ISSN: 2237-826X
ARTIGO | PAPER

Uso do algoritmo "Floresta Aleatória" na identificação do comportamento da população na busca por serviços de saúde após o início da pandemia do novo coronavírus

Use of the random forest algorithm to identify the behavior of the population in the search for health services after the beginning of the new coronavirus pandemic

Vinicius Matheus Pimentel Ariza[1], Mateus Miranda do Nascimento[2], Pedro Picolo Malandrino[3], Bruno Samways dos Santos[4]

[1] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná. ORCID: https://orcid.org/0000-0002-0493-7971*

[2] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná.*

[3] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná. ORCID: https://orcid.org/0000-0001-8945-6191*

[4] *Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, Paraná. ORCID: https://orcid.org/0000-0001-7919-1724*

Autor para correspondência/*Mail to:* Vinicius Matheus Pimentel Ariza, viniciusarizaibi@gmail.com

Recebido/*Submitted*: 26 de abril de 2022; **Aceito/*Approved*:** 30 de junho de 2022

Resumo

Introdução: A mineração de dados é uma das cinco etapas do *Knowledge Discovery in Databases* (KDD), ou Descoberta de Conhecimento em Banco de Dados, e pode ser aplicada em diversas áreas do conhecimento, incluindo a saúde. As preocupações e esforços com a saúde pública vêm ganhando maior relevância, uma vez que o mundo todo vem sofrendo com o enfrentamento da pandemia do novo coronavírus, principalmente em países subdesenvolvidos e em desenvolvimento, como é o caso do Brasil e outras nações do continente americano. **Método:** Visando contribuir para a área, o objetivo deste estudo foi aplicar uma tarefa de mineração de dados, utilizando o algoritmo Floresta Aleatória, para classificar o comportamento da população na busca por serviços de saúde após o início da pandemia do novo coronavírus. A base de dados utilizada foi a *Premise General Population* Covid-19 *Health Services Disruption Survey* 2020, oriunda do projeto Covid-19 *Health Services Disruption Survey* 2020. **Resultados:** Utilizando o algoritmo Floresta Aleatória, obteve-se 87,5% de acurácia na classificação de pessoas que irão ou não recorrer a serviços de saúde em países americanos utilizando métodos de balanceamento de classes. Também se chegou a um valor de sensibilidade *(recall)* de 93% e precisão de 86% nos melhores modelos. **Conclusão:** O modelo desenvolvido e os resultados alcançados podem ser usados para auxiliar autoridades de países americanos no planejamento de políticas públicas de saúde.

Palavras-chave: Floresta aleatória; Mineração de dados;Covid-19; Tarefa de classificação.

Abstract

Introduction: *Data mining is one of the five stages of Knowledge Discovery in Databases (KDD) and its application can be found on a daily basis in the most varied sectors of the economy and study, being widely used in the health area. Health concerns and efforts have been gaining relevance and strength as the entire world has been suffering from the fight against the coronavirus pandemic, especially in underdeveloped and developing countries, such as Brazil and other nations of the American continent.* ***Method:*** *Aiming to contribute to the area, the objective of this study was to apply a data mining task, using the Random Forest algorithm, to classify the population's behavior on the search for health services after the onset of the coronavirus pandemic. The database used was the Premise General Population Covid-19 Health Services Disruption Survey 2020 from the COVID-19 Health Services Disruption Survey 2020 project.****Results:*** *Using the Random Forest algorithm, 87.5% accuracy was obtained in the classification of people who whether or not they will use health services in American countries. A recall value of 93% and precision of 85% were reached in the best models.* ***Conclusions:*** *The model developed and the results achieved can be used to assist authorities in American countries on planning public health policies.*

Keywords: *Random forest; Data mining; Covid-19; Classification task.*

INTRODUÇÃO

A mineração de dados, também conhecida pelo termo em inglês *data mining*, é uma das etapas da descoberta de conhecimento em Banco de Dados (KDD - *Knowledge Discovery in Databases*). Fayyad, Piatetsky-Shapiro, e Smyth (1996, p.83) definem o KDD como "o processo não-trivial de identificação válida, em dados novos, potencialmente úteis e finalmente com padrões compreensíveis".Assim, o objetivo do KDD é obter conhecimento a partir de uma série de dados.

O processo da descoberta de conhecimento a partir de uma base de dados é composto por cinco fases: a seleção dos dados, o pré-processamento, a transformação, a mineração e a interpretação dos resultados. A mineração de dados é uma das etapas do KDD e pode ser definida como o processo de exploração de grandes quantidades de dados em busca de correlações significativas e padrões consistentes (Larose & Larose, 2014).

O uso da mineração de dados está presente nas mais variadas aplicações, como por exemplo, no setor de seguros para identificação de fraudes (Severino & Peng, 2021), no reconhecimento de padrões em emissões domésticas

de dióxido de carbono (Chen, Wu, Zhong, Long, & Meng, 2022), consumo energético (Tang, Wang, Lee, & Yang, 2022), engenharia civil (Naderpour, Mirrashid, & Parsa, 2021), manutenção preditiva (Tessoni & Amoretti, 2022), doenças cardíacas (Pan, Poddar, Mukherjee, & Ray, 2022), desnutrição (Islam et al., 2022), educação superior (Gil, da Cruz Martins, Moro, & Costa, 2021), entre outras áreas. Uma das áreas de pesquisa adepta à mineração de dados é o setor da saúde pública, incluindo principalmente assuntos relacionados à pandemia do novo coronavírus.

Recentemente, o mundo todo apresentou dificuldades no enfrentamento da pandemia do novo coronavírus. A saúde pública de diversos países entrou em colapso em apenas algumas semanas e, em alguns casos, apenas dias. Até mesmo em países desenvolvidos foi observado o colapso dos sistemas de saúde pela alta procura em um curto espaço de tempo e pela alta demanda por leitos ocasionada pela doença, como foi observado na Itália no início do ano de 2020 (Armocida, Formenti, Ussai, Palestra, & Missoni, 2020).

No caso dos países subdesenvolvidos ou em desenvolvimento, incluindo o Brasil, antes mesmo da aparição do novo coronavírus, por diversas vezes observou-se hospitais lotados, atendimento em corredores, falta de leitos e longas filas de espera (Azambuja, 2014). Assim, a pandemia, dentre muitos aprendizados, evidenciou e reforçou a necessidade de melhoria no planejamento e condução de políticas públicas de saúde no mundo todo, especialmente nestes países que apresentavam problemas anteriormente, e uma das formas para isso é conhecer e entender melhor possíveis padrões que possam surgir no comportamento das pessoas destes países, incluindo o aprendizado de máquina.

