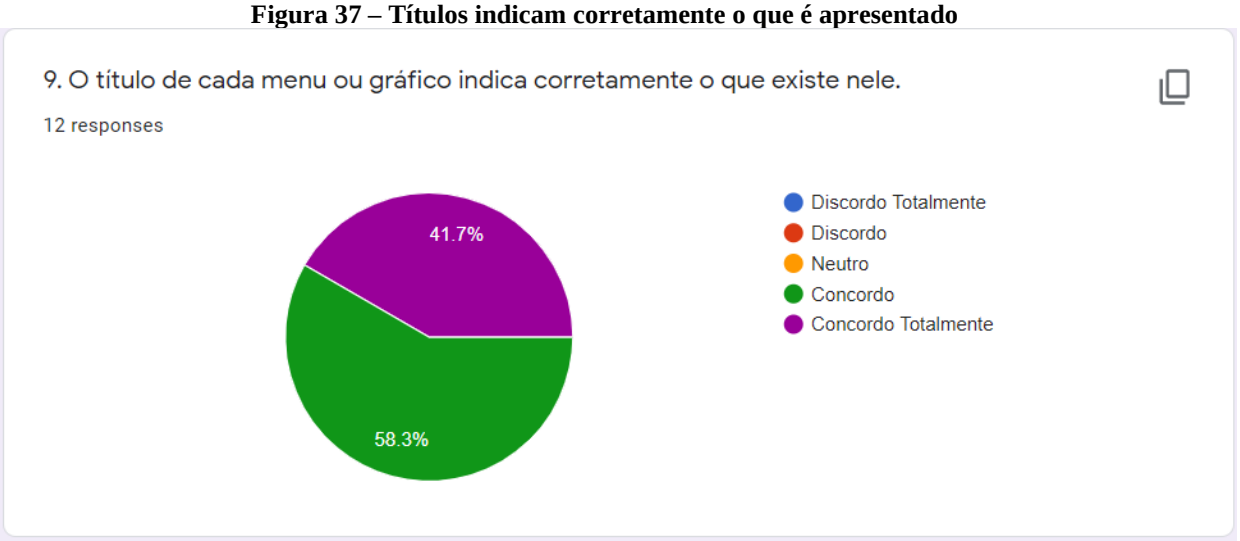
Figura 36 – Gráficos permitem entender as informações apresentadas



Fonte: Elaborada pelo autor

A ilustração seguinte, Figura 37, é relacionada à correta indicação dos títulos utilizados no *dashboard*, com concordância de 100% dos usuários, nos quais 58.3% concordaram e 41.7% concordaram totalmente.

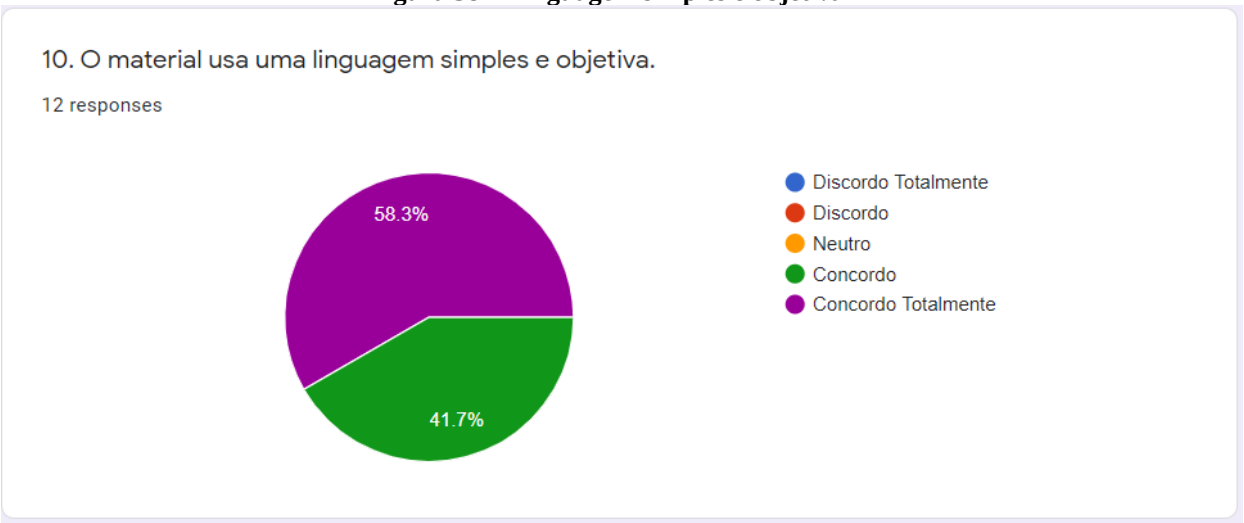
Figura 37 – Títulos indicam corretamente o que é apresentado



Fonte: Elaborada pelo autor

Quanto ao *dashboard* utilizar uma linguagem simples e objetiva, todos os respondentes concordaram com esta afirmação, 58.3% concordaram totalmente e 41.7% concordaram, como é ilustrado no gráfico da Figura 38.

Figura 38 – Linguagem simples e objetiva



Fonte: Elaborada pelo autor

A pergunta seguinte, Figura 39, refere-se a não ser necessária ajuda para entender o *dashboard*. A maioria dos respondentes, 91.7%, concordou com a afirmação, apenas dividindo opiniões entre 41.7% que concordam e 50% que concordam totalmente. 8.3% responderam como neutros. Visto que, além de profissionais da saúde, também há profissionais de outras áreas, isso pode ser caracterizado como um caso em que os respondentes se mantiveram neutros pois não têm o conhecimento específico na área.

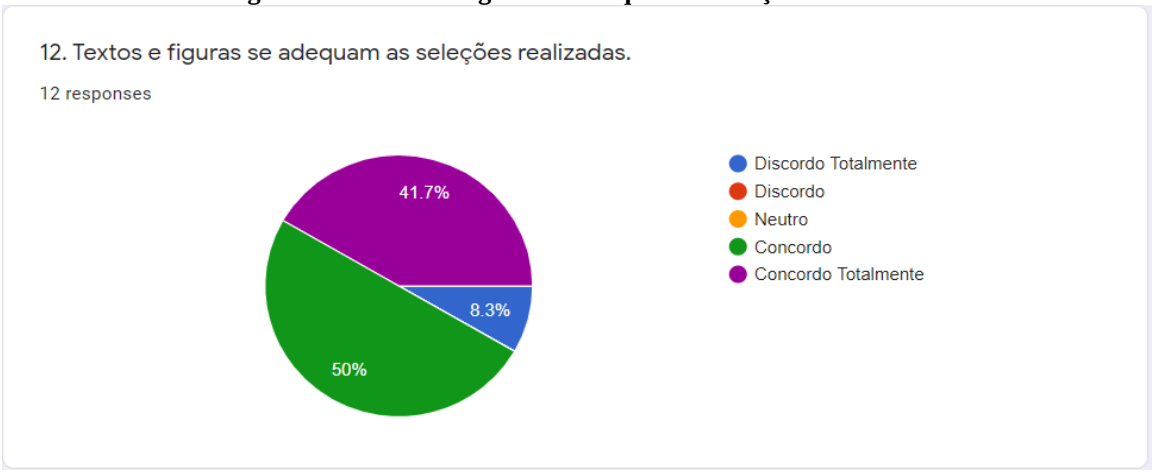
Figura 39 – Não é necessário recorrer a ajuda para entender o sistema

Fonte: Elaborada pelo autor

Para a pergunta seguinte, foi questionado acerca dos textos e figuras se adequarem às seleções realizadas. De acordo com o gráfico ilustrado na Figura 40, foi possível perceber que a maioria, 91.7%, concordou com esta afirmação, porém 8.3% discordaram totalmente. Esta

discordância pode ser relacionada a algum erro de resolução de tela, ou mesmo a algum *bug* do *dashboard* que tenha passado despercebido por todos os outros avaliadores.

