Revista Brasileira de Administração Científica
Brazilian Journal of Scientific Administration

Jul a Set 2022 - v.13 - n.3

ISSN: **2179-684X**

This article is also available online at:
www.sustenere.co

Estimação da frequência de utilização de serviços de saúde durante a pandemia utilizando machine learning

A mineração de dados é uma ferramenta altamente difundida desde o surgimento da computação, realizando a coleta, filtragem, processamento, análise e obtenção de informações relevantes em bases de dados complexas. Dentro da gama de aplicações do data mining destaca-se o machine learning, ou seja, algoritmos projetados em máquinas para que elas aprendam a trabalhar com diversos dados de maneira eficaz. Técnicas de machine learning vêm sendo muito utilizadas para a análise de bases de dados de diversas naturezas, com destaque especial para aquelas geradas com dados colhidos a partir do advento da pandemia do novo coronavírus. Neste artigo, foram utilizadas três técnicas para estimar o número de visitações em serviços de saúde em países de toda a América, e mais tarde comparando cada um deles com o Brasil. As três técnicas utilizadas foram os algoritmos de floresta aleatória, redes neurais e k-vizinhos mais próximos (ou k-nearest neighbors - KNN). A partir da aplicação dos três algoritmos, analisou-se então o erro absoluto médio, o erro quadrático médio, e a raiz do erro quadrático médio para fins de comparação. Com base nos resultados obtidos, foi observado que o país que possuía a população mais distinta da população brasileira no assunto estudado são os Estados Unidos da América, enquanto os dois países mais parecidos são Uruguai e Honduras.

Palavras-chave: COVID-19; Saúde; Regressão; Machine learning.

Estimating the frequency of use of health services during the pandemic using machine learning

Data mining has been a highly widespread tool since the emergence of computing, performing the collection, filtering, processing, analysis and obtaining relevant information in complex databases. Within the range of data mining applications, machine learning stands out, that is, algorithms designed in machines so that they "learn" to work with different data effectively. Machine learning techniques have been widely used for the analysis of databases of different natures, with special emphasis on those generated with data collected from the advent of the new coronavirus pandemic. In this article, three techniques were used to estimate the number of visits to health services in countries across the Americas, and later comparing each of them with Brazil. The three techniques used were random forest algorithms, neural networks and k-nearest neighbors (KNN). From the application of the three algorithms, the mean absolute error, the mean square error, and the root mean square error were then analyzed for comparison purposes. Based on the results obtained, it was observed that the country that had the most distinct population from the Brazilian population in the subject studied is the United States of America, while the two most similar countries are Uruguay and Honduras.

Keywords: COVID-19; Health; Regression; Machine Learning.

Topic: **Pesquisa Operacional**

Reviewed anonymously in the process of blind peer.

Received: **10/07/2022**
Approved: **25/09/2022**

Ana Carolina Piccinini de Alencar Schiavi
Universidade Tecnológica Federal do Paraná, Brasil
http://lattes.cnpq.br/3533970252479548
anaschiavi@alunos.utfpr.edu.br

Arthur Kreling Ozorio
Universidade Tecnológica Federal do Paraná, Brasil
http://lattes.cnpq.br/3307956601604240
pedrosilva@alunos.utfpr.edu.br

Pedro Henrique Terra da Silva
Universidade Tecnológica Federal do Paraná, Brasil
http://lattes.cnpq.br/5500192844287607
brunosantos@utfpr.edu.br

Bruno Samways dos Santos
Universidade Tecnológica Federal do Paraná, Brasil
http://lattes.cnpq.br/5500192844287607
brunosantos@utfpr.edu.br

DOI: 10.6008/CBPC2179-684X.2022.003.0013

Referencing this:
SCRIAVI, A. C. P. A.; OZORIO, A. K.; SILVA, P. H. T.; SANTOS, B. S.. Estimação da frequência de utilização de serviços de saúde durante a pandemia utilizando machine learning. **Revista Brasileira de Administração Científica**, v.13, n.3, p.174-186, 2022. DOI: http://doi.org/10.6008/CBPC2179-684X.2022.003.0013

INTRODUÇÃO

Algumas das maiores transformações na vida humana, segundo Alpaydin (2016), foram resultado da computação, a qual substituiu grande parte das ferramentas e serviços que foram desenvolvidos ao longo dos séculos. Com a maior acessibilidade dos computadores, grande parte das tarefas, antes feitas manualmente, passaram a ser feitas automaticamente pela máquina. Quanto aos conjuntos de dados mais antigos, os dados possuíam um comportamento passivo, serviam apenas para serem processados e retornados aos usuários, porém, hoje, o próprio conjunto de dados conduz a operação, em que ele define o próximo passo, não mais os programadores.

Com um conjunto de dados formado, é possível encontrar padrões de comportamento e, a partir destes padrões, construir uma aproximação do comportamento futuro, prevendo o que acontecerá, além de também facilitar o entendimento do processo (ALPAYDIN, 2016). Isto é chamado de mineração de dados, que pode ser definido como um estudo sobre a coleta, limpeza, processamento, análise e obtenção de informações potencialmente úteis (AGGARWAL, 2015).

Dentro da mineração de dados, destacam-se as técnicas de machine learning (ou do português, "aprendizagem da máquina") que, de acordo com Mahesh (2019), são técnicas utilizadas para ensinar as máquinas a como lidar com os dados de forma mais eficiente por meio de algoritmos, uma vez que não se consegue interpretar e extrair aquilo que é necessário da base de dados. Ainda segundo o autor, a demanda para este processo vem aumentando por conta da grande quantidade de novos conjuntos de dados disponíveis. Corroborando esta afirmação, Sansone et al. (2021), comentam que machine learning pode ser potencializado a partir da geração e armazenamento diário da grande quantidade de dados, além dos aprimoramentos nas áreas de computação.