Com relação às aplicações dos métodos de aprendizado de máquina na saúde pública, a revisão sistemática de dos Santos, Steiner, Fenerich, e Lima (2019) identificou que muitas publicações abordam esta temática, alcançando mais de 250 trabalhos em periódicos desde 2009. Nesta mesma revisão, a Floresta Aleatória foi a terceira mais aplicada pelos autores da área, com 48 aparições, ficando atrás apenas de máquinas de vetores de suporte (do inglês, *support vector machines*, ou SVM), com 64, e das Árvores de decisão (total de 57). Também ficou evidenciado o crescimento das aplicações na área específica de doenças transmissíveis/contagiosas, com um destaque para os anos de 2017 e 2018, último período analisado pelos autores e que, somados, totalizaram 36 artigos. Esta mesma área da saúde pública também foi a mais explorada entre 2009 e 2018, correspondendo a mais de 25% de todas as publicações encontradas (Santos et al., 2019).

Em vista disso, o objetivo deste trabalho é aplicar uma tarefa de mineração de dados, utilizando o algoritmo Floresta Aleatória para desenvolver um modelo capaz de classificar o comportamento da população na busca por serviços de saúde após o início da pandemia do novo coronavírus. Para tal, foi utilizada a base de dados gerados pela pesquisa *Premise General Population* COVID-19 *Health Services Disruption Survey* 2020 comparando-se diferentes parâmetros da Floresta Aleatória e métricas de classificação.

Após esta seção introdutória, o restante do artigo está organizado da seguinte forma: na seção 2 apresenta-se o referencial teórico com explicações sobre as técnicas e mineração de dados; a seção 3 descreve os materiais e métodos utilizados com a sequência das etapas da pesquisa, com a manipulação do conjunto de dados e ferramentas utilizadas; na seção 4 são apresentados os principais resultados, seguido pelas discussões na seção 5. Por fim, na sexta e última seção, são discorridas as conclusões obtidas.

REFERENCIAL TEÓRICO

Dentro da etapa mineração de dados do processo de KDD, existe um conceito que é a capacidade que alguns algoritmos têm de aprendizado a partir dos relacionamentos existentes ou não entre os dados. As duas principais abordagens existentes são o aprendizado supervisionado e o não supervisionado (Goldschmidt, Passos, & Bezerra, 2015).

O aprendizado supervisionado consiste na concepção do modelo de conhecimento a partir de dados expostos em dados de entrada e de saída. Entende-se como entrada o conjunto de valores dos atributos de entrada, também conhecido como atributos preditores. Já o atributo-alvo é a saída desejada que o algoritmo seja capaz de produzir ao receber os atributos de entrada. Portanto, neste tipo de aprendizado, é necessário particionar os conjuntos de dados para treino e teste (Raschka, 2015).

Por outro lado, no aprendizado não supervisionado não há informações sobre a saída que se deseja; assim, o processo de aprendizado se dá por meio da identificação de regularidades que permitam agrupar os dados com base nas similaridades que apresentam entre si. Dessa forma, não se faz necessário o particionamento dos dados (Aggarwal, 2015).

Dentre as tarefas que utilizam o aprendizado supervisionado, a classificação é uma das mais utilizadas na mineração de dados. O algoritmo de classificação Floresta Aleatória, utilizado no desenvolvimento do estudo, como o próprio nome sugere, cria uma combinação de várias Árvores de Decisão de modo aleatório, a fim de obter uma predição estável e com maior acurácia (Breiman, 2001). Assim, deve-se entender o funcionamento das Árvores de Decisão para o melhor entendimento da Floresta Aleatória.

A Árvore de Decisão trabalha com modelagem de processos dentro de um conjunto de decisões hierárquicas sobre os atributos, os quais resultam em uma árvore de regras associadas. Estas decisões tomadas dentro de cada nó da árvore são chamadas de "critério de divisão" e são condições sobre um ou mais atributos nos dados de treinamento (Aggarwal, 2015). Graficamente, conforme ilustra a Figura 1, é possível identificar os nós que representam um determinado teste em um atributo. O nó de partida é também conhecido como raiz. Cada ramo representa os critérios de divisão dos testes, sendo que o nó que não possui descendentes, ou seja, não possui saídas, é chamado de folha. A vantagem é que a Floresta Aleatória combina diversas Árvores de Decisão criadas com seleções aleatórias de seus atributos, convergindo o resultado para a classificação mais realizada pelas diferentes árvores (Breiman, 2001), conforme é possível visualizar por meio da esquematização da Figura 1.

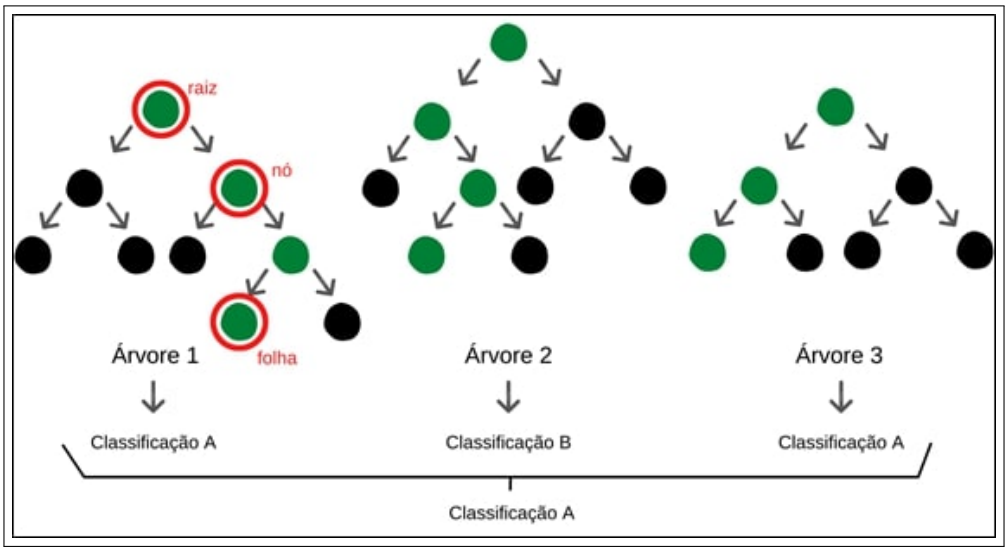

Figura 1. Representação esquemática de uma Floresta Aleatória.
Fonte: Os autores (2022).

Sendo assim, a Figura 1 mostra três Árvores de Decisão na Floresta Aleatória já treinada, de forma que, após diferentes características de construção e regras, cada uma delas retornou como resultado uma classificação. A Árvore 1 retornou a classificação A, enquanto a Árvore 2 retornou a classificação B e, por último, a Árvore 3 também retornou a classificação A. Dessa maneira, como o algoritmo Floresta Aleatória atribui uma classificação final com base nas classificações de todas as árvores e visto que duas das árvores retornaram classificação A e apenas uma retornou classificação B, o algoritmo retorna, como classificação final, a classificação A, dado que é o resultado da maioria das árvores consideradas (voto majoritário).

Esta técnica introduz uma aleatoriedade no desenvolvimento das árvores a partir de um subconjunto de características, resultando em uma grande diversidade da árvore, trocando um alto viés por uma baixa variância (Géron, 2019).

