Análise Preditiva da Escassez de Profissionais de Saúde Mental nos Estados Unidos

Antonio Borssato
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
antonio.borssato@edu.ufes.br

Lucas Alves
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
lucas.o.alves@edu.ufes.br

Rodrigo Fardin
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
rodrigo.fardin@edu.ufes.br

***Resumo* — Este estudo investiga a escassez de profissionais de saúde mental nos Estados Unidos e seu impacto nos serviços essenciais. Com dados da HRSA, o objetivo foi prever a pontuação HPSA utilizando variáveis demográficas e de atendimento. Indicadores socioeconômicos auxiliares foram inicialmente considerados, mas excluídos devido à elevada taxa de valores ausentes. Foram testados três modelos de classificação – Random Forest, Decision Tree e KNN – otimizados por GridSearchCV com validação cruzada. O modelo Random Forest apresentou melhor desempenho, com acurácia de aproximadamente 84,52%. Os resultados demonstram a viabilidade da análise preditiva para orientar a alocação de recursos e embasar políticas de saúde.**

***Palavras-chave* — Saúde mental, HPSA, aprendizado de máquina, validação cruzada, GridSearchCV, análise preditiva.**

I. Introdução

A crescente demanda por serviços de saúde mental nos Estados Unidos, aliada à limitada disponibilidade de profissionais especializados, representa um desafio para o sistema de saúde do país. Regiões com deficiência desses profissionais podem apresentar taxas elevadas de transtornos mentais não tratados, sobrecarregando as unidades de emergência e comprometendo a qualidade de vida da população. Neste contexto, o uso de técnicas de análise preditiva e aprendizado de máquina tem se mostrado fundamental para identificar padrões e priorizar intervenções. Este estudo utiliza dados fornecidos pela Health Resources and Services Administration (HRSA) para modelar e prever a pontuação HPSA – um indicador da necessidade de alocação de profissionais – especificamente na área de saúde mental.

II. Motivação

O problema central deste estudo é a escassez de profissionais de saúde mental, que impacta o acesso a serviços essenciais e agrava os índices de transtornos não tratados. Essa carência, evidenciada pelo mapa de escassez da Fig. 1, onde as regiões em azul escuro indicam maior déficit (e maior HPSA Score), gera uma demanda reprimida que pode resultar em maior utilização de serviços de emergência, menor acompanhamento de pessoas que precisam de cuidados e piora nos indicadores de saúde populacional. Assim, o objetivo deste trabalho é prever a pontuação HPSA, que varia de 0 a 25, utilizando dados históricos e características regionais, possibilitando a identificação de áreas com maior necessidade de intervenção.

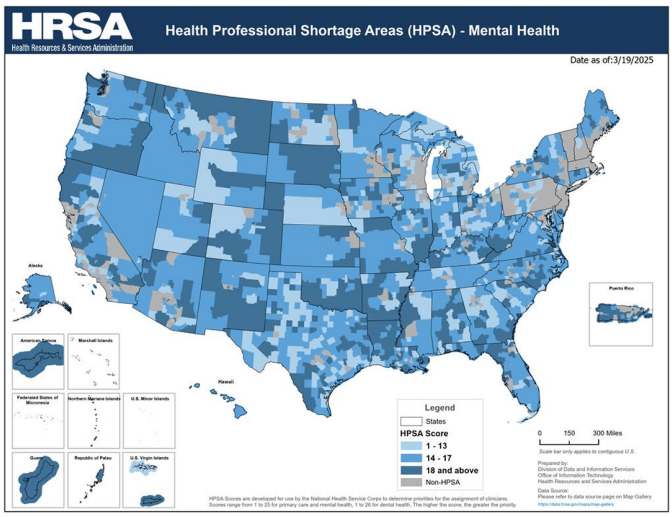

Figura 1. Mapa de Escassez de Profissionais de Saúde Mental nos Estados Unidos.