O tipo de algoritmo utilizado depende do tipo de problema que se busca resolver, o número de variáveis, entre outros fatores. Conforme Oladipupo (2010), machine learning consiste em projetar algoritmos que permitem que um computador aprenda, em que o ato de aprender resume-se em encontrar estatísticas, regularidades ou outros padrões nos dados. Este aprendizado não se compara ao aprendizado humano, mas pode dar uma visão sobre a dificuldade relativa de aprendizagem em ambientes diferentes.

Algo que vem gerando diversas novas bases de dados é a pandemia do novo coronavírus (COVID-19). Conforme Fayyoumi et al. (2020), a COVID-19 é uma doença infecciosa considerada uma zoonose, em que o vírus foi transmitido de um animal para um humano. Ela foi identificada em dezembro de 2019, na cidade de Wuhan, na China e se espalhou por todo o planeta, desenvolvendo a pandemia. De acordo com o Brasil (2021), até agosto de 2021, a doença no país acumulava mais de 20 milhões de casos e mais de 570 mil mortes. Ainda segundo Fayyoumi et al. (2020), diversos autores vêm utilizando técnicas de machine learning para modelar o avanço da doença e poder prever comportamentos. Esta técnica pode ser desenvolvida para prever a extensão da COVID-19 por meio de processamentos em alta velocidade nas bases de dados.

A pandemia fez com que a procura pelos serviços de saúde mudasse. Pessoas que contraíram a doença podem ter passado a necessitar mais de consultas ou podem ter deixado de procurar serviços de

saúdes rotineiros por medo de contrair o vírus na atividade. Além do mais, alguns serviços, como as APS (Atenção Primária à Saúde), encontram-se sobrecarregados por conta da demanda que o vírus causou (SAVASSI et al., 2020). Dessa forma, estimar o número de visitações de acordo com o perfil das pessoas em determinado país é importante para que cada tipo de serviço possa se preparar para receber um aumento de demanda ou sofrer com a diminuição dela. Assim, as questões de organização e efetividade podem melhorar e contribuir para a luta contra o vírus.

Quanto às aplicações de técnicas de machine learning, muitos pesquisadores têm investigado a COVID-19 em termos de diagnóstico, severidade e mortalidade em decorrência da doença, utilizando majoritariamente algoritmos de aprendizado supervisionado devido ao seu uso mais simples e fácil interpretação (ALBALLA et al., 2021). Ressalta-se que os tipos de dados mais utilizados são baseados em imagens (77%) para as tarefas de detecção e classificação, séries temporais (19%) com o intuito de estimar valores futuros relacionado à COVID-19, e apenas 4% com dados clínicos para tarefas de classificação ou estimação (KHAN et al., 2021). Assim, torna-se interessante uma investigação no campo da regressão.

Segundo Gupta (2015), os modelos de regressão linear geralmente são usados para analisar ou predizer a relação entre duas variáveis. A variável que está sendo prevista é chamada de dependente e a variável que está sendo usada para prever o valor dela é chamada de independente. Entretanto, existem diversas outras baseadas em inteligência artificial e mineração de dados que têm sido utilizadas em problemas de saúde pública, incluindo doenças contagiosas. Técnicas como as árvores de decisão, floresta aleatória, redes neurais e k-vizinhos mais próximos (ou do inglês, k-nearest neighbors – KNN) foram amplamente aplicadas na última década (SANTOS et al., 2019), e todas podem ser adaptadas para a tarefa de estimação.

Desta forma, este artigo tem como objetivo estimar, por meio da tarefa de regressão, a frequência da utilização de serviços de saúde no período de pandemia do novo coronavírus (SARS-CoV-2) pelos brasileiros e compará-la com outros 10 países do continente americano, a fim de observar o quão similar os países são em termos de características da população que levam à busca por serviços de saúde.

Após esta seção introdutória, o artigo está dividido em mais três seções. A seção 2 descreve os métodos utilizados na pesquisa, a sequência do desenvolvimento e o pré-processamento e análise exploratória dos dados. Já a seção 3 mostra e discute os resultados obtidos, enquanto a seção 4 inclui a conclusão, limitações e estudos futuros.

METODOLOGIA

Descrição do conjunto de dados

O conjunto de dados utilizado na análise foi “Premise General Population COVID-19 Health Services Disruption Survey 2020”, ou, em português, Pesquisa de Ruptura de Serviços de Saúde COVID-19 de 2020. A pesquisa foi feita pelo Instituto de Métricas e Avaliação de Saúde (*Institute for Health Metrics and Evaluation - IHME*) no ano de 2020 em 76 países. Para a pesquisa, foram feitas perguntas relacionadas a utilização dos

serviços de saúde pelos entrevistados. Todas as informações sobre esta pesquisa podem ser encontradas no website do estudo (IHME, 2020). Ao todo, foram coletadas 52.492 respostas. Sendo assim, o conjunto possuía 2.467.124 dados em sua matriz.

O conjunto de dados possui informações pessoais como país, gênero, idade, situação financeira, geografia, emprego, educação, etnia, religião e quantidade de pessoas que moram na mesma casa de cada entrevistado. Sobre os serviços de saúde, foi questionado sobre a necessidade de procurar um serviço de saúde, quantas vezes foi, se teve a possibilidade de ir ao local, qual o motivo de não ter ido, o motivo de ter ido, em qual local fez a utilização do serviço, se foi necessário tomar medicação, se a pessoa deixou de tomar alguma dose e o motivo dela ter deixado de tomar, que foram perguntadas com base no período de dezembro (2019) a fevereiro (2020) e no período a partir de março (2020).