MATERIAIS E MÉTODOS

Para o presente estudo, utilizou-se o conjunto de dados *Premise General Population* COVID-19 *Health Services DisruptionSurvey* 2020, oriundo do projeto COVID-19 *Health Services DisruptionSurvey* 2020. Tais informações foram obtidas por meio de uma pesquisa realizada por meio de um aplicativo de *smartphone*. Ao todo, a sondagem contou com a participação de 52.492 respondentes, residentes de 76 países, consultados ao longo do mês de julho de 2020 (IHME, 2020).A pesquisa objetivou avaliar, antes e imediatamente após o início da pandemia do novo coronavírus, mudanças nos níveis de prestação de serviços de saúde.O pré-processamento dos dados deste conjunto foi baseado no fluxograma da Figura 2.

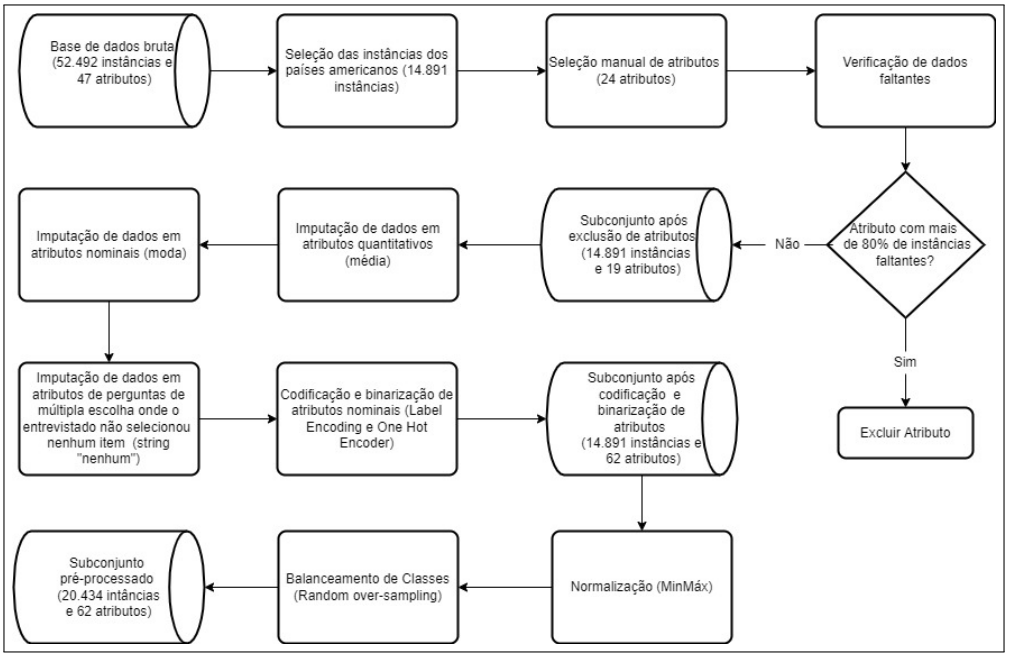

Figura 2. Fluxograma do Pré-Processamento dos Dados.
Fonte: Os autores (2022)

Dada a extensão do conjunto, para a aplicação da tarefa de mineração de dados, foram utilizadas somente respostas de residentes dos países do continente americano. De igual maneira, atributos com informações omitidas por questão de confidencialidade ou com mais de 80% de dados faltantes foram desconsiderados, sendo que, para os atributos restantes, os itens ausentes foram substituídos pela média, pela moda ou por "nenhum" para perguntas de múltipla escolha onde não foi selecionado nenhum item.

Além disso, algumas questões foram feitas em duplicidade aos entrevistados, porém, inicialmente em referência ao período de dezembro de 2019 a fevereiro de 2020 e, posteriormente, tratando de eventos ocorridos após o mês de março de 2020. Dessa forma, o algoritmo teve como atributo-alvo a resposta para a pergunta "desde março, você precisou consultar um profissional de saúde?", usando como atributos de entrada as informações a respeito do entrevistado e as respostas sobre o período anterior a março de 2020. Os demais atributos que se referem ao período posterior a 2020 foram desconsiderados.

Tais mudanças no conjunto de dados foram realizadas por meio de linguagem *Python 3* (Foundation, 2021) e das bibliotecas NumPy (Harris et al., 2020) e Pandas (https://pandas.pydata.org/), esta última que possibilita tanto a exploração quanto a manipulação dos dados. Posteriormente, todos os atributos nominais da base foram transformados em numéricos por meio da biblioteca Scikit-Learn (Pedregosa et al., 2011), usando o método *labelencoding*, à exceção das perguntas de seleção múltipla, onde os dados foram transformados em binários por meio do método *one-hot-encoding*.

Após a transformação dos atributos, foi feita a normalização dos dados por meio do método *MinMax*, onde a escala de cada uma das variáveis é redimensionada para o intervalo entre 0 e 1, por meio da equação (1):

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

Em seguida, o conjunto foi separado em treino e teste, sendo 20% das instâncias para teste, e logo após, foi feito o balanceamento de classes (descrito no início da seção 4) utilizando o método *Randomover-sampling*, da biblioteca *Imbalanced-Learn* (Lemaître, Nogueira, & Aridas, 2016). Em tal método, instâncias que fazem parte da classe minoritária do conjunto de treino são selecionadas aleatoriamente e duplicadas, de maneira a haver um equilíbrio entre as classes. Dessa forma, ao fim, obteve-se um subconjunto de dados com 20.434 instâncias e 62 atributos.Com o subconjunto pré-processado foi dada sequência ao processo de construção do modelo conforme o fluxograma da Figura 3.

Figura 3. Fluxograma das Etapas de Construção do Modelo.
Fonte: Os autores (2022)

A seguir, com a finalidade de avaliar quais atributos possuem maior impacto sobre a determinação do atributo-alvo e de reduzir o número de variáveis presentes na base de dados, foi utilizado o algoritmo *Extra TreesClassifier*. Este algoritmo é semelhante ao algoritmo Floresta Aleatória, porém, do ponto de vista do viés de variância, a lógica por trás do método é que a aleatoriedade do ponto de corte e do atributo combinados com a média do conjunto deve ser capaz de reduzir a variância mais fortemente do que os esquemas de randomização mais fracos usados por outros métodos baseados em árvores (Géron, 2019).

Os atributos selecionados foram os de importância maior ou igual a 0,02, consistindo em um total de 18 atributos, sendo eles descritos na Tabela 1.