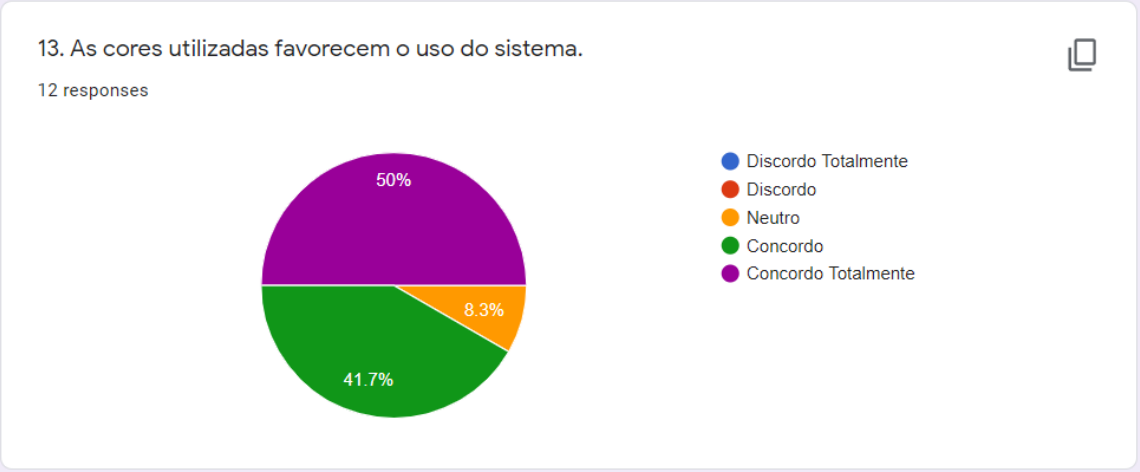
Figura 40 – Textos e figuras se adequam as seleções realizadas



Fonte: Elaborada pelo autor

A pergunta subsequente refere-se ao uso de cores no *dashboard*, e se esta utilização favorece o uso da ferramenta. Para 91.7% dos respondentes, as cores favorecem o uso do *dashboard*, desta porcentagem, 50% concordam totalmente com esta afirmação, enquanto 41.7% apenas concordam. 8.3% se mostraram neutros a esta afirmação. Os dados podem ser observados conforme ilustra a Figura 41.

Figura 41 – Cores favorecem o uso do sistema

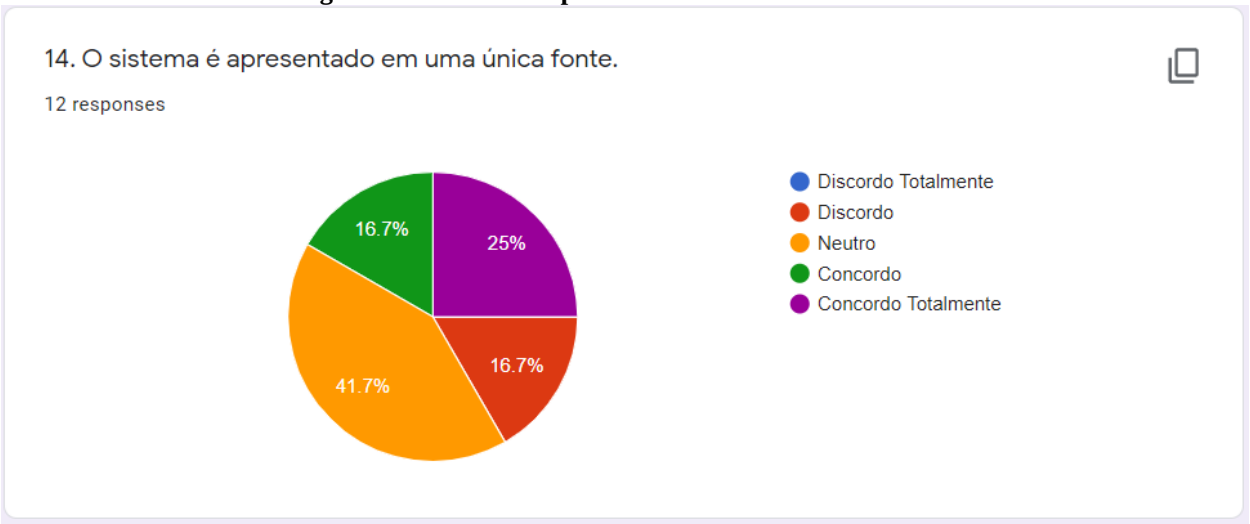


Fonte: Elaborada pelo autor

Ao observar o gráfico que é ilustrado pela Figura 42, foi possível perceber que houve uma grande divergência de respostas no que se trata à pergunta de se o sistema é apresentado em uma única fonte. 41.7% indicaram a opção neutra para esta afirmação, enquanto 25% concordaram totalmente. 16.7% concordaram e 16.7% discordaram. A alta variação desta resposta pode ser indicada por alguns fatores como o tamanho da fonte apresentada em cada

área do *dashboard*, a presença de elementos destacados em negrito e a ocorrência de fontes que utilizam cores diferentes para representar valores.

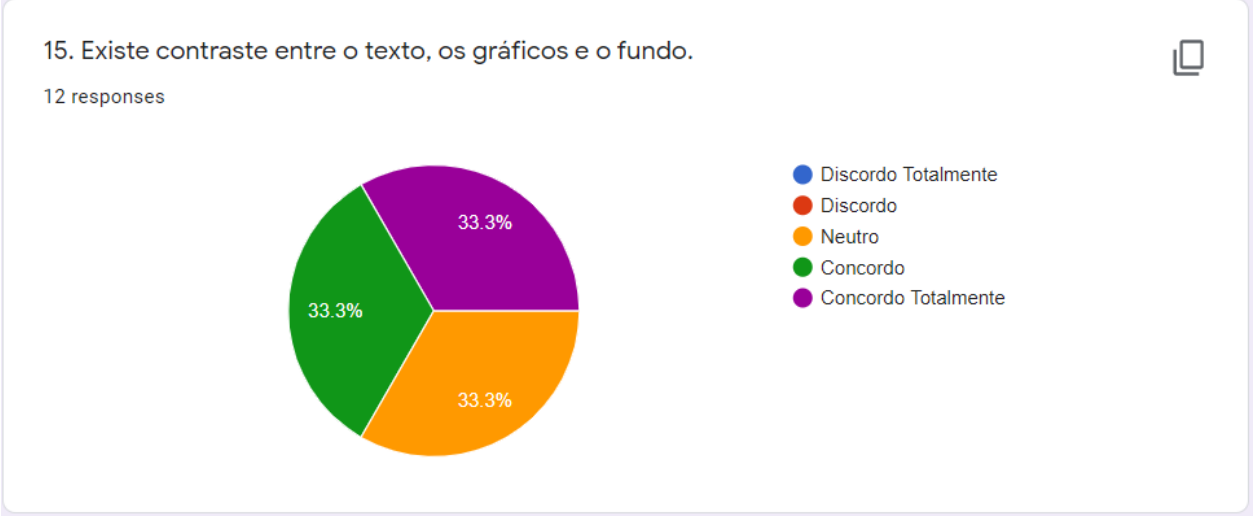
Figura 42 – Sistema é apresentado em uma única fonte



Fonte: Elaborada pelo autor

Para a pergunta referente à existência de contraste entre o texto, os gráficos e o fundo, o grupo ficou dividido igualmente entre três das opções, em que um terço concorda totalmente, um terço concorda e um terço mostrou-se neutro. O *dashboard* utiliza de cores para destacar e diferenciar os elementos e facilitar a visualização por parte do usuário, porém, não há um grande contraste entre eles, de maneira que os dois terços que concordaram tendem a ter levado em consideração o uso de cores, enquanto os 33.3% que se mostraram neutros podem ter apresentado imparcialidade para esta característica. Esta divisão de opiniões é ilustrada pela Figura 43.