III. Revisão da Literatura

A escassez de profissionais de saúde nos Estados Unidos tem sido foco de estudos recentes. Ramezani et al. [4] investigaram em 2023 o impacto da infraestrutura de saúde em nível de condado num programa voltado para a redução de encarceramento de indivíduos com distúrbios mentais. Nesse trabalho os autores discorrem, entre outras coisas, sobre a capacidade dos serviços comunitários de saúde mental, apontando que dos 3.141 condados dos EUA, 1.025 (33%) são áreas com escassez de profissionais de saúde mental e 1.461 (45%) são designados como regiões medicamente desassistidas. Rochford et al. [5] aprofundam essa visão, mostrando que 1 a cada 3 pessoas nos EUA vivem em área com escassez de profissionais da saúde mental, além de apresentarem que essas áreas tendem a ser rurais e com níveis elevados de pobreza. Jamal [6] coloca, por sua vez, o problema como uma crise severa no país, onde 1 a cada 5 adultos experiencia algum transtorno mental no ano e 1 a cada 25 adultos vive com transtorno psicológico grave. No estudo Jamal utiliza machine learning para avaliar o efeito

da telemedicina no gasto médico e na busca por utilização do sistema de saúde, onde ele conclui que a telemedicina pode estar associada ao aumento na procura por cuidados ambulatoriais, mas não a um aumento no gasto médico.

Nota-se, portanto, que os dados geográficos e socioeconômicos apresentados nos estudos citados podem ser variáveis determinantes em modelos preditivos da pontuação HPSA. Felizmente os dados disponibilizados pela HRSA condensam justamente tais informações.

IV. Conjunto de Dados

A base de dados utilizada provém da HRSA e está disponível em HRSA Data. Para a análise, foram utilizados os principais conjuntos de dados:

- All HPSAs – CSV.csv: Contém registros relacionados às áreas de escassez de profissionais, com foco na saúde mental.
- MUA/P – CSV.csv: Lista áreas e populações medicamente carentes.

O conjunto de dados sobre saúde mental possui 34.940 registros e 65 colunas, dentre as quais destacam-se:

HPSA Score: Pontuação atribuída pelo National Health Service Corps, que define a prioridade na alocação dos profissionais.

HPSA Discipline Class: Classificação que, neste estudo, focaliza os serviços de saúde mental.

HPSA Designation Population, HPSA FTE e HPSA Shortage: Variáveis que indicam, respectivamente, a população afetada, o número de profissionais em tempo integral e a necessidade de profissionais adicionais para atingir a razão ideal de atendimento.

Além disso, o conjunto contém informações geográficas e demográficas (por exemplo, indicadores de pobreza e dados de áreas rurais ou urbanas), que fornecem contexto para a análise, mas que, em muitos casos, necessitam de pré-processamento para remoção de redundâncias e informações não preditivas.

Em relação ao conjunto MUA/P, algumas das variáveis de interesse para esse estudo são: Percent of Population with Incomes at or Below 100 Percent of the U.S. Federal Poverty Level, Percentage of Population Age 65 and Over, Infant Mortality Rate, Designation Population in a Medically Underserved Area/Population (MUA/P), Medically Underserved Area/Population (MUA/P) Total Resident Civilian Population e Providers per 1000 Population. Entretanto, a quantidade de valores ausentes nessa base, às vezes acima de 68%, mostrou-se um problema de qualidade dos dados que será comentado nos resultados desse estudo.

V. Metodologia

A abordagem metodológica deste estudo foi estruturada nas seguintes etapas:

A. Pré-processamento e Limpeza de Dados

1) Remoção de identificadores e informações irrelevantes: Foram eliminadas colunas que não possuem valor preditivo, como HPSA ID, BHCMIS Organization Identification Number e HPSA Component Source Identification Number. Informações descritivas (exemplo a HPSA Name, HPSA Address e HPSA City) e códigos redundantes (como os códigos de status e classificação) foram igualmente removidos para evitar sobrecarga no modelo.

2) Filtragem Temporal: Observou-se que até 2008 a pontuação HPSA era zero (como se pode observar na Fig. 2), o que indica falta de variabilidade e relevância para a modelagem. Dessa forma, registros anteriores a 2009 foram descartados, mantendo apenas dados com pontuações significativas.

3) Tratamento de Valores Ausentes: Foram avaliadas as colunas com alta taxa de valores ausentes (acima de 40%). Para variáveis numéricas, considerou-se a imputação por média ou mediana (foi avaliado a distribuição dos dados para escolher qual medida estatísticas utilizar), enquanto para variáveis categóricas a estratégia foi preencher com a moda. Caso o preenchimento não garantisse a integridade dos dados, a coluna foi removida. Assim, colunas como Withdrawn Date, Longitude, Latitude, HPSA Resident Civilian Population e HPSA Formal Ratio foram removidas por possuíram ausência de valores superior a 40% e por não ser possível imputar valores sem a introdução de algum viés por se tratar de dados específicos.