Atributo	Descrição
gender	Gênero
age	Idade
geography	Localização geográfica
financial_situation	Situação financeira
education	Escolaridade
employment_status	Situação empregatícia
ethnicity	Etnia
religion	Religião
gp_hh	Quantas pessoas moram na mesma residência que o entrevistado, incluindo o mesmo
gp_pre_provider_need	De dez-2019 a fev-2020, procurou um provedor de saúde
gp_pre_provider_visit	De dez-2019 a fev-2020, conseguiu ir ao provedor de saúde, não conseguiu, ou, ainda,
gp_pre_provider_num_visit	De dez-2019 a fev-2020, quantas vezes visitou o provedor de saúde
gp_medication	Precisou ou não usar alguma medicação nos últimos seis meses
gp_pre_labor_force	Estava ou não trabalhando de dez-2019 a fev-2020
gp_ppc_00_Nenhum	De dez-2019 a fev-2020, nenhuma condição de saúde levou a procurar atendimento m
gp_ppc_01_Preventive or routine care	De dez-2019 a fev-2020, foi ao provedor de saúde por rotina ou prevenção
gp_mc_00_Nenhum	Não precisou usar medicamento nos últimos seis meses

Tabela 1. Atributos utilizados na base de dados.
Fonte: IHME (2020, tradução livre)

Por fim, para a investigação dos hiperparâmetros, por meio do módulo *GridSearchCV*, da biblioteca *Scikit-Learn*, foram avaliados quais os melhores parâmetros para o conjunto de dados. Os parâmetros testados foram:

- *criterion*: onde é selecionado o critério para avaliar a qualidade de uma divisão, podendo ser gini (para a impureza de *Gini*) ou *entropy* (baseado em entropia);

- *n_estimators*: onde é avaliado o número de árvores da floresta, de maneira que foram testados os estimadores 50, 100, 200 e 250;

- *min_samples_split*: onde se analisa o número mínimo de amostras necessárias às folhas das árvores, sendo avaliados os números 1, 5 e 10 e;

- *min_samples_split*: este que define o número mínimo de amostras necessárias para dividir um nó interno da árvore, sendo testados os números 2, 5 e 10.

Visto que o algoritmo aplica todas as combinações possíveis dos parâmetros desejados, um total de 72 modelos de classificação foram testados.

O desempenho da aplicação da técnica foi avaliado por meio da matriz de confusão e das métricas de acurácia, precisão, *recall* e *f1-score*. Uma matriz de confusão consiste em uma matriz de duas linhas e duas colunas, onde nas linhas estão representadas as categorias reais das instâncias e, nas colunas, as categorias previstas pelo algoritmo, sendo assim possível avaliar o total de erros e acertos para cada categoria, ou seja, número de verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). Dessa forma, acurácia mostra o desempenho do modelo de forma geral, consistindo no total de acertos do algoritmo sobre o total de instâncias, conforme a equação (2).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

Já a precisão avalia o total de verdadeiros positivos em relação ao total de instâncias que o modelo classificou como positivo, verdadeiras e falsas, tal qual mostrado na fórmula (3).

$$Precisão = \frac{VP}{VP + FP}$$

Em contrapartida, o *recall* avalia o total de verdadeiros positivos em proporção ao total de instâncias que deveriam ser classificadas como positivas, elucidado na equação (4).

$$Recall = \frac{VP}{VP + FN}$$

Por último, *f1-score* é a média harmônica entre os dois indicadores, precisão e *recall*, de cálculo mostrado na equação (5).

$$f1 - score = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Visto que cada execução do algoritmo gera uma acurácia diferente (devido aos diferentes conjuntos de treinamento criados em cada análise), para a obtenção de um número que melhor o avalie foi realizada uma validação cruzada *k-fold*, com o parâmetro k igual a 10. Dessa maneira, foram feitos 30 testes, de forma que, em cada um deles, o conjunto de dados é dividido em 10 e a acurácia é calculada individualmente em cada conjunto para a obtenção de um número médio. Além disso, como uma maneira de avaliar se o algoritmo sofre de sobreajuste, ou, do inglês, *overfitting*, foi feita a avaliação do seu desempenho utilizando os dados tanto do conjunto de treino como do conjunto de teste.

RESULTADOS OBTIDOS

Em uma avaliação inicial, onde foi aplicada a técnica Floresta Aleatória na base pré-processada e normalizada com os parâmetros definidos por padrão, chegou-se nos resultados mostrados na Tabela 2 e na matriz de confusão apresentada na Figura 4, onde 0 corresponde aos pacientes que não precisaram procurar atendimento médico e 1 aos pacientes que precisaram procurar atendimento médico.

Categoria	Precisão	Recall	*f1-score*	Suporte
0	0,83	0,90	0,87	2185
1	0,65	0,51	0,57	794
Média	0,74	0,71	0,80	2979
Média ponderada	0,79	0,80	0,72	2979
Acurácia			0,79	2979

Tabela 2. Resultados do teste inicial.
Fonte: Os autores (2022)

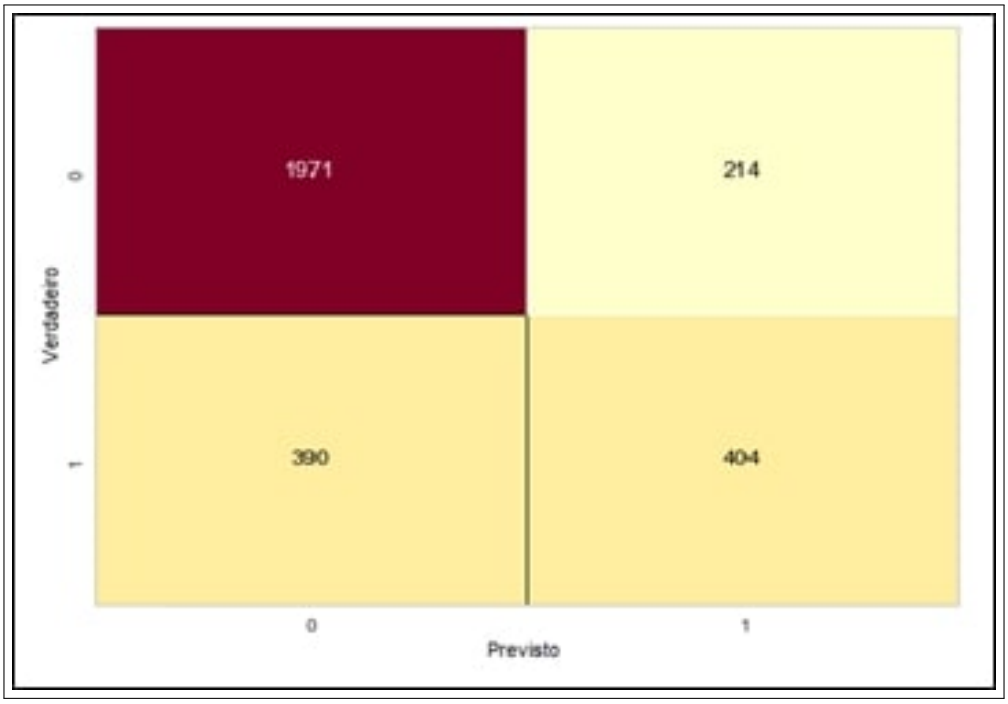

Figura 4. Matriz de confusão.
Fonte: Os autores (2022)

Os resultados iniciais mostraram que o algoritmo possui melhor desempenho em classificar as instâncias quanto à categoria 0 (pessoas que não precisaram procurar atendimento), em detrimento da categoria 1 (pessoas que precisaram procurar atendimento), o que indica um possível desbalanceamento de classes no conjunto de treino, evidenciado também pelo gráfico mostrado na Figura 5.