Figura 43 – Contraste entre textos, gráficos e o fundo



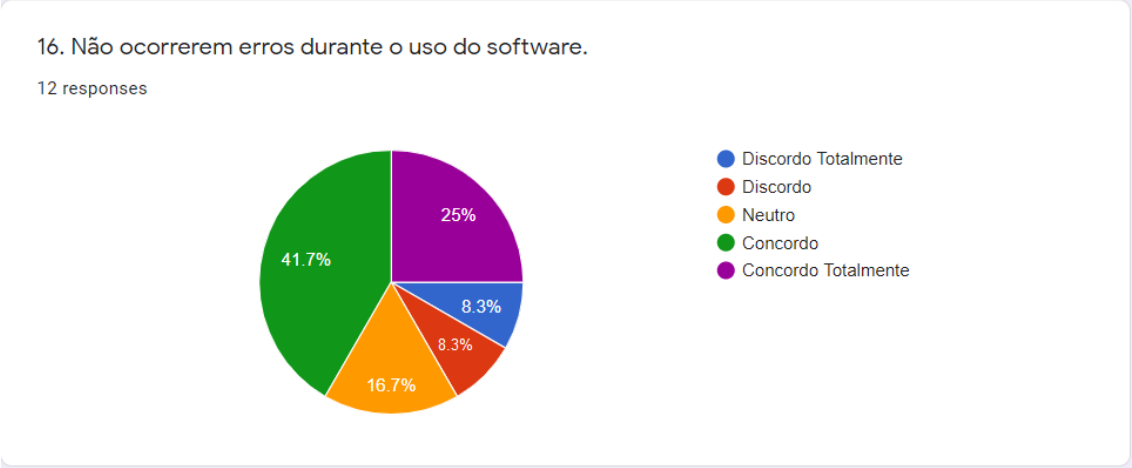
Fonte: Elaborada pelo autor

No que se trata da ocorrência de erros durante o uso do *dashboard*, a opinião dos usuários ficou dividida entre todas as opções. De acordo com o ilustrado na Figura 44, entre os respondentes, 41.7% concordaram com a afirmação, 25% concordaram totalmente, 16.7% se demonstraram como neutros, 8.3% discordaram e 8.3% discordaram totalmente.

Mesmo levando em consideração que a maioria, 66.7%, afirmou não haver erros, 16.6% dos usuários se mostraram contrários à afirmação, de maneira que existe a possibilidade de alguns erros terem passado despercebidos por todos os demais usuário.

Nos 16.7% que se mostraram neutros, é difícil inferir uma correta constatação, visto a questão proposta.

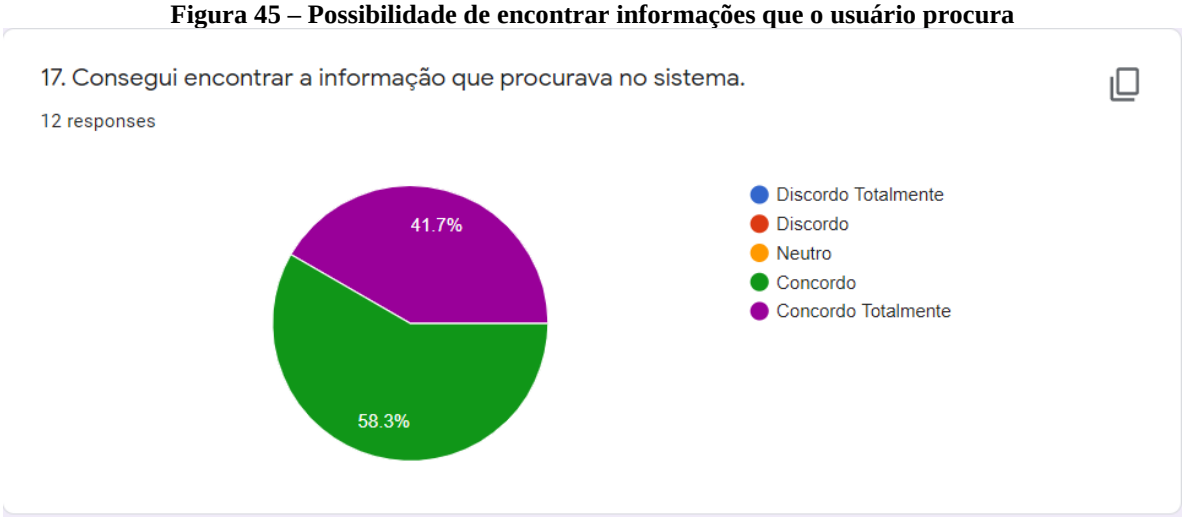
Figura 44 – Não ocorrem erros durante o uso do *dashboard*



Fonte: Elaborada pelo autor

Referente à facilidade por parte do usuário de encontrar informações no *dashboard*, 58.3% concordaram com a afirmação e 41.7% concordaram totalmente, mostrando, assim, uma aceitação de 100%, como ilustra o gráfico da Figura 45.

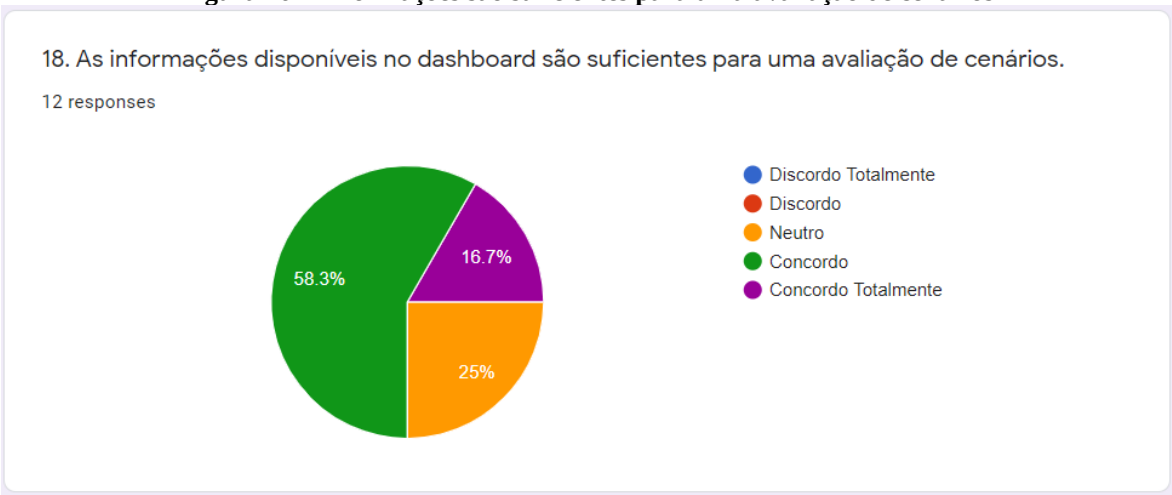
Figura 45 – Possibilidade de encontrar informações que o usuário procura



Fonte: Elaborada pelo autor

A pergunta seguinte questiona se as informações disponíveis no *dashboard* são suficientes para uma avaliação de cenários e, de acordo com as respostas, 75% se mostraram positivos à afirmação, em que 58.3% concordaram e 16.7% concordaram totalmente. 25% dos usuários demonstraram-se neutros à afirmação que, neste caso, é possível inferir que pode ter interferência do conhecimento técnico por parte do respondente. Na Figura 46, é ilustrada a concordância da maioria dos usuários acerca da questão.

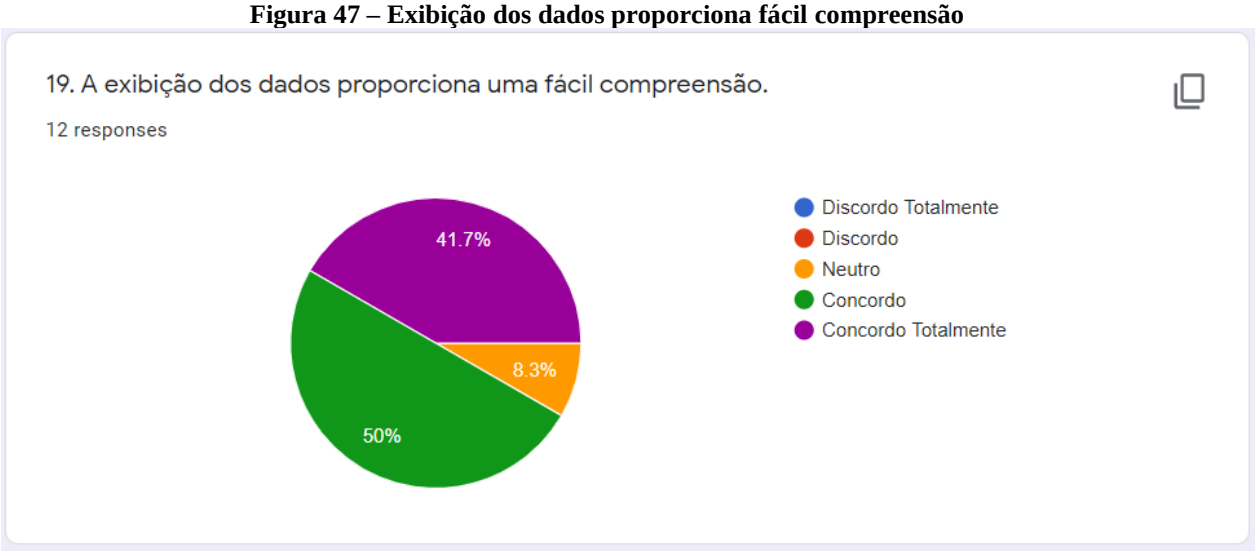
Figura 46 – Informações são suficientes para uma avaliação de cenários



