Trabalho de Sistemas de Apoio à Decisão
Business Intelligence

Celso Antonio Uliana Junior[1], Felipe Salles Lopes[1], Lucas Avanzi[1]

[1]Faculdade de Computação (FACOM) – Universidade Federal do Mato Grosso do Sul (UFMS) - Campo Grande - MS - Brazil

celso.a.u@gmail.com, felipe28101997@gmail.com, lucasavanzi97@gmail.com

Abstract. *This paper provides a data analysis over the subject of coronavirus dissemination which its goal is to support decisions for a federal level manager. This paper was not meant to provide a complex solution in order to stop or refrain the coronavirus dissemination, but rather try to identify the main parameters that contribute to the spread of the virus and, therefore, support the decision-makers to make the right call that could mitigate the disease advances. We present analyses for three issues: death toll arising in consequence of the hospital beds maximum occupancy, the correlation between the isolation rate and the number of daily cases in Brazil, and the correlation between the temperature and cases daily in big cities around the world. The content of this paper is divided into four sections: introduction, methods and materials, results, and the made analyses discussions, and conclusion.*

Resumo. *Este artigo fornece uma análise de dados sobre o assunto da disseminação do coronavírus com o objetivo de apoiar decisões para um gestor de nível federal. Este artigo não é feito para fornecer uma solução complexa para interromper ou frear a disseminação do coronavírus, mas sim, para tentar identificar os principais parâmetros que contribuem para o aumento do contágio da doença e, assim, apoiar os gestores a tomar decisões que possam mitigar o avanço da doença. Nós exibimos análises para três questões: aumento de mortes por conta da lotação de leitos hospitalares, correlação entre a taxa de isolamento e o número de contágio diário no Brasil, e correlação entre a temperatura e o contágio diário em grandes cidades no mundo. O conteúdo deste artigo está dividido em quatro capítulos: introdução, materiais e métodos utilizados, resultados e discussão das análises realizadas, e conclusão.*

1. Introdução

Desde o começo do ano, o mundo sofre com uma grande epidemia, que posteriormente se tornou uma pandemia, de uma doença chamada *COVID-19*. Esta doença é causada pelo coronavírus *SARS-CoV-2* e pode apresentar desde sintomas menos graves e mais comuns como tosse seca, febre e cansaço; assim como sintomas mais graves, porém mais raras: dificuldade de respirar ou falta de ar, e dor ou pressão no peito, além de poder ter casos assintomáticos. Apesar de não ser uma doença tão letal, a velocidade de contágio é muito rápida; em 2 meses a doença já contagiou mais de 40 mil pessoas enquanto o vírus *Sars*, que tem a mesma origem da atual pandemia (China), demorou 10 meses para infectar 8 mil pessoas[1].

O objetivo desse artigo é fornecer informações extraídas com base em dados coletados sobre o coronavírus utilizando aprendizado de máquina para que tais informações possam ser úteis no apoio a decisão de gestores ao nível federal e, logo, estes tomarem medidas para combater esse rápido avanço do vírus. Neste artigo, três questões para análise foram determinadas: aumento de mortes por conta da lotação de leitos hospitalares, correlação entre a taxa de isolamento e o número de contágio diário no Brasil, e a correlação entre a temperatura e contágio diário em grandes cidades ao redor do mundo.

Na primeira questão, é realizada uma breve análise sobre os dados coletados a respeito dos leitos hospitalares, número de casos e mortes da doença COVID-19 em quatro países selecionados, são estes: Brasil, Espanha, Estados Unidos e Itália. Tais países foram selecionados pelo motivo de serem um dos mais afetados em todo o mundo. Em seguida, na segunda questão, utilizamos o aprendizado de máquina para realizar a correlação entre as séries temporais de número de casos diários da doença e a taxa de isolamento, estudar e analisa-la brevemente e utilizar ambos dados resultantes para treinar um modelo e investigar se este é melhor que o treinado com apenas a quantidade de casos. Destacamos o uso dos algoritmos de regressão linear, redes neurais e *SVR* (*Epsilon-Support Vector Regression*) para esta questão. E por último, na terceira questão, comparamos a trajetória de casos com a temperatura de nove cidades selecionadas com o intuito de checar se esses parâmetros têm relação. Para cada questão, exibimos os materiais e métodos usados, assim como discutimos as respostas obtidas.