Técnicas de Machine Learning

Existem vários modelos de aplicação de regressão baseados em machine learning que podem substituir ou complementar análises de regressão simples ou regressão múltiplas. Alguns deles são a árvore de decisão, a floresta aleatória, as redes neurais e o KNN.

Floresta Aleatória

Para a técnicas de florestas aleatórias, é importante um entendimento inicial sobre as árvores de decisão. O modelo da árvore de decisão, de acordo com Mahesh (2019), é um gráfico hierárquico que representa as escolhas (regras) e seus resultados (nós inferiores ou nós folhas) em um formato de árvore. Os nós neste gráfico determinam a decisão tomada e os quadros representam os critérios para esta decisão. A Figura 1 mostra um modelo de árvore de decisão para uma tarefa de classificação.

A Figura 1 define o nó raiz baseado na importância de uma determinada variável (ou atributo), mostrando uma árvore conhecida como árvore de classificação. Para a regressão, tem-se a diferença sobre os nós folha, no qual é tirada a média das amostras pertencentes a aquele nó (YANG et al., 2017).

Figura 1: Representação de árvore de decisão. **Fonte:** Adaptado de Mahesh (2019).

A floresta aleatória pode ser considerada como um conjunto de árvores de decisão (RASCHKA, 2015). Na regressão, elas são formadas por meio de árvores que crescem dependendo de um vetor aleatório, de modo que o preditor da árvore assume valores numéricos em oposição aos rótulos de classe (BREIMAN, 2001). A ideia de fazer esses conjuntos é de possibilitar a combinação de classificadores fracos (weak learners) para construir um modelo mais robusto, chamado de classificador forte (strong learner), que possui um melhor erro generalizado e é menos suscetível ao sobre ajuste (overfitting). Este tipo de algoritmo pode

ser resumido em quatro passos: 1. Desenhar uma amostra de bootstrap aleatória de tamanho n; 2. Criar uma árvore de decisão a partir desta amostra, em que, em cada nó deve-se selecionar aleatoriamente características sem substituição e separar o nó utilizando a característica que promove a melhor separação de acordo com o objetivo da função, por instância, por meio da maximização do ganho de informação; 3. Repetir os passos 1 e 2 k vezes; 4. Agregar à predição por meio de cada árvore para atribuir o rótulo da classe por voto da maioria. A Figura 2 traz a representação da floresta aleatória.

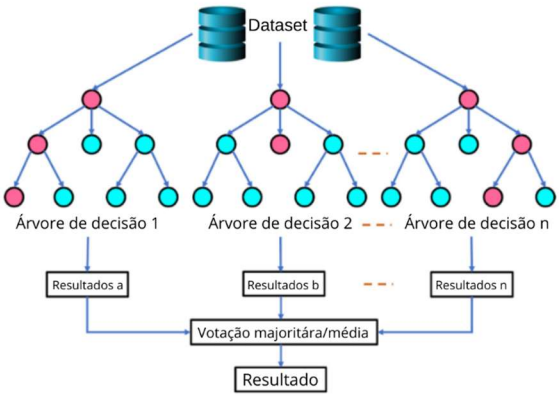

Figura 2: Representação de floresta aleatória. **Fonte:** LIU et al. (2021).

Redes Neurais

As redes neurais (ou redes neurais artificiais), segundo Mahesh (2019), são uma série de algoritmos que se esforçam para reconhecer relações subjacentes em uma base de dados por meio de um processo que imita a atividade de uma rede neural do cérebro humano. Conforme é mostrado na Figura 3, as redes recebem um input no que é chamado "camada de entrada" da rede neural para que ela possa gerar o melhor resultado possível a partir da transmissão para os neurônios da camada de saída (LIOU et al., 2008).

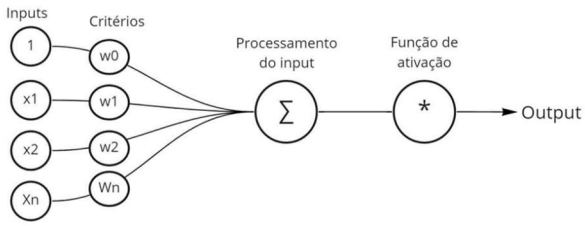

Figura 3: Representação das redes neurais. **Fonte:** Adaptado de MAHESH (2019).

No caso das redes neurais, as funções de ativação de um neurônio mais comuns são a linear (ou identidade), sinal, sigmoide e tangente hiperbólica (AGGARWAL, 2018). Para esta pesquisa, foi aplicada a função ReLU (Rectified Linear Unit).

K-Vizinhos Mais Próximos

O algoritmo k vizinhos mais próximos, ou do original, k-nearest neighbors (KNN) é um "típico exemplo de aprendiz preguiçoso (lazy learner), por conta de não aprender uma função discriminativa do conjunto de treinamento, memorizando-o ao invés" (RASCHKA, 2015). Ele pode ser resumido nos seguintes passos: (i) escolher o número de k; (ii) Escolher e calcular a métrica de distância; (iii) classificar com a classe mais

frequente a partir da avaliação do subconjunto dos k-vizinhos mais próximos da amostra de teste. Esta definição pode ser aplicada para a tarefa de classificação.

A técnica do KNN, porém, pode ser usada tanto para classificação quanto para regressão. Quando usada para regressão (ou estimação), o valor da instância de teste é calculado como uma soma ponderada das respostas (vizinhos) válidos para todos os k vizinhos, e o peso é inversamente proporcional à distância do registro de entrada. A partir da escolha de k, as estimações baseadas nos exemplos de treino podem ser feitas com as médias da variável de interesse (saída), como especificado por apud Al-Dosary et al. (2019). A Figura 4 mostra um exemplo de KNN para o caso da regressão.

Para o caso da Figura 4, o ponto vermelho representa a nova instância, e com um k = 5, tira-se a média da variável y dos cinco pontos (registros) mais próximos, ou seja, aqueles identificados a partir da matriz de atributos de entrada X.