Fonte: Elaborada pelo autor

Complementar à anterior, a questão ilustrada pela Figura 47 refere-se à facilidade de compreensão dos dados por parte do usuário. 50% das pessoas responderam concordar com a afirmação, enquanto 41.7% concordaram totalmente, deixando apenas 8.3% com opinião neutra. É possível constatar que a maioria dos usuários, 91.7%, compreendeu facilmente os dados que são exibidos pelo *dashboard*.

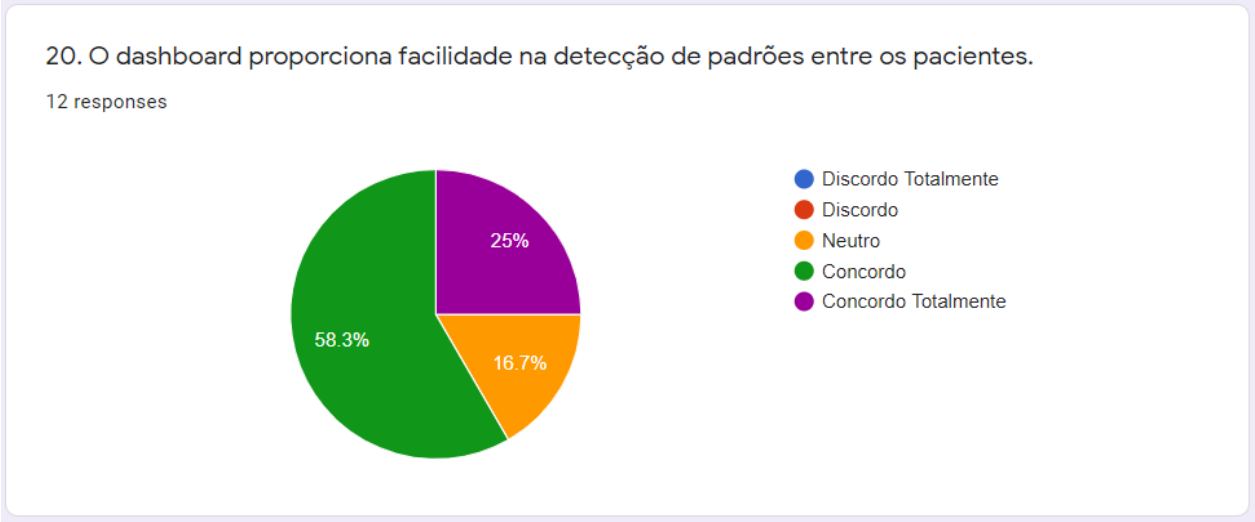
Figura 47 – Exibição dos dados proporciona fácil compreensão



Fonte: Elaborada pelo autor

A pergunta seguinte, questiona sobre o *dashboard* proporcionar facilidade no que se trata da detecção de padrões entre os pacientes. Acerca deste questionamento, a maioria dos usuários mostraram-se concordantes, onde 58.3% concordaram e 25% concordaram totalmente. Deixando ainda uma margem de 16.7% de usuários que se apresentaram neutros. A Figura 48 ilustra a alta taxa de concordância por parte dos usuários.

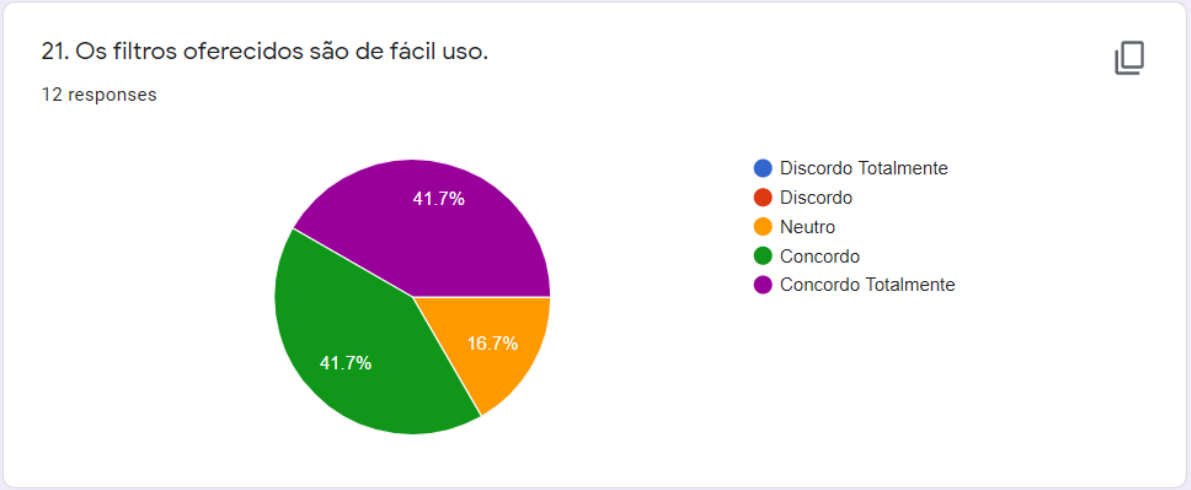
Figura 48 – *Dashboard* proporciona facilidade na detecção de padrões entre pacientes



Fonte: Elaborada pelo autor

Quanto aos filtros oferecidos aos usuários, é questionado sobre a facilidade de utilização, de acordo com a ilustração no gráfico da Figura 49, 83.4% mostraram achar fácil o uso dos filtros, destes, o grupo ficou dividido, onde 41.7% concordaram e 41.7% concordaram totalmente. Ainda houve um grupo de pessoas que se mostraram neutros ao uso dos filtros.

Figura 49 – Filtros oferecidos são de fácil uso

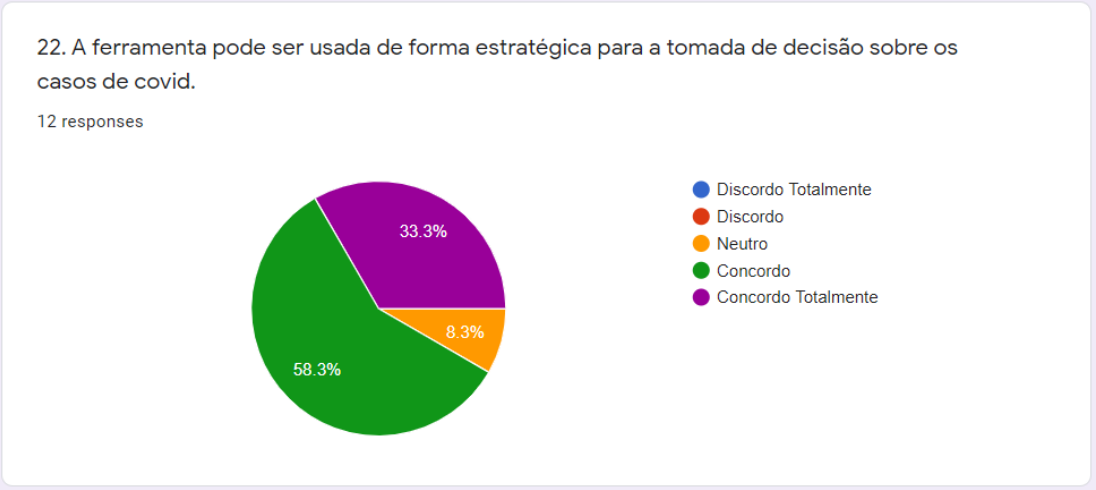


Fonte: Elaborada pelo autor

A questão subsequente indaga os usuários sobre o *dashboard* ser uma ferramenta que pode auxiliar de forma estratégica para a tomada de decisão no que se trata dos casos de COVID-19. Para esta questão, 58.3% dos usuários mostraram concordar, um terço dos usuários concordaram totalmente e 8.3% foram neutros sobre a questão.