2. Materiais e métodos

Neste capítulo, são apresentados as ferramentas e algoritmos utilizados, assim como as fontes dos dados obtidos e a maneira na qual foram obtidos e transformados para a sua devida utilização. As ferramentas utilizadas são apresentadas na seção 2.1, onde serão brevemente descritas o propósito de suas utilizações. Em 2.2, o conjunto de dados utilizado para a realização de experimentos e análises é apresentado. Por fim, na seção 2.3 será descrito os experimentos e análises realizados nos conjuntos de dados.

2.1. Ferramentas

Para este trabalho foram utilizadas diversas ferramentas de diversos propósitos para auxiliar na realização do mesmo. Primeiramente foi utilizado a ferramenta *Pentaho Data Information* para a transformação de certos conjuntos de dados. No entanto, essa ferramenta só foi utilizada no começo do trabalho, já que os conjuntos de dados não possuíam complexidade suficiente para o uso desta ferramenta, portanto esta foi descontinuado ao longo do trabalho.

Para a escrita de códigos, Python, Jupyter ou SQL, foram utilizadas quatro ferramentas distintas: *PyCharm*, *Sublime Text 3*, *pgAdmin 4* e *Jupyter Notebook*. O *PyCharm* foi utilizado para a escrita de códigos Python; o *Sublime Text 3* foi utilizado

para escrever códigos em geral, principalmente SQL; o *pgAdmin 4* foi empregado para a criação do banco de dados, assim como a sua estrutura e DMLs (linguagem de manipulação), em geral; por último, o *Jupyter Notebook* foi utilizado para testes de códigos Python.

Mais especificamente, para o aprendizado de máquina, foi utilizado as bibliotecas *scikit-learn* ou *sklearn* de onde foram importados os algoritmos de aprendizagem, avaliadores de modelo e as métricas. Também foi utilizado as bibliotecas *Numpy*, *Pandas*, *Psycopg2* e *Matplotlib* para manipulação e apresentação dos dados consumidos e gerados.

Para o armazenamento e realização de consulta dos dados, foi utilizado o SGBD (Sistema de gerenciamento de banco de dados) *PostgreSQL*. Nesse SGBD foi modelado o nosso *DataWarehouse* (Figura 1), que nos possibilita uma busca multidimensional no banco.