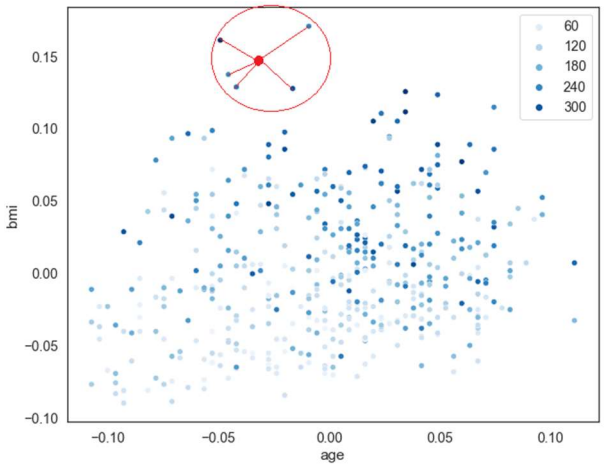

Figura 4: Representação das redes neurais.

Sequência do desenvolvimento da pesquisa e do tratamento de dados

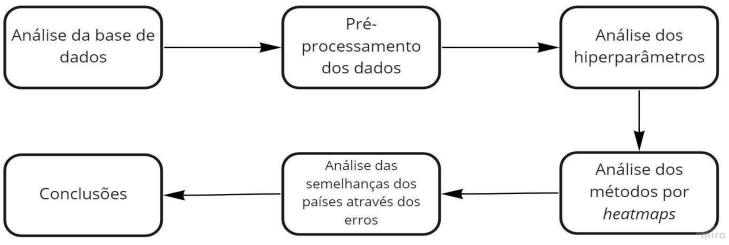

Figura 5: Sequência de implementação da pesquisa.

A sequência do desenvolvimento da pesquisa está apresentada e detalhada na Figura 5. Para cumprir com o objetivo do artigo, foram utilizadas algumas das variáveis disponíveis na base de dados. Elas e seus respectivos tipos (dependente e independente) estão apresentados no Quadro 1.

Quadro 1: Variáveis utilizadas no artigo e seus tipos.

Variável	Tipo
Idade	Independente
Gênero	Independente
Situação financeira	Independente
Educação	Independente
Emprego	Independente
Precisou utilizar serviço de saúde?	Dependente
Motivos pelos quais precisou utilizar o serviço de saúde	Independente

O período escolhido para análise foi a partir de março de 2020, uma vez que buscou-se incluir a atuação significativa da COVID-19 na necessidade da utilização destes serviços. Os algoritmos utilizados para esta pesquisa foram para a tarefa de regressão, uma vez que se desejava estimar numericamente uma variável a partir de um conjunto de entrada com as variáveis dependentes.

Primeiramente, o conjunto foi pré-processado. A primeira ação foi excluir da base de dados todas as variáveis que não eram necessárias, deixando apenas aquelas que seriam analisadas. Posteriormente, os valores vazios na variável "quantidade de idas a serviços de saúde" decorrentes da resposta "não" nas variáveis "precisou utilizar algum serviço de saúde?" e "teve condições de ir a um serviço de saúde?" foram preenchidas com o valor 0 (zero), simbolizando que a pessoa em não teria visitado serviços de saúde desde março de 2020. Após isso, nesta mesma variável, para fins técnicos, foram excluídas as respostas que indicavam "10 ou mais" e as que ainda continham valores vazios, ficando, assim, a quantidade de pessoas que foram de 0 (zero) a 9 (nove) vezes a serviços de saúde no período a partir de março de 2020.

Na sequência, visto que em cada resposta da variável "motivos pelos quais a pessoa foi a serviços de saúde", os motivos estavam contidos todos juntos em formato de texto com respostas agregadas, sem assim necessário criar colunas com respostas no formato binário para uma melhor análise. Dessa forma, o conjunto de dados recebeu uma nova coluna para cada motivo contido na variável e, cada uma delas recebeu os valores dicotomizados, correspondendo à procura ou não de um serviço de saúde por aquele motivo. Um exemplo desta binarização está apresentada na Tabela 1, para três pacientes e quatro motivos. Nela é possível observar que, quando o paciente fez a visitação, recebe o valor 1 e, quando não, o valor 0.

Tabela 1: Exemplo da binarização da variável "motivos pelos quais foi a serviços de saúde".

Paciente	Câncer	Covid-19	Consulta de rotina	Asma
1	0	1	0	1
2	1	0	1	0
3	0	0	1	0

Então, todas as variáveis nominais ainda restantes no conjunto, salvo "país", também foram binarizadas. A variável "país" não foi binarizada por conta de o nome do país ser essencial na divisão dos subconjuntos de treino e teste para os métodos de regressão, exercendo apenas um identificar para filtro dos conjuntos de treino e teste. Após processado, o conjunto de dados passou a possuir 51.995 instâncias e 60 variáveis, com 59 sendo preditoras, e uma variável de saída (número de visitas).

As análises foram feitas utilizando a linguagem de programação Python 3 (PYTHON FOUNDATION, 2021), juntamente com as bibliotecas pandas, numpy, matplotlib, seaborn, math e sklearn. O conjunto de treino foi composto pelos dados do Brasil e os de teste pelos 10 países americanos analisados (Chile, Costa Rica, Colômbia, Estados Unidos da América, El Salvador, Equador, Honduras, México, Panamá e Uruguai) e o próprio Brasil. O Brasil foi usado tanto para o conjunto de treino quanto para o conjunto de teste, com o objetivo de verificar possíveis underfititngs ou overfittings. Em relação aos outros países, quando testados, aqueles que obtiveram menor erro, podem indicar uma maior semelhança em termos de características e comportamento da população em relação à brasileira.