Na Figura 50, é ilustrado o gráfico com o alto nível de concordância dos usuários. Assim, é possível compreender que o *dashboard* pode auxiliar de maneira positiva profissionais da área no que se trata de tomada de decisão.

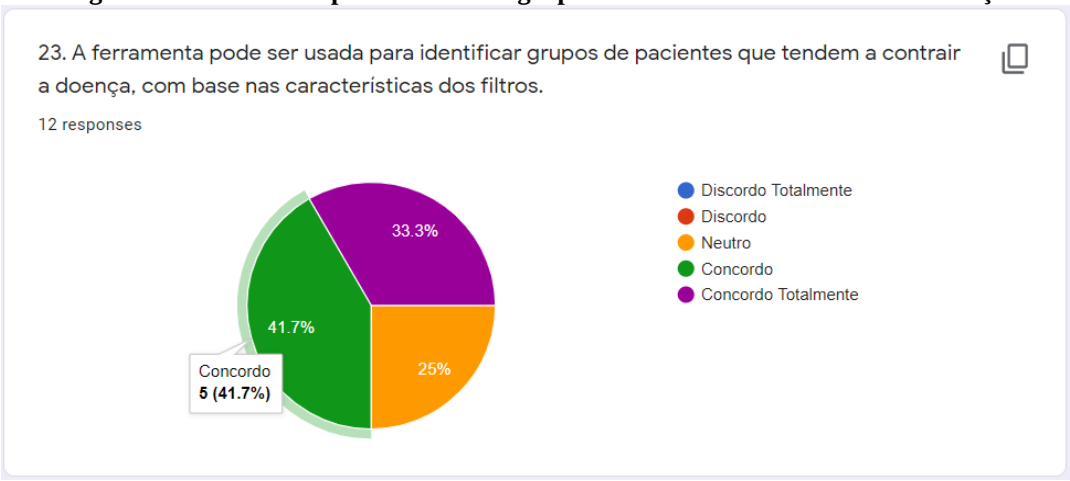
Figura 50 – Ferramenta estratégica para tomada de decisão



Fonte: Elaborada pelo autor

A questão final do questionário, ilustrada na Figura 51, indaga os usuários sobre o *dashboard* auxiliar na identificação de grupos com tendência a contrair a doença. Para esta questão, 41.7% dos usuários informaram concordar com a possibilidade de identificar grupos tendenciosos, enquanto 33.3% concordaram totalmente e 25% mostraram-se neutros a esta questão.

Figura 51 – *Dashboard* pode identificar grupos com tendência de contrair a doença



Fonte: Elaborada pelo autor

Ao fim do questionário, é exposto aos usuários um campo opcional descritivo para críticas e sugestões, para que o *dashboard* possa estar em constante aprimoramento. Dentre as doze pessoas do grupo de participantes, três responderam ao questionário, em que as respostas podem ser localizadas no Quadro 11.

Quadro 11 – Críticas e sugestões dos usuários sobre o uso do *dashboard*

Pessoa entrevistada	Resposta
Anônimo 1	“exceto em "visão geral", eu não acho que seja necessário colocar o gráfico [pacientes por estado/cidade] uma vez que o filtro dessas páginas não permite deixar o filtro cidade ou estado vazio e trazer mais de uma. Acredito que um marcador de quantas pessoas são a amostra para o filtro selecionado já resolva.”
	“a versão para impressão seria usar a opção que o navegador permite? Não encontrei um botão dedicado para impressão em si.”
	“O gráfico de barras é bom de acompanhar grandes volumes de dados, mas se pudesse ter um contador geral das informações contidas no gráfico. Por exemplo, o gráfico por sexo mostra 500 e 400 F e M. Poderia ter um somatório informando 900 abaixo ou acima, porque os indicadores na página da [visão geral] contabiliza todos os dados, e não é possível aplicar filtros como nos demais para ter essa visão geral em números.”
	“a opção de [pacientes com exames], é uma informação importante, mas acho que você podia comparar com algo para ser mais relevante. Por exemplo, "população da cidade vs população testada", ou "testes realizados por período", por exemplo, como ocorreram as ondas de teste (e quando ficou mais intensa a campanha de teste). Talvez com essa informação você consiga nos dar a temperatura de qual o impacto de testar mais, se desacelera o contágio ou ajuda a diminuir a intensidade da doença em evoluções graves e óbitos.”
	“Obs: Claro que contando que este *dataset* te passe alguma informação sobre data do teste, da coleta ou do resultado.”
Anônimo 2	“Ferramenta bem interessante e fácil de usar e entender, penso que se o valor/percentual aparecesse ao clicar em cima da coluna do gráfico seria ainda mais fácil a compreensão dos dados.”

Pessoa entrevistada	Resposta
Anônimo 3	“Necessário ajustes e atualização dos dados do combo após mudanças. Exemplo: após selecionar um estado, é possível a seleção da cidade, caso após a seleção de um município, for feito alteração do estado, os dados dos municípios não são alterados, ficando a lista de municípios do estado anterior.”

Fonte: Elaborado pelo autor

Um dos participantes, aqui chamado de “Anônimo 1”, expos quatro pontos importantes: inicialmente, comentou sobre a baixa necessidade de gráficos de estado e cidade nas demais opções do menu, que não a opção “Visão Geral”, além de acrescentar um marcador para informar quantas pessoas são amostradas para o filtro.

Na sequência, informou que não localizou a opção de impressão, indagando se deveria fazer o uso da opção fornecida pelo navegador. Esta resposta relaciona-se à ilustração do gráfico da Figura 32, na qual houve diferentes resultados, sendo que a maioria, 50%, apresentou opinião neutra. Porém, optou-se por não exibir uma opção de impressão decorrente dos impactos ao meio ambiente descritos pelo Instituto *Information Management* (2020).

O terceiro ponto exposto é referente ao gráfico de barras ser bom para acompanhar grandes volumes, porém, trouxe a necessidade de ser incluso um contador para os valores totais exibidos, visto que, nos gráficos, são exibidos os valores somente para as colunas apresentadas.

O quarto e último ponto exposto ressalta que os dados referentes a “Pacientes com Exames” sejam importantes, porém seria necessário relacionar com outro dado que apresente relevância, como população da cidade ou quantidade de testados, para que, assim, possa ser visto se o volume de testes tem impacto em desacelerar o contágio ou mesmo diminuir a intensidade dos casos.

Como observação ainda, o entrevistado indaga sobre a existência de data de teste, data de coleta e data de resultado nas bases de dados e, em resposta, é possível afirmar, com base nas bases de dados analisadas, que a informação de data de coleta está presente para todos os exames, porém a data de teste e a data de resultado não estão disponíveis nas bases de dados.

O próximo entrevistado, “Anônimo 2”, relata que o *dashboard* é uma ferramenta interessante, fácil de usar e entender, e dá como sugestão de melhoria a opção de, ao clicar sobre uma das colunas, o dado de porcentagem seja exibido para facilitar ainda mais a compreensão por parte do usuário.

A última pessoa a responder o campo de críticas e sugestões, “Anônimo 3”, relata que existe a necessidade de ajustes nos campos de seleção de estado e cidade, pois, ao alterar um dos estados, o campo de cidade não estaria atualizando os dados.

Visto as críticas e sugestões acima relatadas, é possível identificar qual o rumo tomar para a evolução do projeto do *dashboard* e, assim, manter um processo de constante melhoria para melhor servir ao usuário final.