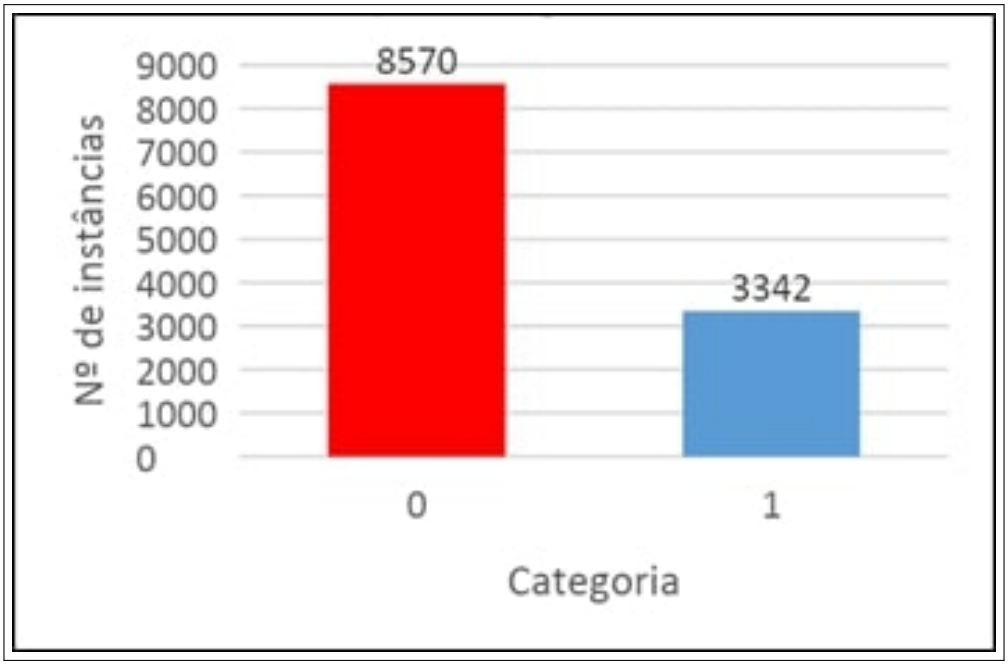

Figura 5. Gráfico do número de instâncias do conjunto de treino por categoria.
Fonte: Os autores (2022)

Dessa maneira, após a aplicação do método *Random over-sampling* para balanceamento de classes, obteve-se os seguintes resultados, mostrados na Tabela 3 e Figura 6. O desempenho do algoritmo se mostrou consistente entre as categorias, com a acurácia evoluindo de 79% para 88%.

Categoria	Precisão	*Recall*	*f1-score*	Suporte
0	0,91	0,85	0,88	2091
1	0,86	0,92	0,89	1996
Média	0,89	0,89	0,88	4087
Média ponderada	0,89	0,88	0,88	4087
Acurácia			0,88	4087

Tabela 3. Resultados após o balanceamento de classes.
Fonte: Os autores (2022)

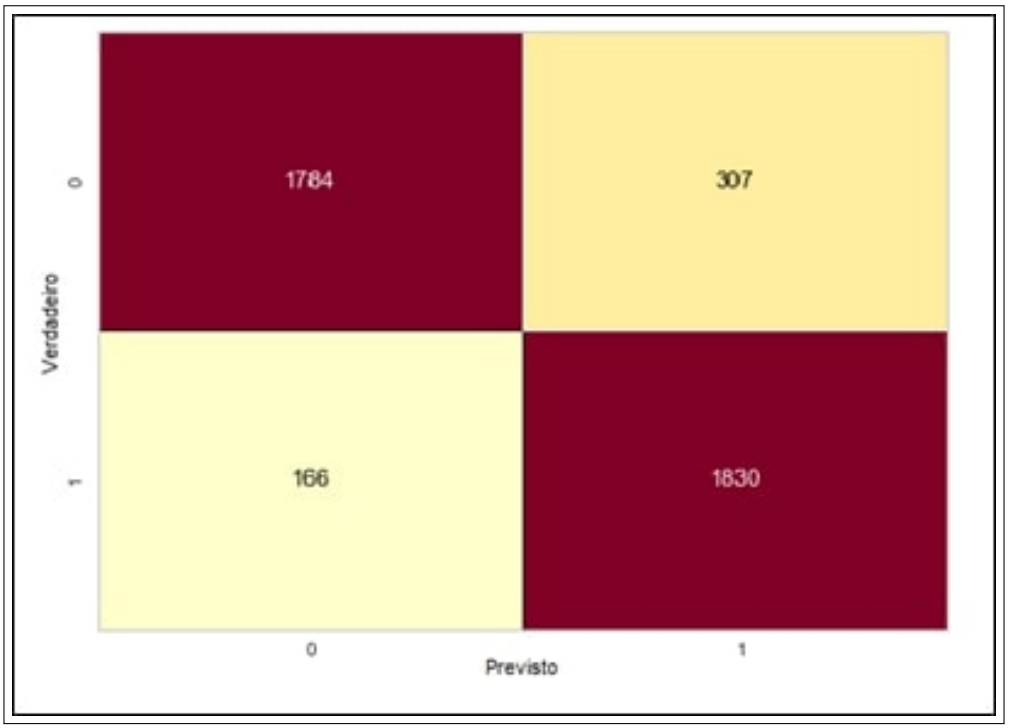

Figura 6. Matriz de confusão após o balanceamento de classes.
Fonte: Os autores (2022)

Em seguida, foram analisados os diferentes parâmetros da Floresta Aleatória (comentados na seção 3) que maximizam a acurácia, sendo estes mostrados na tabela Tabela 4, assim como a acurácia obtida para cada uma das combinações de parâmetros é mostrada no gráfico disposto na Figura 7.

Parâmetro	Valor ou tipo padrão	Valor ou tipo selecionado
criterion	*Gini impurity*	*Entropy*
n_estimators	100	250
min_samples_leaf	1	1
min_samples_split	2	2

Tabela 4. Parâmetros selecionados após avaliação de hiperparâmetros.
Fonte: Os autores (2022)

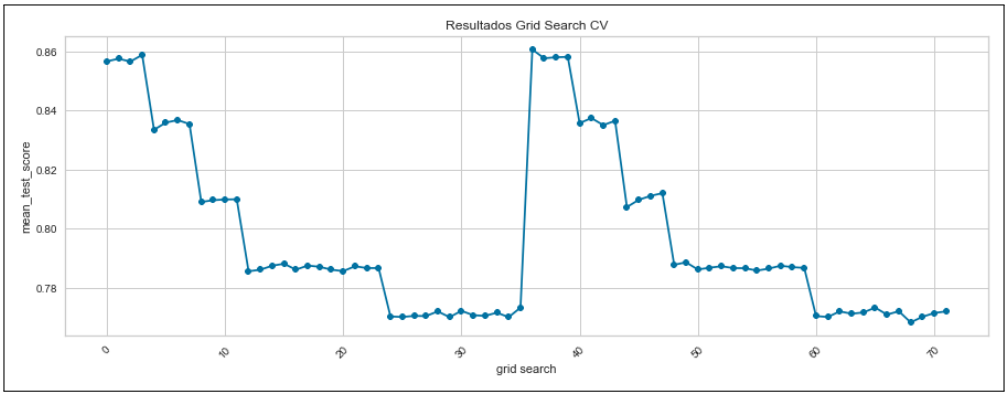

Figura 7. Gráfico da acurácia em cada combinação de parâmetros.
Fonte: Os autores (2022)

A acurácia máxima obtida por meio da otimização de hiperparâmetros foi de 87% (Figura 7), praticamente sem alteração se comparado aos resultados após o balanceamento de classes. Após a aplicação dos parâmetros obtidos utilizando o conjunto de teste para avaliação, chegou-se nos resultados mostrados na Tabela 5 e na matriz de confusão exibida na Figura 8.