B. Transformação de Dados

1) One-Hot Encoding: Para variáveis categóricas, como por exemplo ”HPSA Status”e ”Metropolitan Indicator”, foram transformadas em variáveis binárias para evitar interpretações incorretas de hierarquia e garantir que os dados categóricos possam ser usados com eficácia em modelos de aprendizado de máquina.

2) Padronização dos Dados: Aplicou-se a padronização (utilizando o StandardScaler da biblioteca Scikit-learn) às variáveis numéricas, garantindo que todas tenham média zero e desvio padrão igual a 1. Essa etapa garante que todos os recursos estejam na mesma escala, evitando que qualquer recurso domine o processo de aprendizado devido à sua maior magnitude.

C. Modelagem Preditiva

O objetivo foi prever a pontuação HPSA com base nas variáveis preditivas. A modelagem envolveu a divisão dos dados em conjuntos de treinamento (70%) e teste (30%) e a utilização de GridSearchCV com validação cruzada (StratifiedKFold) para a otimização dos modelos. Os modelos escolhidos são os seguintes:

- **Random Forest (RandomForestClassifier):** Esse modelo foi utilizado porque ele considera naturalmente as interações entre variáveis e é útil para dados categóricos e codificados por um hot e rótulo.
- **Decision Tree (DecisionTreeClassifier):** Foi utilizado por proporcionar uma alta interpretabilidade e execução rápida, facilitando a compreensão dos padrões decisórios.
- **K-Nearest Neighbors (KNeighborsClassifier):** Foi utilizado por ser um modelo simples e eficiente, capaz de captar relações locais, dependendo da escolha do número de vizinhos e da métrica de distância.

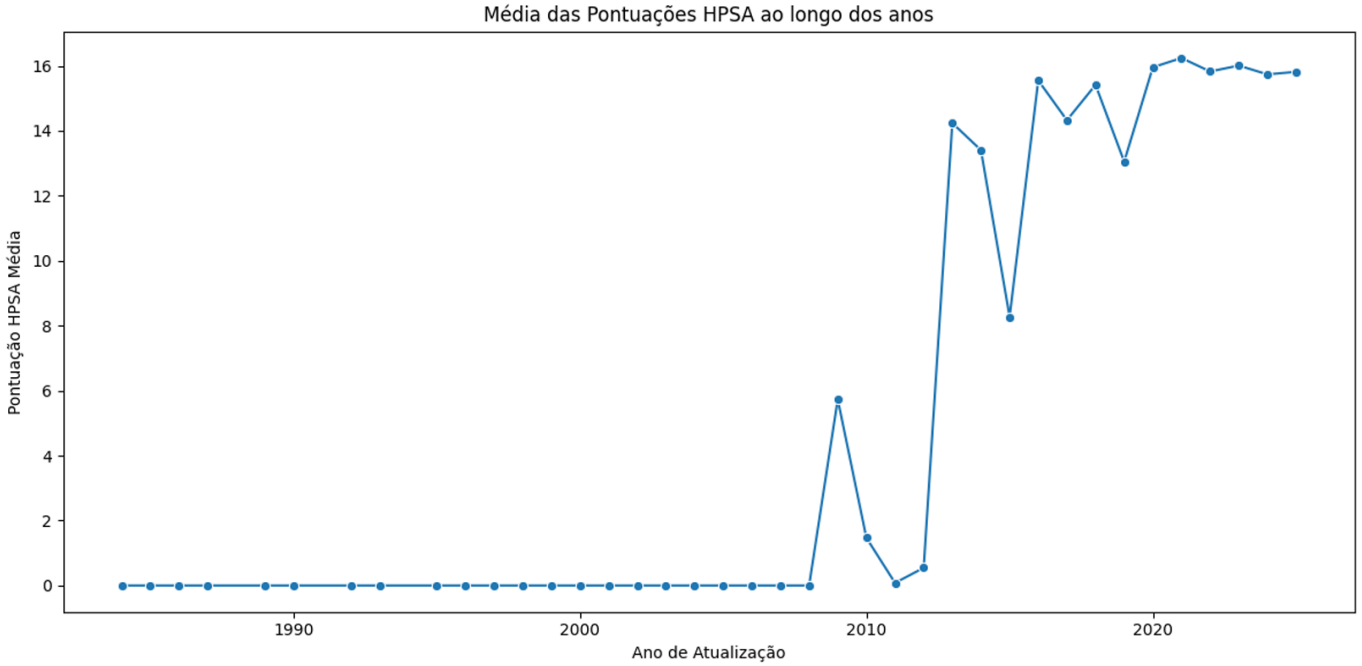

Figura 2. Média das pontuações HPSA ao longo dos anos.