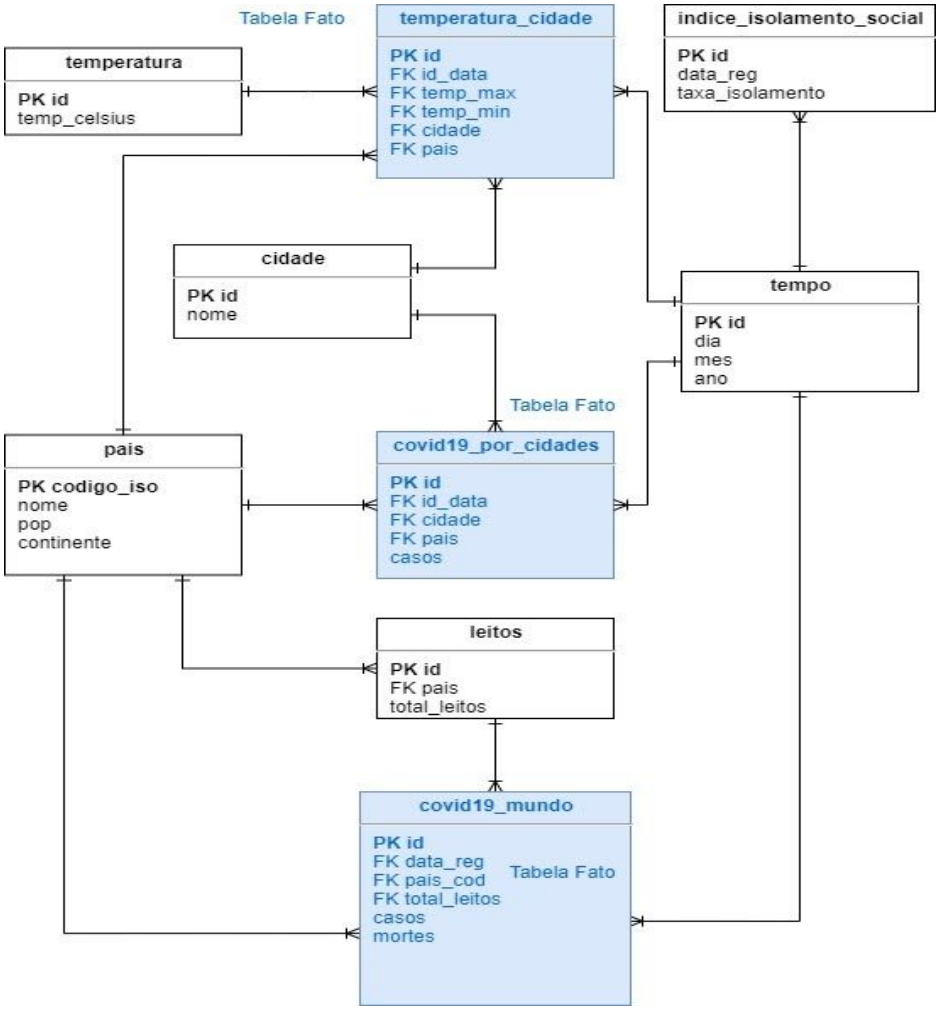

Figura 1. Modelo *DataWarehouse*

Para o controle de versão do trabalho/código, foi utilizado as ferramentas *GIT Bash*, que permite a inserção de comandos *GITs* em um terminal separado do *cmd* do Windows, e *GitKraken*, que foi utilizado apenas para auxiliar no controle de versão.

Finalmente, foi utilizado para a criação e edição de documentos o Google Docs e Microsoft Word, e para visualização de arquivos CSV ou XLSX o Excel.

2.2. Conjunto de Dados

Foram coletados dados de diversas fontes para cada uma das questões levantadas para a análise. Na primeira questão, os dados obtidos em relação aos casos e mortes de pessoas em decorrência do COVID-19 são oriundos do portal de dados aberto da União Europeia (*EU Open Data Portal*)[2], ao passo que os dados obtidos relacionados aos leitos hospitalares foram retirados de diversas fontes. Para a segunda questão, usamos o mesmo material relacionado ao coronavírus, mas, em acréscimo a isso, coletamos os dados da *Inloco*[3], uma empresa de segurança de localização, relacionados à taxa de isolamento da população brasileira. Por último, na terceira pergunta, buscamos inúmeras fontes para obter materiais necessários para cada cidade que escolhemos para a análise dos casos de contágio da doença, em contrapartida, coletamos o material para as temperaturas de cada cidade diariamente no site de meteorologia *AccuWeather*[4] (com exceção de Camberra, na qual foram coletadas as temperaturas no site do governo da Austrália[5]).

Todos os materiais obtidos, tantos os brutos quanto os transformados, foram armazenados em um repositório central Github[6], assim como todo o material resultante deste trabalho. Todos os dados obtidos foram injetados no *DataWarehouse* e os códigos SQL para criação e inserção também estão disponíveis no repositório.

2.3. Experimento e análises

A partir dos dados obtidos e transformados descritos na seção 2.2, foram escolhidos os algoritmos regressão linear, SVR e redes neurais. A regressão linear foi escolhido por ser um modelo bem simples, por outro lado, os modelos SVR e redes neurais foram escolhidos por serem bastante abordados nas aulas da disciplina. Fizemos uso da janela deslizante para todos os algoritmos.