Categoria	Precisão	*Recall*	*f1-score*	Suporte
0	0,93	0,84	0,88	2091
1	0,84	0,93	0,88	1996
Média	0,89	0,89	0,88	4087
Média ponderada	0,89	0,88	0,88	4087
Acurácia			0,88	4087

Tabela 5. Resultados após a aplicação dos melhores parâmetros encontrados.
Fonte: Os autores (2022)

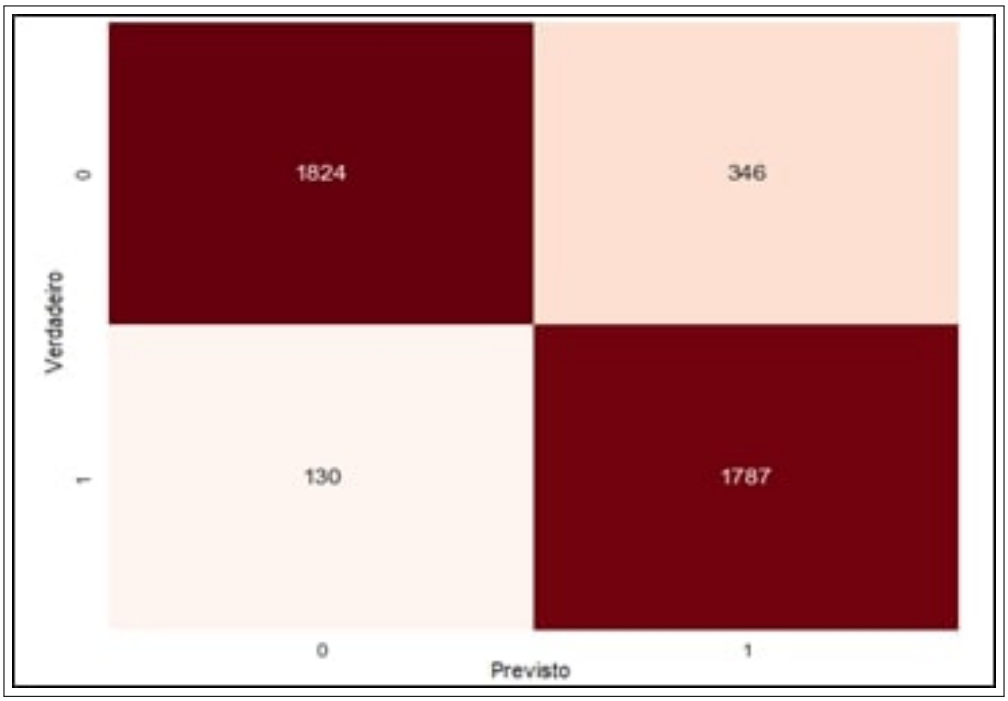

Figura 8. Matriz de confusão após a aplicação dos parâmetros obtidos.
Fonte: Os autores (2022)

Por fim, foi feito o teste do desempenho do algoritmo usando a técnica da validação cruzada fazendo-se uso dos dados do conjunto de treino, cujos resultados são mostrados na Tabela 6 e na matriz de confusão disposta na Figura 9.

Valor	Acurácia	Precisão	*Recall*
número de testes	30	30	30
média	0,870541	0,834732	0,906273
desvio padrão	0,001650	0,001562	0,002131
mínimo	0,867621	0,831947	0,902165
mediana	0,870435	0,834410	0,907180
máximo	0,873555	0,838175	0,911718

Tabela 6. Resultados após validação cruzada utilizando o conjunto de treinamento.
Fonte: Os autores (2022)

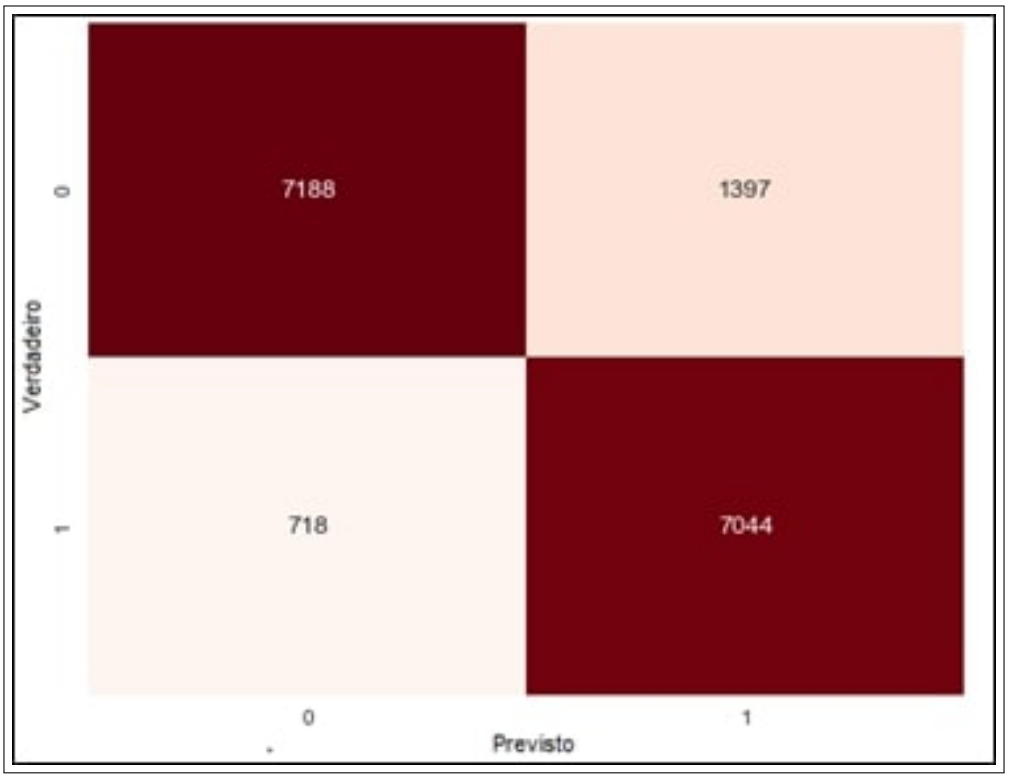

Figura 9. Matriz de confusão após validação cruzada usando dados de treino.
Fonte: Os autores (2022)

Diante dos resultados obtidos, é possível observar que a acurácia obtida com os dados de treino e com os dados de teste são parecidos, respectivamente 88% e 87,5%. Com isso, o modelo não sofre de sobreajuste, mantendo o mesmo desempenho quanto aos dados conhecidos (treino) e também quanto aos dados desconhecidos (teste).Por fim,a Figura 10 ilustra os parâmetros selecionados para compor o subconjunto de dados por meio do método *Extra TreesClassifier*.