7 CONCLUSÃO

No decorrer deste estudo, foi investigado sobre a ocorrência de doenças respiratórias no ser humano e, relacionado a estas enfermidades, uma pesquisa foi feita no que se refere à família de vírus Coronavírus, com foco principal no vírus Sars-Cov-2, conhecido também por COVID-19, visto a grande ameaça à saúde que se mostrou nos anos de 2020 e 2021. Juntamente com a pesquisa sobre o vírus, foi pesquisado sobre métodos de mineração de dados e descoberta de conhecimento, em paralelo com pesquisas sobre a linguagem de programação R, que serviu como ferramenta para a exibição final dos dados.

Na continuidade deste trabalho, foi feita a união das quatro bases de dados selecionadas e, na sequência, a aplicação do processo de KDD à base então unificada. Com os dados já preparados, a etapa seguinte consistiu em utilizar a linguagem de programação R para construir visualizações de possíveis cenários e, como última etapa, foi construída uma interface que exibe essas visualizações para o usuário final, tornando possível a análise por parte de profissionais e, assim, auxiliando na tomada de decisão.

Para fins de validação do *dashboard* desenvolvido, a ferramenta foi disponibilizada para participantes do Grupo de Pesquisa em Computação Aplicada da Universidade Feevale, em que, após utilizarem o *dashboard*, responderam a um questionário visando a usabilidade da ferramenta e a contribuição que ela oferece para a tomada de decisão.

Como base nos resultados do questionário, a ferramenta foi vista como de fácil uso e compreensão, por parte do usuário, e pode, sim, ser utilizada para auxiliar no processo estratégico de tomada de decisão, pois, de acordo com os resultados coletados no questionário, os dados exibidos pelos gráficos são suficientes para uma avaliação de cenários e, com base nas visualizações proporcionadas pelo *dashboard*, é possível identificar grupos de tendência no que se refere à contaminação por coronavírus.

Para fins de afirmar com maior eficácia a relevância técnica do *dashboard*, é exposto a necessidade da avaliação do *dashboard* por parte de profissionais da saúde, pois este grupo tende a ser o principal usuário da ferramenta.

Da maneira como foi implementado o uso das bases de dados, é possível agregar novas bases, desde que estejam de acordo com a formatação utilizada e descrita na presente pesquisa, visto desta característica de expansão, a cada nova base disponibilizada, mais características serão exibidas acerca dos grupos de pacientes.

Para fim de proporcionar a identificação de mais grupos de tendências e a descoberta de novas características nos pacientes, foram sugeridas melhorias por parte dos usuários.

Algumas delas relacionadas ao uso dos dados de estados e cidades, outras destacando a possibilidade de ser utilizada a data de realização dos exames. Todas estas informações que estão disponíveis nas bases podem contribuir para o desenvolvimento de novos filtros, assim, refinando ainda mais os resultados.

De acordo com os resultados coletados no questionário, percebe-se que o *dashboard* oferece diversos caminhos para a continuidade do projeto, desde as melhorias em determinadas funções até o desenvolvimento de novas visualizações e novas formas de se utilizar os dados oferecidos pelas bases. Dentre as melhorias citadas para a continuidade do projeto, uma delas foi destacando a importância de confrontar determinados dados do *dashboard*, colocando diferentes dados lado a lado, para que o usuário possa determinar com maior facilidade características entre os dados visualizados.

Outro ponto destacado como melhoria futura, é a importância da utilização de ontologias afim de melhorar a visibilidade do *dashboard* para a área da saúde.

A partir do que foi exposto, juntamente com a ferramenta desenvolvida e as sugestões e críticas coletadas, conclui-se que o *dashboard* proposto é útil ao usuário final no que tange a usabilidade e relevância dos dados apresentados, proporcionando ao usuário da ferramenta o conhecimento no que relaciona aos grupos com maior e menor ocorrências de contágio pelo Coronavírus, auxiliando, assim, na detecção de cenários e na tomada de decisão quanto aos grupos de tendência.

REFERÊNCIAS

AGNI, Eduardo. **As oito regras de ouro do design de interfaces**. Conheça os princípios que podem tanto orientar a concepção quanto a avaliação da maioria dos sistemas interativos. UXDesign, [S.l.], outubro 2015. Disponível em: <https://uxdesign.blog.br/as-oito-regras-de-ouro-do-design-de-interfaces-836fb166d36b>. Acesso em: 22 mai. 2021.

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 2. ed. Cambridge: The MIT Press, 2010.

AZEVEDO, Julia. E Cycle. **O que são zoonoses?** [S.l.], 2020. Disponível em: <https://www.ecycle.com.br/7902-zoonoses.html#:~:text=Zoonoses%20s%C3%A3o%20doen%C3%A7as%20infecciosas%20transmitidas%20de%20animais%20para,naturalmente%20transmiss%C3%ADveis%20entre%20animais%20vertebrados%20e%20seres%20humanos%E2%80%9D>. Acesso em: 8 nov. 2020.

BARRETO, Sérgio S. Menna. **Volumes Pulmonares**. Jornal Brasileiro de Pneumologia, v. 28, n. 3, p. S83–S94, 2002.

BASTIEN, Christian; SCAPIN, Dominique L. **Ergonomic criteria for the evaluation of human-computer interfaces**. 1993.

BIOEMFOCO. **Semelhanças e diferenças entre COVID-19 e Influenza**. [S.l.], 2020. Disponível em: <http://bioemfoco.com.br/noticia/semelhancas-e-diferencas-entre-covid-19-e-influenza/>. Acesso em: 7 nov. 2020.

BOONE, HN; BOONE, DA. **Analyzing Likert Data**. The Journal of Extension, 2012, 50, 1-5. (2012). Disponível em: <https://joe.org/joe/2012april/tt2.php>. Acesso em: 25 mai. 2021.

BORGES, Helyane Bronoski. **Redução de Dimensionalidade de Atributos em Bases de Dados de Expressão Gênica**. Curitiba, PR: 2006. 123 p. Dissertação (Mestrado) – Programa de Pós Graduação em Informática. Pontifícia Universidade Católica do Paraná, 2006.

BRASIL, Secretaria de vigilância em saúde − ministério da saúde. **Epidemiológico**. Boletim Epidemiológico, v. 47, n. 19, p. 1–9, 2016.

BRASIL. Coronavírus Brasil. **Painel Coronavírus**. [S.l.], 2020. Disponível em: <https://covid.saude.gov.br/>. Acesso em: 9 nov. 2020 2:25.

BRASIL. **Entenda a diferença entre Coronavírus, Covid-19 e Novo Coronavírus**: Os primeiros casos desse agente foram registrados na cidade de Wuhan, na China. Disponível em: <https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/03/entenda-a-diferenca-entre-coronavirus-covid-19-e-novo-coronavirus>. Acesso em: 25 out. 2020.

BREIMAN, Leo. **Random forests. Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CABENA, Peter et al. **Discovering data mining**: from concept to implementation. Prentice-Hall, Inc., 1998.

CARVALHO, Juliano Varella de. **Reconhecimento de Caracteres Manuscritos Utilizando Regras de Associação**. Campina Grande, PB: 2000. Dissertação (Mestrado) - Centro de Ciências e Tecnologia, Universidade Federal da Paraíba, 2000.

CENTERS FOR DESEASE CONTROL AND PREVENTION. **Common Human Coronaviruses**. [S.l.], 2020. Disponível em: <https://www.cdc.gov/coronavirus/downloads/Common-HCoV-fact-sheet-508.pdf>. Acesso em: 7 nov. 2020.

COSTA, D.; JAMAMI, M. **Bases fundamentais da espirometria**. Rev. bras. fisioter. v. 5, n. 2, p. 95–102, 2001.

COSTA, William Jackson da. **INCADATABR**: UMA BIBLIOTECA EM R PARA MANIPULAÇÃO DE DATASETS DO INCA. 2019. Universidade Feevale, [S.l], 2019. Disponível em: <https://tconline.feevale.br/tc/files/0001_4737.pdf>. Acesso em: 28 out. 2020.

COX, N; FUKUDA, K. **Infectious Disease Clinics of North America**: Incluenza. Science Direct, março 1998. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0891552005704062>. Acesso em: 12 out. 2020.

CUTHBERT, Lori. "National Geographic Brasil". **Como infecções como as do coronavírus passam de animais para pessoas?** [S.l.], 2020. Disponível em: <https://www.nationalgeographicbrasil.com/ciencia/2020/04/doencas-zoonoticas-zoonoses-infecao-coronavirus-animais-humanos-covid-19-ebola>. Acesso em: 7 nov. 2020.