O modelo regressão linear ajusta um modelo linear para minimizar a soma residual quadrática entre os alvos observados no conjunto de dados e alvos previstos pela aproximação linear. Os parâmetros para esse modelo foram todos os padrões.

O modelo redes neurais *MLPRegressor* (*Multilayer Perceptron Regressor)* utiliza *backpropagation* para atualização de pesos e foi utilizado como parâmetros o número de camadas ocultas, alfa, taxa de aprendizado inicial, número máximo de iterações, estado aleatório e tolerância para otimização além dos parâmetros padrões do modelo.

O modelo *SVR* é responsável por dividir diferentes dados no hiperplano e foi utilizado como parâmetros, tipo do 'kernel' do algoritmo, parâmetro de regularização, épsilon e número máximo de iterações além dos parâmetros padrões do modelo.

Já o método de validação cruzada utilizado pelos algoritmos de aprendizado de máquina foi o *Leave One Out* que basicamente consiste em um *K Folds* com valor k=n, é um método de validação custoso e foi escolhido devido à baixa quantidade de dados no dataset.

Foi utilizado as métricas *MSE (Mean Squared error)*, R^2 *(R-Squared)* e *adjusted* R^2 para avaliação dos modelos utilizados neste trabalho.

3. Resultados e discussão

3.1. Primeiro objetivo

O primeiro objetivo consiste em montar gráficos gerados a partir dos dados coletados, e analisa-los utilizando os vetores de mortes acumuladas, casos acumulados e a capacidade máxima de leitos disponível por país. O objetivo é comprovar ou não a hipótese de que a capacidade de leitos influência no número de mortes em decorrência da doença.

Observando o gráfico, pode-se perceber que o Brasil tem uma quantidade de leitos muito inferior comparado aos outros países escolhidos, incluindo aqueles que têm população inferior ao Brasil. Também que é possível observar que em todos os países selecionados, quando o número de pessoas contaminadas ultrapassa a capacidade máxima de leitos hospitalares, o número de mortes começa a crescer, em alguns gráficos acentuadamente, e em outros a elevação é mais discreta (Figura 2). É importante notar que o gráfico do Brasil está em escala logarítmica.