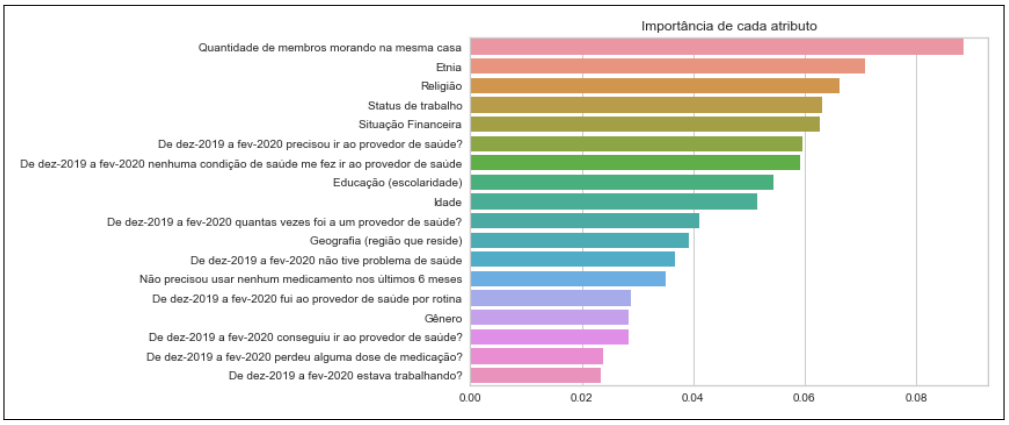

Figura 10. Importância de cada atributo selecionado.
Fonte: Os autores (2022)

Nota-se pela Figura 10 que os cinco atributos mais importantes escolhidos pelo método estão ligados com dados

socioeconômicos.

DISCUSSÕES

Do ponto de vista da metodologia, os resultados obtidos sugerem que o conjunto em seu formato original, com um total de 8.570 instâncias que não precisaram visitar um profissional da saúde e 3.342 que precisaram, o modelo teve um desempenho razoável no geral, com 79% de acurácia, porém foi incapaz de identificar as pessoas que de fato precisaram (51% de sensibilidade). Desta forma, a classe-alvo (classe "1") não era compreendida pelo modelo, provavelmente em função da menor quantidade de instância na etapa de treinamento. Em contrapartida, o modelo conseguiu acertar 90% das pessoas que não buscaram por ajuda, com uma precisão de 83%. Assim, surgiu a necessidade de se buscar alternativas para que os modelos pudessem compreender de uma maneira mais assertiva a classe de interesse deste artigo.

A partir do balanceamento e depois da investigação de hiperparâmetros, houve um aumento não só no valor da acurácia, pois também conseguiram atingir melhores resultados para o *recall* e precisão, chegando a mais de 90% nestas duas métricas, contrastando com a queda obtida no *recall* para a classe contrária (classe "0"). Houve um cuidado ao se avaliar estes resultados, pois como o método de *oversampling* utilizado foi baseado na reamostragem com reposição, podendo levar o modelo ao sobreajuste. Entretanto, a análise feita com as métricas utilizando os conjuntos de treino e teste sugeriram que não existiu esta condição, destacando-se assim o bom poder preditivo dos modelos com *oversampling*, não parecendo ocorrer um alto viés ou variância dos resultados. Neste sentido, a metodologia aqui aplicada tem uma contribuição prática relevante, pois uma investigação inicial sobre este conjunto de dados não encontrou nenhum trabalho que explorou esta base com o interesse particular sobre as pessoas que precisaram visitar algum profissional da saúde.

Verificando-se os atributos mais importantes (Tabela 10), a quantidade de pessoas morando na mesma casa pode ser um fator crucial na procura ou necessidade de assistência médica. Para uma pessoa que possui responsabilidades de gerenciamento familiar, a dedicação às pessoas que necessitam de cuidados ou supervisão pode acabar aumentando a necessidade de procura por profissionais da saúde, seja por questão física ou psicológica. Para o caso da pandemia, o distanciamento social se torna desafiador uma vez que o vírus se espalha mais facilmente em locais fechados e com um grande número de pessoas. Uma revisão sistemática feita por Madewell, Yang, Longini Jr, Halloran, e Dean (2020) indicou que, em ambientes fechados, muitas pessoas podem ser infectadas ou transmitir o vírus e, quando infectadas, muitas vezes ficam em isolamento domiciliar, ainda representando ameaças na dispersão do vírus. Assim, estas condições podem elevar a necessidade de procurar por atendimentos.

Quanto à etnia e raça, o Centro para Controle e Prevenção de Doenças *(Centers for Disease Control and Prevention* CDC (2020)) afirma que há uma grande disparidade quanto aos aspectos econômicos, sociais e consequências da COVID-19, afetando de forma diferente os grupos raciais e étnicos minoritários. A partir de informações sobre imunização, verificou-se que cinco fatores sociais são determinantes para estes grupos minoritários, como: vizinhança e ambiente físico; saúde e sistemas de saúde; ocupação e condições de trabalho; renda; e educação. Cada um destes fatores pode aumentar o risco de exposição à COVID-19, doenças, hospitalização, saúde a longo prazo e consequências sociais, e morte (CDC, 2020).

Em relação à religião, existe uma preocupação em termos. Schnabel (2021) noticiou sobre a importância da religião em tempos de crise, principalmente na pandemia do novo coronavírus. Em uma análise com 12 mil americanos, a pesquisa conduzida pelo mesmo autor mostrou que a saúde mental foi menos impactante entre religiosos americanos, especialmente evangélicos. Em contrapartida, estas mesmas pessoas que tiveram um menor estresse mental com a COVID-19, também se preocuparam menos com as medidas de saúde pública e o perigo com as consequências físicas causadas pelo vírus (Schnabel, 2021).

Dessa maneira, a aplicação das técnicas mostradas nesses estudos poderiam ser de grande utilidade para o planejamento e gestão de um sistema de saúde, tal qual o SUS (Sistema Único de Saúde). A partir de dados simples, muitos dos quais já presentes na base de dados governamental, é possível identificar grupos e regiões cuja ação dos órgãos públicos de saúde é crítica, o que possibilita a adoção de medidas preventivas direcionadas a esses grupos, como campanhas educacionais, que visam permitir que, por meio da informação, evite-se o agravamento do quadro epidemiológico. Ressalta-se também que a utilidade do algoritmo como guia de políticas governamentais depende também do teste de seu desempenho com populações de diferentes características, pois a acurácia pode sofrer alterações com dados de teste de populações muito diferentes das aqui utilizadas para treino, sendo este um ponto a ser avaliado em estudos futuros.