1) Otimização de Hiperparâmetros: O GridSearchCV testa exaustivamente todas as combinações de hiperparâmetros, garantindo a melhor configuração para maximizar o desempenho do modelo. Para o *Random Forest*, buscou-se otimizar parâmetros como *n_estimators*, *max_depth* e *min_samples_split*. No caso do *Decision Tree*, foram avaliados *max_depth*, *min_samples_split* e *criterion*, enquanto para o *KNN* os parâmetros *n_neighbors*, *weights* e *metric* foram testados.

2) Validação Cruzada: Utilizamos a validação cruzada de k vezes (foi utilizado um $k = 5$) para avaliar a generalização de nossos modelos. A validação cruzada escolhida foi o *StratifiedKFold*, que assegura que cada subconjunto seja representativo da distribuição original dos dados, reduzindo o risco de *overfitting* e aumentando a confiabilidade da avaliação.

D. Métricas de Avalição

As principais métricas utilizadas para avaliar o desempenho dos modelos preditivos aplicados neste estudo foram:

1) Acurácia: Representa a proporção de previsões corretas em relação ao total de previsões. Levando em consideração que as pontuações HPSA apresentam desbalanceamento das classes, somente essa métrica é insuficiente.

2) Precisão (Precision): Indica a proporção de verdadeiros positivos entre todas as instâncias classificadas como positivas. Esta métrica é util quando os custos de falsos positivos são elevados, pois reflete a exatidão das previsões positivas do modelo.

3) Recall (Sensibilidade): Mede a capacidade do modelo em identificar corretamente todas as instâncias positivas. É importante quando a detecção de casos positivos é crítica, mesmo que isso implique em um aumento de falsos positivos.

4) F1-Score: A média harmônica entre precisão e recall, o F1-Score fornece uma medida balanceada da performance do modelo. Esta métrica é útil quando se deseja um equilíbrio entre precisão e recall, principalmente em situações onde há um *trade-off* entre as duas.

Estas métricas foram empregadas para comparar o desempenho dos três modelos testados (Random Forest, Decision Tree e KNN), permitindo uma avaliação detalhada e fundamentada na eficácia de cada abordagem para a previsão da pontuação HPSA.

VI. Resultados

Após as etapas de pré-processamento, transformação e modelagem, os modelos aplicados demonstraram a viabilidade de prever a pontuação HPSA com desempenho satisfatório.

A. Desempenho dos Modelos

A seguir, apresentamos os resultados dos três modelos testados e os hiperparâmetros usados:

Tabela I
Desempenho dos Modelos de Classificação

Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	0.845151	0.845080	0.845151	0.844647
Decision Tree	0.835824	0.836981	0.835824	0.836072
K-Nearest Neighbors	0.734075	0.730905	0.734075	0.731681

Melhores hiperparâmetros dos resultados da Tabela I:

- **Random Forest**: `max_depth=None`, `min_samples_split=2`, `n_estimators=100`
- **Decision Tree**: `criterion='gini'`, `max_depth=None`, `min_samples_split=2`

- **K-Nearest Neighbors**: `metric='manhattan', n_neighbors=7, weights='distance'`

Os modelos Random Forest e Decision Tree demonstraram desempenho superior (com uma ligeira superioridade do Random Forest), com acurácias acima de 83%, em contraste com o KNN, que apresentou resultados inferiores (acurácia de 73,41%). Essa diferença sugere que modelos baseados em árvores são mais eficazes para captar as relações complexas entre as variáveis preditivas neste contexto, como pode ser visto na Fig. 3.