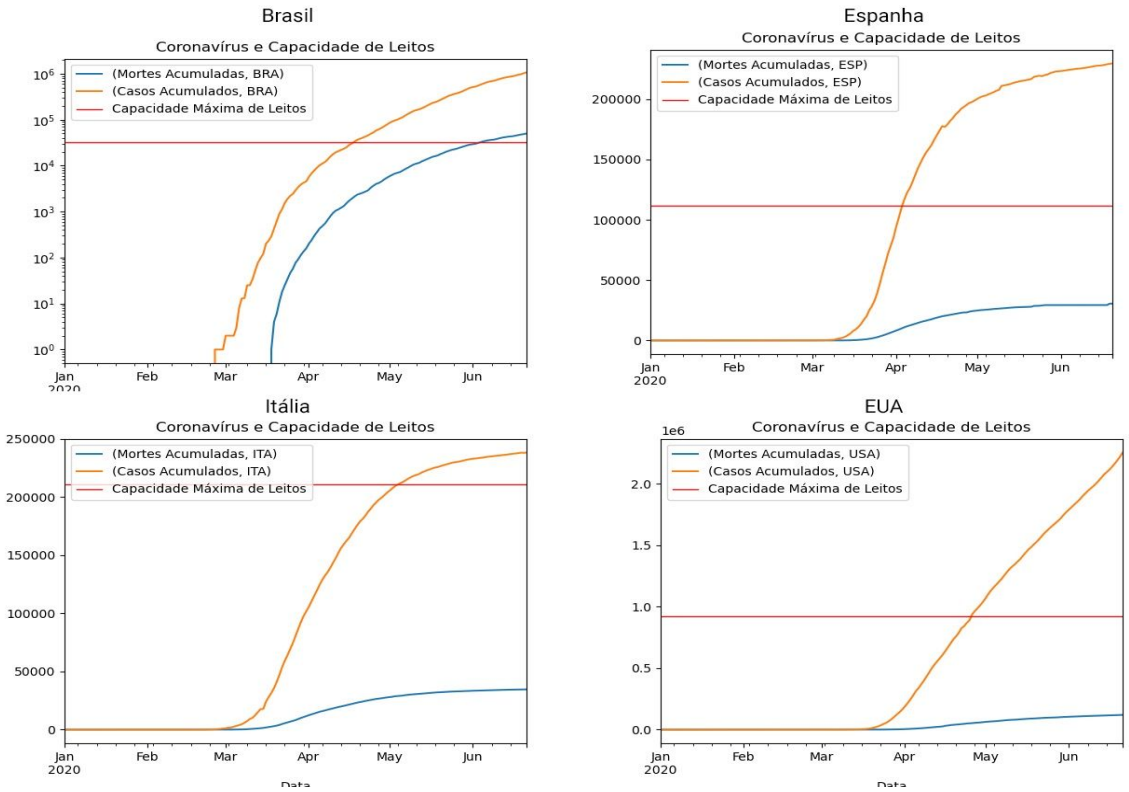

Figura 2. Gráficos coronavírus e leitos por país

No entanto, apesar de todas as circunstâncias apontarem para a comprovação de nossa hipótese, deve-se atentar ao fato que os leitos hospitalares não são exclusivos para o COVID-19 - o que até contribui para a hipótese considerando que teria ainda menos espaço exclusivo para portadores da doença -, além do fato que nem todas as pessoas contagiadas precisam de leitos. Contudo, concluímos que, por meio desses dados, é muito provável que a lotação dos leitos hospitalares tenham influenciado no aumento de mortes acumuladas.

3.2. Segundo objetivo

O segundo objetivo consiste em utilizar a aprendizagem de máquina para realizar a correlação entre duas séries temporais: o número de casos diários do coronavírus e a taxa de isolamento da população (Figura 3). Em seguida, utilizar para analisar a correlação brevemente para aplicar a informação resultante no treinamento de um modelo e, então, investigar se este é melhor que o treinado com apenas o número de pessoas contagiadas.

Foi decidido usar a correlação de *Pearson* como uma métrica de correlação entre as duas séries. Ao utilizar todos os dados disponíveis no *DataWarehouse*, obtemos a

correlação de Pearson no valor aproximado de -0,0389 o que indica, mesmo que pouco, que as séries estão relacionadas inversamente.

Ao analisar novamente a correlação de *Pearson* dessa vez ignorando os 25 primeiros dias da série temporal, isto é, ignorando o crescimento inicial do gráfico tendo em vista a natureza exponencial da série temporal de casos diários de COVID-19 e da baixa taxa de isolamento inicial pelo desconhecimento ou despreocupação da população, temos um valor por volta de -0,55 que indica uma correlação aparentemente maior que a do valor da série inteira.

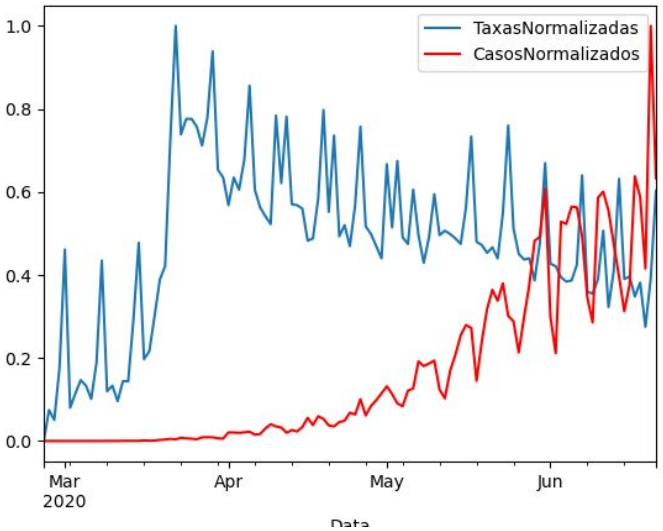

Figura 3. Série temporal de taxa de isolamento social normalizado

Utilizamos três modelos nos quais foram obtidos métricas para analisar se ao adicionar a taxa de isolamento social normalizada na janela deslizante do aprendizado, seria obtido um modelo mais preciso no qual descreve melhor o problema.