CONCLUSÕES

O presente artigo objetivou a aplicação de um modelo de classificação, a partir do algoritmo Floresta Aleatória, para classificar o comportamento da população na busca por serviços de saúde após o início da pandemia de Covid-

19. Os dados explorados neste trabalho provêm dos resultados da pesquisa *Premise General Population* COVID-19 *Health Services DisruptionSurvey* 2020, oriunda do projeto COVID-19 *Health Services DisruptionSurvey* 2020.Após a seleção dos dados, pré-processamento e aplicação do algoritmo de mineração de dados, o melhor resultado encontrado foi com uma acurácia aproximada de 88% com dados desconhecidos (conjunto de teste). Quanto ao *recall*, o modelo conseguiu alcançar um valor de 0,89 em média, sendo o mesmo valor para a precisão, mostrando que ele é eficiente tanto na detecção de pessoas que buscaram cuidados médicos, quanto daqueles que não buscaram. A média harmônica da precisão e *recall* foi de 0,88 para as duas classes.

Com relação aos atributos mais importantes, foram encontrados os atributos (em ordem de relevância para o modelo): "Quantidade de membros morando na mesma casa"; "Etnia"; "Raça"; "Status de trabalho"; e "Situação financeira". Desta forma, estudos específicos e aprofundados nestas variáveis podem ser realizados para se encontrar fatores também qualitativos, ampliando o debate sobre questões demográficas e a relação com a procura ou necessidade de sistemas de saúde.

Os modelos de classificação pesquisados mostraram-se eficazes em predizer o comportamento dos pacientes em buscar atendimento de saúde em um futuro próximo com base em seus dados sobre estilo de vida, características pessoais e comportamentos associados à saúde. Como sugestões para estudos futuros, indica-se investigar outros algoritmos que possam contribuir para a construção de modelos classificadores ainda mais eficientes, focando principalmente em seus hiperparâmetros, como também nos atributos mais relevantes encontrados pela Floresta Aleatória.

REFERÊNCIAS

Aggarwal, C. (2015). *Data mining.* California: Springer.

Armocida, B., Formenti, B., Ussai, S., Palestra, F., & Missoni, E. (2020). The italian health system and the covid-19 challenge. *The Lancet Public Health*, *5*(5), e253. doi: 10.1016/S2468-2667(20)30074-8

Azambuja, C. (2014). Importância das medidas de gestão hospitalar no controle da superlotação hospitalar [TCC]. *Universidade Federal de Santa Maria*. Recuperado de https://repositorio.ufsm.br/bitstream/handle/1/11730/Azambuja_Claudio_Roberto_Carvalho_de.pdf?sequence=1&isAllowed=y

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. doi: 10.1007/9781441993267_5

CDC. (2020). *Introduction to covid-19 racial and ethnic health disparities.* Recuperado de https://www.cdc.gov/healthequity/whatis/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcommunity%2Fhealth-equity%2Fracial-ethnic-disparities%2Findex.html

Chen, P., Wu, Y., Zhong, H., Long, Y., & Meng, J. (2022). Exploring household emission patterns and driving factors in japan using machine learning methods. *Applied Energy*, *307*, 118251. doi: 10.1016/j.apenergy.2021.118251

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-53. doi: 10.1609/aimag.v17i3.1230

Foundation, P. S. (2021). *Python.* Recuperado de https://www.python.org/

Gil, P. D., da Cruz Martins, S., Moro, S., & Costa, J. M. (2021). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, *26*(2), 2165-2190. doi: 10.1007/s10639-020-10346-6

Goldschmidt, R., Passos, E., & Bezerra, E. (2015). *Data mining: conceitos, técnicas, algoritmos, orientações e aplicações.* Rio de Janeiro: Elsevier.

Géron, A. (2019). *Mãos à obra: aprendizado de máquina com scikit-learn tensor flow.* Rio de Janeiro: Alta Books.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with numpy. *Nature*, *585*(7825), 357-362. doi: 10.1038/s41586-020-2649-2

IHME. (2020). *Premise general population covid-19 health services disruption survey 2020.* Recuperado de https://ghdx.healthdata.org/record/ihme-data/premise-general-population-covid-19-health-services-disruption-survey-2020

Islam, M., Rahman, M., Islam, M., Roy, D., Ahmed, N., Hussain, S., ... Maniruzzaman, M. (2022). Application of machine learning based algorithm for prediction of malnutrition among women in bangladesh. *International Journal of Cognitive Computing in Engineering*, *3*, 46-57. doi: 10.1016/j.ijcce.2022.02.002

Larose, D., & Larose, C. (2014). *Discovering knowledge in data.* New Jersey: John Wiley Sons.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2016). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *7*, 1-5. doi: 10.48550/arXiv.1609.06570

Madewell, Z. J., Yang, Y., Longini Jr, I. M., Halloran, M. E., & Dean, N. E. (2020). Household transmission of sars-cov-2: a systematic review and meta-analysis. *JAMA Network Open*, *3*(12), e2031756-e2031756. doi: 10.1001/jamanetworkopen.2020.31756

Naderpour, H., Mirrashid, M., & Parsa, P. (2021). Failure mode prediction of reinforced concrete columns using machine learning methods. *Engineering Structures*, *248*, 113263. doi: 10.1016/j.engstruct.2021.113263

Pan, C., Poddar, A., Mukherjee, R., & Ray, A. K. (2022). Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. *Biomedical Signal Processing and Control*, *76*, 103666. doi: 10.1016/j.bspc.2022.103666

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Raschka, S. (2015). *Python machine learning.* Reino Unido: Packt Publishing.

Santos, B. S. d., Steiner, M. T. A., Fenerich, A. T., & Lima, R. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers and Industrial Engineering*, *138*, 106120. doi: 10.1016/j.cie.2019.106120

Schnabel, L. (2021). *Religion both helped and hurt during the pandemic.* Recuperado de https://www.scientificamerican.com/article/religion-both-helped-and-hurt-during-the-pandemic/

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074. doi: 10.1016/j.mlwa.2021.100074

Tang, W., Wang, H., Lee, X.-L., & Yang, H. (2022). Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, *240*, 122500. doi: 10.1016/j.energy.2021.122500

Tessoni, V., & Amoretti, M. (2022). Advanced statistical and machine learning methods for multi-step multivariate time series forecasting in predictive maintenance. *Science*, *200*, 748-757. doi: 10.1016/j.procs.2022.01.273

Como citar este artigo (APA):
Ariza, V. M. P., Nascimento, M. M., Malandrino, P. P., Santos, B. S. (2022). Uso do algoritmo "Floresta Aleatória" na identificação do comportamento da população na busca por serviços de saúde após o início da pandemia do novo coronavírus. *AtoZ: novas práticas em informação e conhecimento, 11*, 1 – 15. Recuperado de: http://dx.doi.org/10.5380/atoz.v11.84017

NOTAS DA OBRA E CONFORMIDADE COM A CIÊNCIA ABERTA

CONTRIBUIÇÃO DE AUTORIA

Papéis e contribuições	Vinicius Matheus Pimentel Ariza	Mateus Miranda do Nascimento	Pedro PicoloMalandrino	Bruno Samways dos Santos
Concepção do manuscrito	X		X	
Escrita do manuscrito	X	X	X	X
Metodologia		X		X
Curadoria dos dados	X			
Discussão dos resultados	X	X	X	X
Análise dos dados	X	X	X	X

EQUIPE EDITORIAL