DALE, Becky; STYLIANOU. Nassos. **Coronavírus**: como o 'excesso de mortes' pode revelar o verdadeiro número de vítimas da pandemia de covid-19. BBC News Brasil, 18 junho 2020. Disponível em: <https://www.bbc.com/portuguese/internacional-53092095> Acesso em: 30 ago. 2020.

FAPESP. **Banco de dados compartilhados** FAPESP. 2020. Disponível em: <https://repositoriodatasharingfapesp.uspdigital.usp.br/>. Acesso em: 6 mar. 2021.

FARGE, Emma. **OMS diz que pandemia de covid-19 é "uma grande onda", não é sazonal**. Agência Brasil, Genebra, julho 2020. Disponível em: <https://agenciabrasil.ebc.com.br/internacional/noticia/2020-07/oms-diz-que-pandemia-decovid-19-e-uma-grande-onda-nao-e-sazonal>. Acesso em: 26 ago. 2020.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery**: An overview. In: FAYYAD et al. Advances in Knowledge Discovery and Data Mining. G. Cambridge-Mass: AAAI/MIT Press, 1996.

FELLOWS, Ian. Deducer: **A Data Analysis GUI for R. Journal of Statistical Software**, Innsbruck, v. 49, n. 8, jun. 2012. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v049i08/v49i08.pdf>. Acesso em: 17 mai. 2021.

FORLEO-NETO, E et al. **Influenza.** In: Revista da Sociedade Brasileira de Medicina Tropical, v. 36, n. 2, p. 267-274, 2003.

GÓES, Anderson Roges Teixeira; STEINER, Maria Teresinha Arns**. O processo KDD aplicado na extração de regras**: Um estudo de caso da área médica. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2012, Rio de Janeiro. Anais... Rio de Janeiro: SOBRAPO, 2012. p. 3741.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining**: Um Guia Prático. Rio de Janeiro: Elsevier Editora Ltda., 2005.

GONÇALVES, Siumara. **Da pneumonia na China à pandemia, o caminho do coronavírus até o ES**. Jornal A Gazeta, Vitória, março 2020. Disponível em: <https://www.agazeta.com.br/es/gv/da-pneumonia-na-china-a-pandemia-o-caminho-docoronavirus-ate-o-es-0320>. Acesso em: 26 ago. 2020.

GOULART, Flavio A. de Andrade. **Doenças crônicas não transmissíveis**: estratégias de controle e desafios e para os sistemas de saúde. Organização Pan-Americana da Saúde/Organização Mundial da Saúde, p. 96, 2011.

GROSSI, Gustavo. **O que é Creative Commons?** Saiba tudo sobre a licença autoral mais famosa para conteúdo web! Comunidade Rockcontent. [S.l.], 2017. Disponível em: <https://comunidade.rockcontent.com/o-que-e-creative-commons/>. Acesso em: 7 mar. 2021.

GRUBER, Arthur. **A origem do Sars-CoV-2**. Pfarma, [S.l.], 16 abr. 2020. Disponível em: <https://pfarma.com.br/coronavirus/5439-origem-covid19.html>. Acesso em: 25 out. 2020.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: Concepts and Techniques. 3. ed. Waltham: Elsevier, 2012.

HECKLER, Wesllei Felipe. **Análise preditiva sobre pacientes do "projeto de extensão reabilitação pulmonar"** da universidade feevale. 2018. Universidade Feevale, [S.l.], 2018. Disponível em: <https://tconline.feevale.br/NOVO/tc/files/0001_4637.pdf> Acesso em: 10 out. 2020.

HEIN, Alexandra**. WHO declares coronavirus global 'pandemic'**. Fox News, Nova Iorque, 11 março 2020. Disponível em: <https://www.foxnews.com/health/who-declares-coronavirusglobal-pandemic>. Acesso em: 2 set. 2020.

HORTA, Luis. **Influenza**. In: Sinopse pediátrica. [S.l.], 2016. Disponível em: <http://sinopsepediatrica.blogspot.com/2016/04/influenza.html>. Acesso em: 13 out. 2020.

INSTITUTO INFORMATION MANAGEMENT. **Menos papel e mais árvores**: digitalização é uma das principais medidas para a proteção do meio ambiente. 5 de Junho de 2020. Disponível em: <https://docmanagement.com.br/06/05/2020/menos-papel-e-mais-arvores-digitalizacao-e-uma-das-principais-medidas-para-a-protecao-do-meio-ambiente/>. Acesso em: 23 abr. 2021.

IULIANO et al. **Estimates of global seasonal influenza-associated respiratory mortality**: a modelling study. The Lancet, dezembro 2017. Disponível em: <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)33293-2/fulltext> Acesso em: 5 ago. 2020.

LIMA, Rodrigo Ramos. Fiocruz. Especial Covid-19 | **A Covid-19 e a relação entre humanos e animais**: zoonoses e zooterapias. [S.l.], 2020. Disponível em:

<http://coc.fiocruz.br/index.php/pt/todas-as-noticias/1816-especial-covid-19-a-covid-19-e-a-relacao-entre-humanos-e-animais-zoonoses-e-zooterapias.html>. Acesso em: 8 nov. 2020.

LIU, Bing. **Web Data Mining**: Exploring Hyperlinks, Contents and Usage Data. 2. ed. Chicago: Springer, 2011.

MAIMON, O.; ROKACH, L. **Data mining and knowledge discovery handbook**. 2. ed. Springer, janeiro 2010.

NICI, Linda et al. **American Thoracic Society/European Respiratory Society Statement on Pulmonary Rehabilitation**. American Journal of Respiratory and Critical Care Medicine, v. 173, n. 1, p. 1390–1413, 2006.

NIELSEN, Jakob. **Projetando websites**. Gulf Professional Publishing, 2000.

NIELSEN, Jakob; TAHIR, Marie. **Homepage usability**: 50 enttarnte Websites. Pearson Deutschland GmbH, 2004.

PARIZOTTO, Rosamelia et al. **Elaboração de um Guia de Estilos para Serviços de Informação em Ciência e Tecnologia via Web.** 1997.

PELLEGRINO, R et al. **Interpretative strategies for lung function tests**. European Respiratory Journal, v. 26, n. 5, p. 948–968, 2005.

PÚBLICO. Covid-19: **A pandemia abalou sistemas de saúde de todo o mundo, diz a OMS**. Portugal, Lisboa: Coronavírus, 2020. Disponível em: <https://www.publico.pt/2020/08/31/mundo/noticia/covid19-pandemia-abalou-sistemassaude-mundo-oms-1929833>. Acesso em: 2 set. 2020.

PUJARI, Arun K. **Data mining techniques**. Universities press, 2001.

REVISTA DA CIDADE. **Gripe e resfriado podem ser ameaças para pessoas com doenças crônicas**. Aracaju. 2018. Disponível em: <http://www.jornaldacidade.net/saude/2018/07/301761/gripe-e-resfriado-podem-ser-ameacas-para-pessoas-com-doencas.html>. Acesso em: 12 out. 2020.

REVISTA ISTOÉ. Pandemia já matou mais de 3,7 milhões de pessoas no mundo. [S.l.]. 2021. Disponível em: <https://www.istoedinheiro.com.br/pandemia-ja-matou-mais-de-37-milhoes-de-pessoas-no-mundo>. Acesso em: 2 jul. 2021.

REZENDE, Solange Oliveira. **Mineração de Dados**. In: XXV Congresso da Sociedade Brasileira de Computação, 2005. Anais do XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo: SBC, 2005. p. 397-433.

RODRIGUES, Joaquim Carlos; CARDIERI, Joselina M. Andrade; BUSSAMRA, Maria Helena Carvalho de Ferreira; NAKAIE, Cleyde Myriam Aversa; ALMEIDA, Marina Buarque de; FILHO, Luiz Vicente Ferreira da Silva; ADDE, Fabíola Villac. **Provas de função pulmonar em crianças e adolescentes**. Jornal Pneumologia, v. 28, n. Supl 3, p. S 207-S 221, 2002.