Todas essas as métricas citadas foram adquiridas junto de um gráfico que representa vários modelos e seus alvos (Y) preditos com o alvo (*Y*) verdadeiro, posteriormente, guardamos essas métricas em um arquivo. Analisando esse arquivo, é possível observar que, em geral, os resultados foram similares, exceto o *MLPRegressor* que performou melhor na análise que utiliza a taxa de isolamento na grande maioria dos casos, e que a regressão linear começa a performar pior com um tamanho de janela grande, o que geralmente para todos os modelos um tamanho de janela pequeno é recomendado. Concluímos que com um conjunto de dados maior, e possivelmente com a utilização de outras séries temporais que têm relação com a de número de casos, seria possível melhorar o modelo.

3.3. Terceiro objetivo

Finalmente, no terceiro objetivo comparamos a trajetória de casos de contágio diários com a temperatura de nove cidades selecionadas. As cidades selecionadas foram:

Buenos Aires, Camberra, Los Angeles, Madri, Miami, Nova Iorque, Rio de Janeiro, Roma e São Paulo. Foram selecionadas tanto cidades de climas mais frios, quanto cidades de climas mais quentes, porém que já tivesse com casos de coronavírus por um tempo considerável. A hipótese aqui expõe que uma temperatura mais quente atrapalha na disseminação do vírus. Para testar esta hipótese, construímos gráficos com os dados de contágio e de temperatura, e mudamos a escala dos casos para a melhor visualização do gráfico.

Ao obter o gráfico de cada cidade, não foi possível observar nenhuma relação clara entre esses dois parâmetros. Além disso, com o clima, muda-se os hábitos das pessoas. Uma de nossas premissas para essa hipótese se diz respeito do fato das pessoas se higienizarem menos no frio, o que aumentaria o contágio, contudo não achamos evidência para essa premissa. Outra premissa corresponde ao aumento de número de resfriados por conta do frio, porém, novamente, não pudemos comprovar. Possivelmente, se usássemos mais parâmetros, como a umidade, quiçá poderíamos ver algo mais evidente. Portanto, concluímos que não há relação aparente com base nesses dados entre a temperatura e a disseminação do vírus.

4. Conclusão

Sendo esse assunto um bem complexo, sabemos que diversos outros fatores influenciam na curva de crescimento do vírus, como, por exemplo: natureza exponencial da curva, número de leitos, distanciamento social e muitos outros. Outros dados podem também ter grande relação com o estudo do COVID-19 que não foram utilizados neste artigo, mas possivelmente tem alguma relação com os dados do vírus mesmo que pequena, citando alguns desses possíveis dados: dados econômicos do país (renda mínima e salário médio), dados de educação de uma região (número de pessoas alfabetizadas, taxa da população que tem ensino superior, etc) e talvez até índice de desenvolvimento humano. Sendo assim, uma análise nesses dados é um trabalho complexo e robusto, que necessitaria uma infraestrutura e métodos muito mais perto do estado da arte de *DataWarehouse* e *Machine Learning,* mas como trabalho final da disciplina de SIAD estamos contentes nos resultados obtidos, no conhecimento e experiência adquirido. Acreditamos que conseguimos explorar bastante o assunto e a área de *Machine Learning*, assim como o aprendizado novas ferramentas e uma maior noção de um assunto complexo e atual.

Referências

[1] BIOLAB ANÁLISES CLÍNICAS (Brasil). Coronavírus: por que ele se espalhou muito mais rápido que o vírus da SARS?. São Paulo, SP, 9 fev. 2020. Disponível em: http://laboratoriosobrinho.com.br/site/conteudo/322-coronavirus-por-que-ele-se-espalhou-muito-mais.html. Acesso em: 12 jul. 2020.