Análise dos hiperparâmetros

Para a aplicação da floresta aleatória, primeiramente foram utilizados os números de estimadores (árvores de decisão) de 500 a 1.000, em um intervalo de 100. Destes, foram comparados os valores de erro quadrático médio (mean square error - MSE) ao longo do número de estimadores (Figura 6). Pode-se analisar que o menor erro ocorreu quando foram utilizados 1.000 estimadores. Assim, o modelo de floresta aleatória foi implementado com o parâmetro "n_estimators" = 1.000.

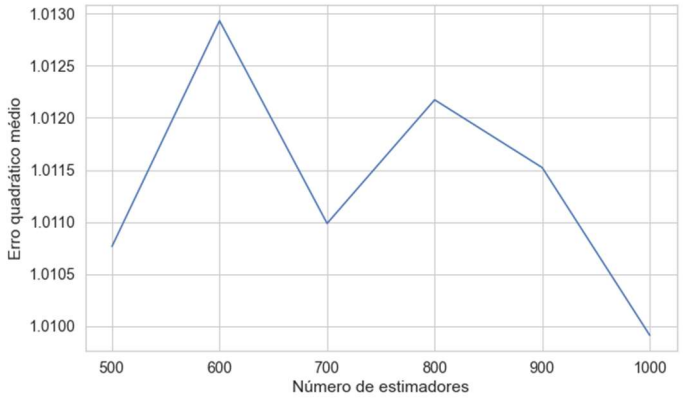

Figura 6: MSE dos estimadores de 500 a 1.000 na floresta aleatória.

No método de redes neurais, o número de iterações variou de 500 a 2.000, em um intervalo de 100. Assim como no método de floresta aleatória, o MSE foi analisado para comparar cada estimador. Analisando-se a Figura 7 percebeu-se que o menor erro ocorreu quando o número de iterações foi 500.

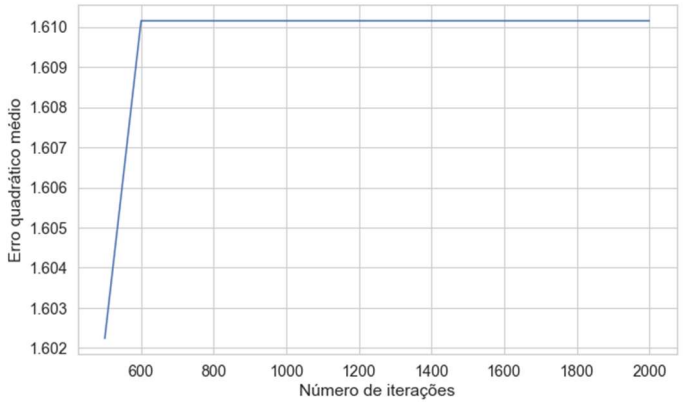

Figura 7: MSE variando as iterações entre 500 e 2.000 nas redes neurais.

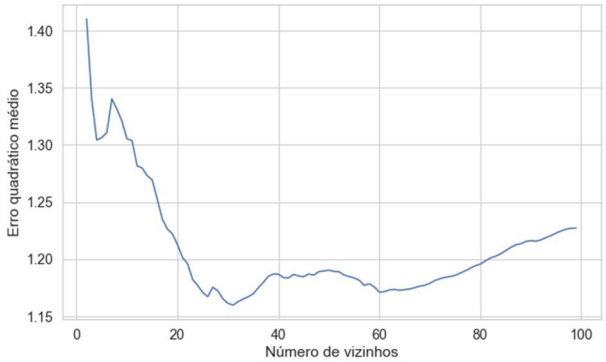

Figura 8: MSE dos vizinhos de 2 a 100 para o KNN.

No método KNN, foram utilizados de 2 a 100 vizinhos (valor de k vizinhos) para verificar a efetividade de cada um. Escolheu-se o mesmo erro (MSE) para realizar a comparação entre os vizinhos, tendo como menor resultado o número de vizinhos igual a 30 (Figura 8).

Para todas as técnicas anteriores (com os seus respectivos hiperparâmetros), foram analisados nos conjuntos de treino (Brasil) e de teste (outros países) o erro absoluto médio (mean absolute error – MAE), MSE e a raiz do erro quadrático médio (root mean squared error – RMSE).

RESULTADOS

Com um mapa de calor (do inglês, heatmap) gerado a partir dos valores do MSE nos três métodos, foi possível observar que o método da floresta aleatória tem os menores valores de erro (Figura 9). Percebe-se na Figura 9 que existe uma diferença aparente entre as três técnicas utilizadas. O menor erro encontrado foi para floresta aleatória (identificado por RF na Figura 9), com um menor valor para países como o Uruguai e Honduras, (0,6 e 0,69, respectivamente).

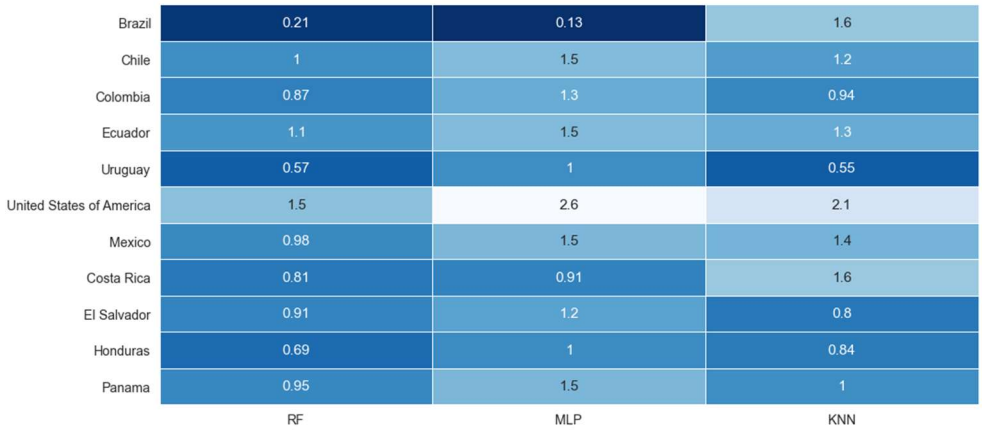

Figura 9: Comparação dos resultados do MSE das três técnicas.