SLAYER, Stephanie J. et al. **Prioritizing Zoonoses for Global Health Capacity Building** - Themes from One Health Zoonotic Disease Workshops in 7 Countries, 2014–2016. [S.l.], dez. 2017.

SOCIEDADE PAULISTA DE PNEUMOLOGIA E TISIOLOGIA. **Pneumologia:** atualização e reciclagem. 7. ed. São Paulo: Roca, 2008.

SPERANDIO, Evandro Fornias et al. **Distúrbio ventilatório restritivo sugerido por espirometria**: associação com risco cardiovascular e nível de atividade física em adultos assintomáticos. Jornal Brasileiro Pneumologia, v. 42, n. 1, p. 22–28, 2016.

STEINER; Maria Teresinha Arns et al. **Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados**. Gestão & produção, v. 13, n.2, p.325-337, 2006.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining**. Rio de Janeiro: Editora Ciência Moderna, 2009.

VIDALE, Giulia. **44% dos brasileiros sofrem com problemas respiratórios**. Revista Veja, São Paulo, agosto 2015. Saúde. Disponível em: <https://veja.abril.com.br/saude/44-dos-brasileiros-sofrem-com-problemas-respiratorios/>. Acesso em: 1 ago. 2020.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining**: Practical Machine Learning Tools and Techniques. 2. ed. Burlington: Elsevier Inc., 2005.

APÊNDICE A – QUESTIONÁRIO DE VALIDAÇÃO DO *DASHBOARD*

O questionário para a validação do *dashboard* apresenta três perguntas iniciais referentes à dados pessoais do avaliado, as perguntas subsequentes têm sua recomendação baseada em experiências práticas de vários pesquisadores de usabilidade, como Nielsen (2000, 2004), Bastien e Scapin (1993) e, tomando como base as “regras de ouro” para projetos de interfaces de Bem Shneiderman, conforme descrito por Agni (2015), além das guias de estilos de Parizotto (1997), ao final do questionário há um campo descritivo opcional para que o avaliado possa contribuir com críticas e sugestões.

Relacionado às perguntas apresentadas no questionário, as três primeiras questionam sobre a área de atuação, a profissão e o tempo de experiência na área. As perguntas subsequentes são de múltipla escolha, onde os usuários podem optar por respostas como:

- Discordo totalmente
- Discordo
- Neutro
- Concordo
- Concordo totalmente

A seguir, são descritas as perguntas que foram respondidas pelos usuários.

1. O tempo para carregar a página não interfere no meu trabalho.
2. Em caso de demora, é informado o tempo restante para concluir a tarefa.
3. Os links e menus são facilmente encontrados.
4. É fácil encontrar a informação desejada no sistema.
5. O sistema oferece um layout que permite se localizar facilmente.
6. É oferecida versão imprimível das informações.
7. O texto descrito é legível.
8. Os gráficos me permitem entender as informações apresentadas.
9. O título de cada menu ou gráfico indica corretamente o que existe nele.
10. O material usa uma linguagem simples e objetiva.
11. O sistema é fácil de usar, não sendo necessário recorrer à ajuda para encontrar as informações.
12. Textos se adequam as seleções realizadas.
13. As cores utilizadas favorecem o uso do sistema.
14. O sistema é apresentado em uma única fonte.

15. Existe contraste entre o texto, os gráficos e o fundo.
16. Não ocorrem erros durante o uso do software.
17. Consegui encontrar a informação que procurava no sistema.
18. As informações disponíveis no *dashboard* são suficientes para uma avaliação de cenários.
19. A exibição dos dados proporciona uma fácil compreensão.
20. O *dashboard* proporciona facilidade na detecção de padrões entre os pacientes.
21. Os filtros oferecidos são de fácil uso.
22. A ferramenta pode ser usada de forma estratégica para a tomada de decisão sobre os casos de covid.
23. A ferramenta pode ser usada para identificar grupos de pacientes que tendem a contrair a doença, com base nas características dos filtros.

Por fim, foi adicionado um campo descrito como “Sugestões e críticas”, onde os respondentes podem contribuir com o questionário e o *dashboard*.

APÊNDICE B – CONFIGURAÇÃO E INSTALAÇÃO DO SISTEMA NA AMAZON

Com o servidor selecionado e em funcionamento, foi utilizada a própria interface do Amazon EC2 para acessar o terminal de controle do servidor e, então, iniciar as instalações necessárias para o ambiente suportar a aplicação. Antes de iniciar a instalação dos pacotes necessários, foi executado o comando "sudo apt-get update" para atualizar o repositório do ubuntu, e, em seguida, as duas primeiras instalações foram do R e shinydashboard. Eles foram instalados com os comandos "sudo apt-get install r-base" e "sudo apt-get install r-cran-shinydashboard".

Com o R e shinydashboard instalados, o próximo pacote instalado foi o shiny-server, responsável por disponibilizar a aplicação para a web. Para a instalação, foram executados a sequência de comandos "sudo apt-get install gdebi-core", em seguida, "sudo wget http://download3.rstudio.org/ubuntu-12.04/x86_64/shiny-server-1.3.0.403-amd64.deb", e, por fim, "sudo gdebi shiny-server-1.3.0.403-amd64.deb". Finalizada a instalação dos pacotes principais, os próximos pacotes a serem instalados estão descritos em ordem no Quadro 12.

Quadro 12 – Lista de comandos para instalação de pacotes no Ubuntu

Pacote	Comando
www-browser	**sudo apt-get install www-browser**
links2	**sudo apt-get install links2**
elinks	**sudo apt-get install elinks**
links	**sudo apt-get install links**
lynx	**sudo apt-get install lynx**
w3m	**sudo apt-get install w3m**

Fonte: Elaborado pelo autor

Para transferir a aplicação desenvolvida localmente para o servidor, foi utilizado o comando: "scp /d/CoronaViewer/CoronaViewer.zip ubuntu@ec2-52-15-175-87.us-east-2.compute.amazonaws.com:/home/ubuntu/CoronaViewer.zip"

Sobre o funcionamento do comando acima, o "scp" (*Secure Copy*) envia um ou mais arquivos de uma origem até um destino. Neste caso, a origem foi o computador onde a aplicação foi desenvolvida. Para simplificar o envio, foi criado um arquivo no formato zip com todos os dados, incluindo as bases, enviado ao *link* de destino oferecido pelo serviço da Amazon. Ao final deste *link*, foi adicionado o comando ":/home/ubuntu", que se refere à pasta no servidor que receberá a cópia do arquivo.

Tendo finalizada a transferência do arquivo, na pasta “/home/ubuntu”, é executado o comando “unzip CoronaViewer.zip” para descompactar a aplicação. Na sequência, na pasta “/srv/shiny-server/” executa-se dois comandos: “mv /home/ubuntu/CoronaViewer.” e “chmod -R 777 CoronaViewer”. O primeiro move na pasta já descompactada para a pasta de projetos do servidor, já o segundo dá permissão de leitura e escrita para todos os usuários do sistema em todas as pastas, subpastas e arquivos dentro do projeto CoronaViewer.

O último passo, antes de subir a aplicação, é editar o arquivo “shiny-server.conf” que está disponível na pasta “/etc/shiny-server/”. Para editar o arquivo, é utilizado o comando “sudo nano /etc/shiny-server/shiny-server.conf”. A única alteração necessária é na linha 5 do arquivo, em que foi substituído o valor “listen 80” por “listen 3838”. Tendo feito a substituição, é necessário salvar as alterações, pressionando o comando “Ctrl X”, em seguida digitar a letra “y” para confirmar ou “n” para negar, e por fim “Enter”.

Depois destes passos, é possível subir a aplicação. Ao entrar na pasta “/srv/shiny-server”, é executado o comando “R”, que abre um painel permitindo interagir diretamente com a linguagem R. Dentro deste painel, executou-se o comando “shiny::runApp(‘CoronaViewer’)”. Como passo final, pode ser verificado o funcionamento da aplicação digitando o endereço do link fornecido pela Amazon, juntamente com o trecho “:3838/CoronaViewer”. Para o caso deste projeto, o *link* de acesso pode ser identificado como: http://ec2-52-15-175-87.us-east-2.compute.amazonaws.com:3838/CoronaViewer/.