[2] UNIÃO EUROPEIA (União Europeia). European Centre for Disease Prevention and Control (org.). COVID-19 coronavirus data: Covid-19 cases worldwide. Solna, Suécia, 31 dez. 2019. Disponível em: http://data.europa.eu/88u/dataset/covid-19-coronavirus-data. Acesso em: 29 abr. 2020.

[3] INLOCO (Brasil). Índice de isolamento social. [S. l.], 2 abr. 2020. Disponível em: https://www.inloco.com.br/covid-19. Acesso em: 19 jun. 2020.

[4]ACCU WEATHER (USA). ACCU Weather. Weather Forecast. United States of America, 1 jan. 2020. Disponível em: https://www.accuweather.com/. Acesso em: 28 abr. 2020.

[5]AUSTRALIAN GOVERNMENT (Canberra). Bureau of Meteorology. Canberra, Australian Capital Territory January 2020 Daily Weather Observations. Canberra, ACT, 1 jan. 2020. Disponível em: http://www.bom.gov.au/climate/dwo/202001/html/IDCJDW2801.202001.shtml. Acesso em: 28 abr. 2020.

[6]LOPES, Felipe Salles Lopes. Trab-siad. Campo Grande, MS, 17 maio 2020. Disponível em: https://github.com/El-Raptor/trab-siad/. Acesso em: 17 maio 2020.

MINISTÉRIO DA SAÚDE (Brasil). Governo do Brasil (org.). Coronavírus // Brasil. Brasil: Governo do Brasil, 19 mar. 2020. Disponível em: https://covid.saude.gov.br/. Acesso em: 27 abr. 2020.

CENTRAL INTELLIGENCE AGENCY (USA). Central Intelligence Agency (org.). The World Factbook: Hospital Beds Density. USA, 1 jan. 2019. Disponível em: https://www.cia.gov/library/publications/resources/the-world-factbook/fields/360.html. Acesso em: 6 maio 2020.

GOBIERNO DE ESPAÑA (España). Gobierno de España (org.). Sanidad en datos: Atención sanitaria. Madrid, ESPAÑA, 31 dez. 2019. Disponível em: https://www.mscbs.gob.es/estadEstudios/sanidadDatos/home.htm. Acesso em: 7 maio 2020.

AMERICAN HOSPITAL ASSOCIATION (USA) (org.). Fast facts on U.S. hospitals. [*S. l.*], 1 jan. 2020. Disponível em: https://www.aha.org/statistics/fast-facts-us-hospitals. Acesso em: 6 maio 2020.

JUSTEN, Álvaro Justen. COVID-19. Curitiba, BRA, 24 mar. 2020. https://twitter.com/turicas. Disponível em: https://brasil.io/dataset/covid19/caso_full/. Acesso em: 28 abr. 2020.

DATADISTA (Espanha). DATADISTA (org.). Datasets: COVID-19. [*S. l.*], 28 abr. 2020. Disponível em: ,. Acesso em: 28 abr. 2020.

NEW YORK TIMES (New York City). New York Times (org.). Covid-19 data. New York City, NY, 21 jan. 2020. Disponível em: https://github.com/nytimes/covid-19-data. Acesso em: 28 abr. 2020.

DIPARTIMENTO DELLA PROTEZIONE CIVILE (Roma). Presidenza del Consiglio dei Ministri (org.). COVID-19 Italia - Monitoraggio situazione. Italia, 18 fev. 2020. Disponível em: https://github.com/pcm-dpc/COVID-19. Acesso em: 28 abr. 2020.

WIKIPEDIA (Argentina) (org.). COVID-19 pandemic data/Argentina medical cases. Buenos Aires, ARGENTINA, 12 mar. 2020. Disponível em: https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/Argentina_medical_cases. Acesso em: 28 abr. 2020.