Após a análise do MSE de cada um dos métodos, foi gerado um novo heatmap, tendo como referência a floresta aleatória, e comparou-se percentualmente a variação deste erro para cada país, com os outros dois métodos. Isso foi feito com o objetivo de identificar a variação dos resultados destes outros dois métodos em relação à floresta aleatória como destacado na Figura 10.

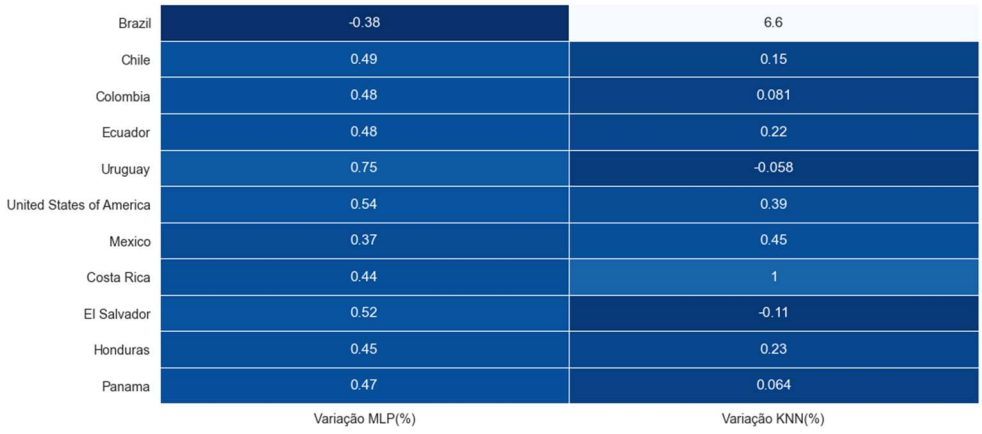

Figura 10: Valores de variação percentual do MSE das redes neurais e KNN, comparados com floresta aleatória.

Os resultados apresentados na Figura 10 indicaram que a variação foi positiva para ambos os métodos e na grande maioria dos países, o que indica que os valores de MSE encontrados pelas redes neurais e pelo KNN foram maiores do que o valor encontrado por meio da aplicação da floresta aleatória.

Na sequência, foram gerados os gráficos de barra para o erro MAE e RMSE, utilizados tanto em estimação como em previsão de série temporal, como aplicado no trabalho de Borges e Mattos (2021), a fim de observar as diferenças e semelhanças entre a população dos países analisados e a população brasileira. Neste caso, o Brasil também foi utilizado como teste para fins de análise da eficiência dos métodos. A partir da métrica do MAE, é possível observar que o país com um comportamento e características mais parecido ao Brasil é o Uruguai e o mais distinto são os Estados Unidos da América (Figura 11).

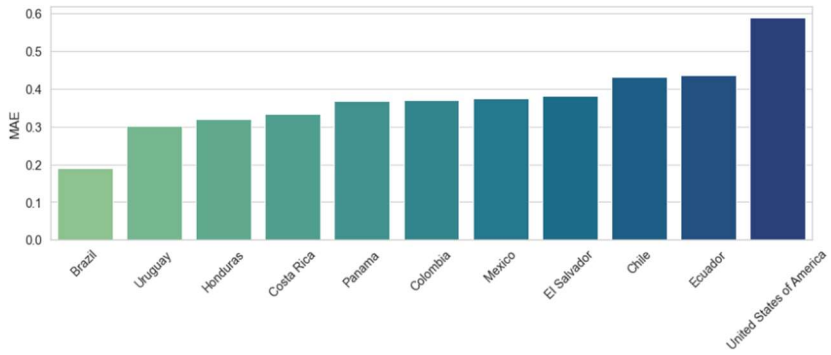

Figura 11: Valores do MAE dos países por floresta aleatória.

A partir do RMSE, conforme mostrado na Figura 12, nota-se os desvios dos erros em relação ao Brasil, com o Uruguai, Honduras e Costa Rica obtendo desvios de até 0,9 visitas, enquanto o cidadão estadunidense tem um desvio maior, de até 1,2 visitas.

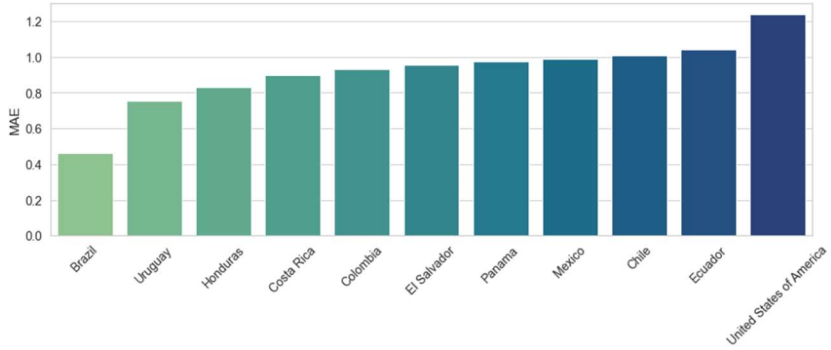

Figura 12: Valores RMSE dos países por floresta aleatória.

Ao avaliar a comparação entre os três métodos, é possível levantar a hipótese de que a floresta aleatória apresentou o melhor desempenho por selecionar variáveis por meio das árvores de decisão, enquanto o as redes neurais tradicionais e o KNN utilizam todas as variáveis fornecidas originalmente. Além disso, o erro do Brasil foi alto para o KNN, o que pode indicar uma possível situação de underfitting. Isto pode acontecer quando o modelo não se adapta bem nem ao conjunto treino, nem ao conjunto de teste. Já as redes neurais apresentaram um erro menor quanto testado com o conjunto brasileiro, mas os erros aumentaram para os outros países, indicando aqui um overfitting.

O Uruguai se mostrou o país com a população mais parecida ao Brasil em decorrência do menor erro

dos algoritmos para este país. Em termos de aprendizado de máquina, isto mostra que a etapa de treinamento dos algoritmos, com os exemplos contidos no conjunto de cidadãos brasileiros, conseguiu compreender com um menor erro em relação aos outros países, o comportamento em relação ao número de visitações aos centros de saúde.

Como adição à discussão, foi realizada uma análise das variáveis por meio do teste qui-quadrado X2 para avaliar algumas das variáveis que mais contribuíram para estimar o número de visitas. Como as variáveis de entrada foram binarizadas, para cada variável, cada categoria dela se tornou uma variável específica (vide Tabela 1), que uma variável que continha tipos de doenças, expandiu para o número de doenças. Assim, as variáveis que mais influenciaram na estimação poderiam estar atreladas à mesma dimensão, como idade, sexo, doenças, entre outras. A Figura 13 apresenta as dimensões mais importantes para se estimar o número de visitas, com o valor do eixo y sendo a frequência de variáveis que esta dimensão obteve.

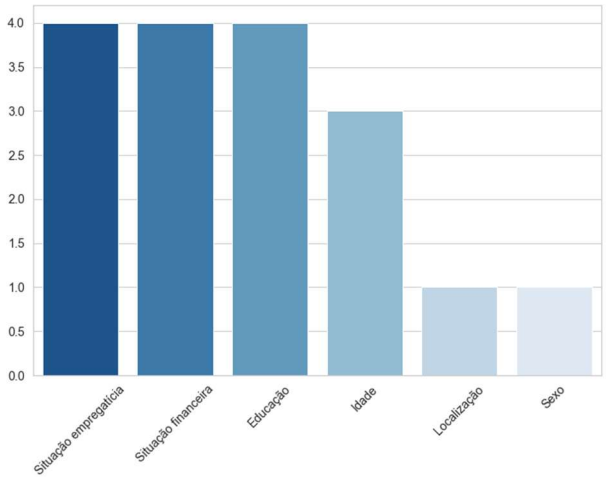

Figura 13: Dimensões mais importantes para se estimar o número de visitas.

Como pode ser observado na Figura 13, a situação empregatícia teve quatro variáveis entre as mais importantes, juntamente com a situação financeira e nível de escolaridade. Sabe-se que a procura por hospitais ou postos de saúde aumentou muito durante a pandemia, gerando um desafio para diversas nações.

O trabalho de Acosta (2020) mostrou que o Índice de Rigor de resposta do governo contra a pandemia foi menor para os países como Uruguai, Brasil e Costa Rica, podendo levar os habitantes destes países a procurar os sistemas de saúde devido às dúvidas geradas no início da pandemia. Este mesmo estudo, por meio de uma análise de correspondência múltipla, enfatizou que houve um maior êxito na luta contra a pandemia nesta época por países como Uruguai, Costa Rica e Cuba, pois estes possuem uma população total menor do que vários outros da América Latina. A taxa de mortalidade baixa também foi associada com um nível médio de urbanização e baixo nível de pobreza. Outra evidência é que o número de mortes neste período devido à COVID-19 foi encontrado em uma proporção média da população que vive com menos de um dólar por dia, destacando a importância das variáveis financeiras Acosta (2020). Entretanto, não foram encontradas maiores referências sobre os motivos das populações de Uruguai, Honduras e Costa Rica terem uma maior similaridade com o Brasil na área da saúde em tempos de pandemia.

É importante ressaltar que o motivo da visita (doenças) não foi significativo para se estimar o número de visitas em estabelecimentos de saúde, mostrando assim que as pessoas vão ao hospital pelos mais diversos tipos de doenças, sendo mais importante as variáveis demográficas as que mais se destacam para estas pessoas (vide Figura 13).

CONCLUSÕES

Foi desenvolvido, nessa pesquisa, um estudo de análise de regressão, a fim de analisar similaridades entre o Brasil e outros dez países da América, a respeito da frequência de utilização dos serviços de saúde durante a pandemia.

Para isso, foi realizado um pré-processamento dos dados da "Premise General Population COVID-19 Health Services Disruption Survey 2020", seguido de um treinamento e de testes para se definir os melhores parâmetros dos algoritmos. Por último, foram aplicados três métodos (floresta aleatória, redes neurais e KNN) para se analisar as características e comportamento dos brasileiros em relação aos outros países.

Com as análises feitas, foi possível concluir que o país com o comportamento da população mais distinta da brasileira são os Estados Unidos e o mais parecido é o Uruguai. Estes resultados foram encontrados para diferentes erros e para os três algoritmos aplicados. O melhor resultado (erro) encontrado foi com floresta aleatória.

Os resultados obtidos por meio da tarefa de regressão evidenciam a importância da análise e tratamento de dados para prever determinados eventos. No caso desta pesquisa, o estudo serviu para evidenciar similaridades e diferenças entre a população brasileira e de outros países a respeito de seu comportamento com relação aos cuidados com a saúde.

Também foi possível perceber que a análise de regressão é uma importante ferramenta de estimação de dados nos dias de hoje, podendo ser utilizada em muitas outras áreas do conhecimento para se observar tendências, podendo se estender até à análise de comportamentos de populações inteiras, como no estudo realizado nesta pesquisa.

As limitações desta pesquisa se baseiam na aplicação de apenas três técnicas de regressão, como também na limitação de exemplos para treinamento em relação a cada país. Também foram selecionados os atributos que tiveram uma maior disponibilidade de dados. Para estudos futuros, técnicas alternativas poderiam ser utilizadas, como máquinas de vetores de suporte (support vector machines), como também a utilização de outras variáveis de entrada, tentando incluir outras características e comportamentos destas amostras avaliadas.

REFERÊNCIAS

ACOST, L. D.. Capacidade de respuesta frente a la pandemia de COVID-19 em América Latina y el Caribe. **Revista Panamericana de Salud Publica**, v.44, p.109, 2020. DOI: https://doi.org/10.26633/RPSP.2020.109

AGGARWAL, C. C.. **Data Mining**. Cham: Springer, 2015.

AGGARWAL, C. C.. **Neural Networks and Deep Learning**. Cham: Springer, 2018.

AL-DOSARY, N. M. N.; AL-HAMED, S. A.; ABOUKARIMA, A. M.. K-nearest neighbors method for prediction of fuel consumption in tractor-chisel plow systems. **Engenharia Agrícola**, v.39, n.6, p.729-736, 2019. DOI:

https://doi.org/10.1590/1809-4430-Eng.Agric.v39n6p729-736/2019

ALBALLA, N.; AL-TURAIKI, I.. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. **Informatics in Medicine Unlocked**, v.24, p.100564, 2021. DOI: https://doi.org/10.1016/j.imu.2021.100564

ALPAYDIN, E.. **Machine Learning**: The New AI. Cambridge: The MIT Press, 2016.

BRASIL. **Painel Coronavirus Brasil**. MS, 2021.

BORGES, F. S.; MATTOS, V. L. D.. Modelo de Previsão para uma série temporal de dados de cabotagem. **Revista Mundi: Engenharia, Tecnologia e Gestão**, v.6, n.3, p.01-353, 2021. DOI: http://dx.doi.org/10.21575/25254782rmetg2021vol6n31642

BREIMAN, L.. Random forests. **Machine Learning**, v.45, p.5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324

SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P.. Data mining and machine learning techniques applied to public health problems: a bibliometric analysis from 2009 to 2018. **Computers and Industrial Engineering**, v.138, p.106120, 2019. DOI: https://doi.org/10.1016/j.cie.2019.106120

FAYYOUMI, E.; IDWAN, S.; ABOSHINDI, H.. Machine learning and statistical modelling for prediction of novel COVID-19 patients case study: Jordan. **International Journal of Advanced Computer Science and Applications**, v.11, n.5, p.122-126, 2020. DOI: https://doi.org/10.14569/IJACSA.2020.0110518

GUPTA, S.. A Regression modeling technique on data mining. **International Journal of Computer Applications**, v.116, n.9, p.27-29, 2015. DOI: https://doi.org/10.5120/20365-2570

IHME. Institute for Health Metrics and Evaluation. **Premise data corporation**: premise general population COVID-19 health services disruption survey 2020. 2020.

KHAN, M.; MEHRAN, M. T.; HAQ, Z. U.; ULLAH, Z.; NAQVI, S. R.; IHSAN, M.; ABBASS, H.. Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. **Expert Systems with Applications**, v.185, 2021. DOI: https://doi.org/10.1016/j.eswa.2021.115695

LIOU, F.-M.; TANG, Y.-C.; CHEN, J.-Y.. Detecting hospital fraud and claim abuse through diabetic outpatient services. **Health Care Management Science**, v.11, n.4, p.353-358, 2008. DOI: https://doi.org/10.1007/s10729-008-9054-y

LIU, Y.; ESAN, O. C.; PAN, Z.; AN, L.. Machine learning for advanced energy materials. **Energy and AI**, v.3, p.100049, 2021. DOI: https://doi.org/10.1016/j.egyai.2021.100049

MAHESH, B.. Machine learning algorithms: a review. **International Journal of Science and Research**, v.9, n.1, p.381-386, 2019. DOI: https://doi.org/10.21275/ART20203995

OLADIPUPO, T.. **Types of machine learning algorithms**: new advances in machine learning. Rijeka: InTech, 2010.

PYTHON FOUNDATION. **Python**. 2021.

RASCHKA, S.. **Python machine learning**. Birmingham: Packt, 2015.

SANTOS, B. S.; STEINER, M. T. A.; FENERICH, A. T.; LIMA, R. H. P.. Data mining and machine learning techniques applied to public health problems: a bibliometric analysis from 2009 to 2018. **Computers and Industrial Engineering**, v.138, p.106120, 2019. DOI: https://doi.org/10.1016/j.cie.2019.106120

SANSONE, V. T. B.; VECCHIA, R. D.. Construção do modelo preditivo de desligamento de colaboradores. **Revista Brasileira de Administração Científica**, v.12, n 4, p.171-189, 2021. DOI: https://doi.org/10.6008/CBPC2179-684X.2021.004.0012

SAVASSI, L. C. M.; DIAS, B. A.; ABREU, A. B. J.; COSTA, A. C.; PERDIGÃO, R. M. C.; FERREIRA, T. P.. Ensaio acerca das curvas de sobrecarga da COVID-19 sobre a atenção primária. **Journal of Management & Primary Health Care**, v.12, p.1-13, 2020. DOI: https://doi.org/10.14295/jmphc.v12.1006

YANG, L.; LIU, S.; TSOKA, S.; PAPAGEORGIOU, L. G.. A regression tree approach using mathematical programming. **Expert Systems with Applications**, v.78, p.347-357, 2017. DOI: https://doi.org/10.1016/j.eswa.2017.02.013