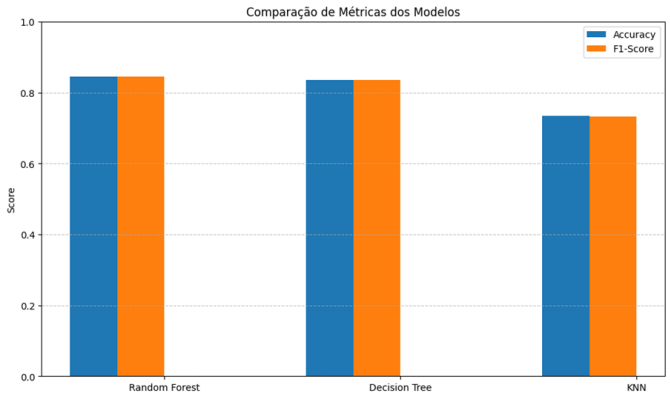

Figura 3. Comparação de Métricas dos Modelos.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	361
3	0.73	1.00	0.85	11
4	0.00	0.00	0.00	1
5	0.75	0.60	0.67	10
6	0.74	0.78	0.76	40
7	0.95	0.83	0.89	47
8	0.87	0.82	0.84	56
9	0.85	0.79	0.82	130
10	0.79	0.72	0.75	145
11	0.92	0.90	0.91	429
12	0.87	0.84	0.85	580
13	0.87	0.84	0.86	496
14	0.85	0.85	0.85	581
15	0.89	0.88	0.88	869
16	0.86	0.86	0.86	1350
17	0.84	0.88	0.86	1609
18	0.87	0.86	0.86	1338
19	0.84	0.83	0.83	743
20	0.74	0.76	0.75	369
21	0.44	0.53	0.48	171
22	0.30	0.19	0.23	67
23	0.21	0.21	0.21	24
24	0.00	0.00	0.00	8
25	0.00	0.00	0.00	0
accuracy			0.85	9435
macro avg	0.67	0.66	0.67	9435
weighted avg	0.84	0.85	0.84	9435

Figura 4. Classification Report – Random Forest.

B. Análise do Classification Report – Random Forest

Como o Random Forest foi o melhor modelo, analisamos detalhadamente seu classification_report (evidenciado na Fig. 4):

- Classes com alta representatividade (e.g classes 0 e 11 ao 19) apresentaram valores de precisão, recall e f1-score superiores a 0.83, indicando que o modelo classificou corretamente a maioria dos casos nessas classes.
- Classes menos representadas (e.g., classes 3, 5, 7, 8, 9, 10, 12, 13) também demonstraram bom desempenho, com f1-scores entre 0.75 e 0.91.
- Algumas classes com suporte muito baixo (e.g., classes 4, 22 ao 25) apresentaram resultados nulos ou baixos, o que é esperado dada a escassez de amostras para treinamento.
- No geral, a acurácia global do modelo foi de 85%, com médias macro de precisão, recall e f1-score em torno de 0.67 e médias ponderadas próximas a 0.84–0.85, reforçando a robustez do modelo para a maioria dos casos.
- Essas métricas demonstram que o modelo Random Forest é capaz de classificar com eficácia as diferentes pontuações HPSA, evidenciando a relevância dos indicadores utilizados e a capacidade do modelo de lidar com a heterogeneidade dos dados.

C. Interpretação das Importância das Características

Usando nosso modelo Random Forest, conseguimos traçar as importâncias das características, conforme visto na Fig. 5. Podemos obverser que a HPSA Designation Population é a variável mais relevante, seguida da HPSA Shortage e HPSA Estimated Underserved Population também apresentam forte influência. Outros indicadores, como % of Population Below 100% Poverty e HPSA FTE, também contribuíram para a determinação das pontuações HPSA. Essa análise ressalta que tanto as características demográficas quanto as informações de atendimento possuem papel crucial na modelagem, justificando a estratégia de incluir variáveis que refletem a demanda e a capacidade de atendimento em áreas com escassez de profissionais de saúde mental.

VII. Discussão

Este estudo evidenciou que a análise preditiva, combinada com técnicas de pré-processamento e transformação de dados, pode oferecer insights relevantes sobre a escassez de profissionais de saúde mental nos Estados Unidos. Entre as principais conclusões, destacam-se:

- **Qualidade dos Dados:** A remoção de informações redundantes e o tratamento cuidadoso dos valores ausentes foram essenciais para a construção de modelos robustos. A tentativa de incorporar dados socioeconômicos do conjunto MUA/P foi inviabilizada pela alta taxa de valores ausentes neste conjunto de dados (acima de 68% e 73% em colunas críticas), comprometendo a representatividade e a confiabilidade das análises. Assim, optou-se por focar nas variáveis com dados completos.
- **Aplicabilidade dos Modelos Preditivos:** A capacidade dos modelos em prever com precisão a pontuação HPSA

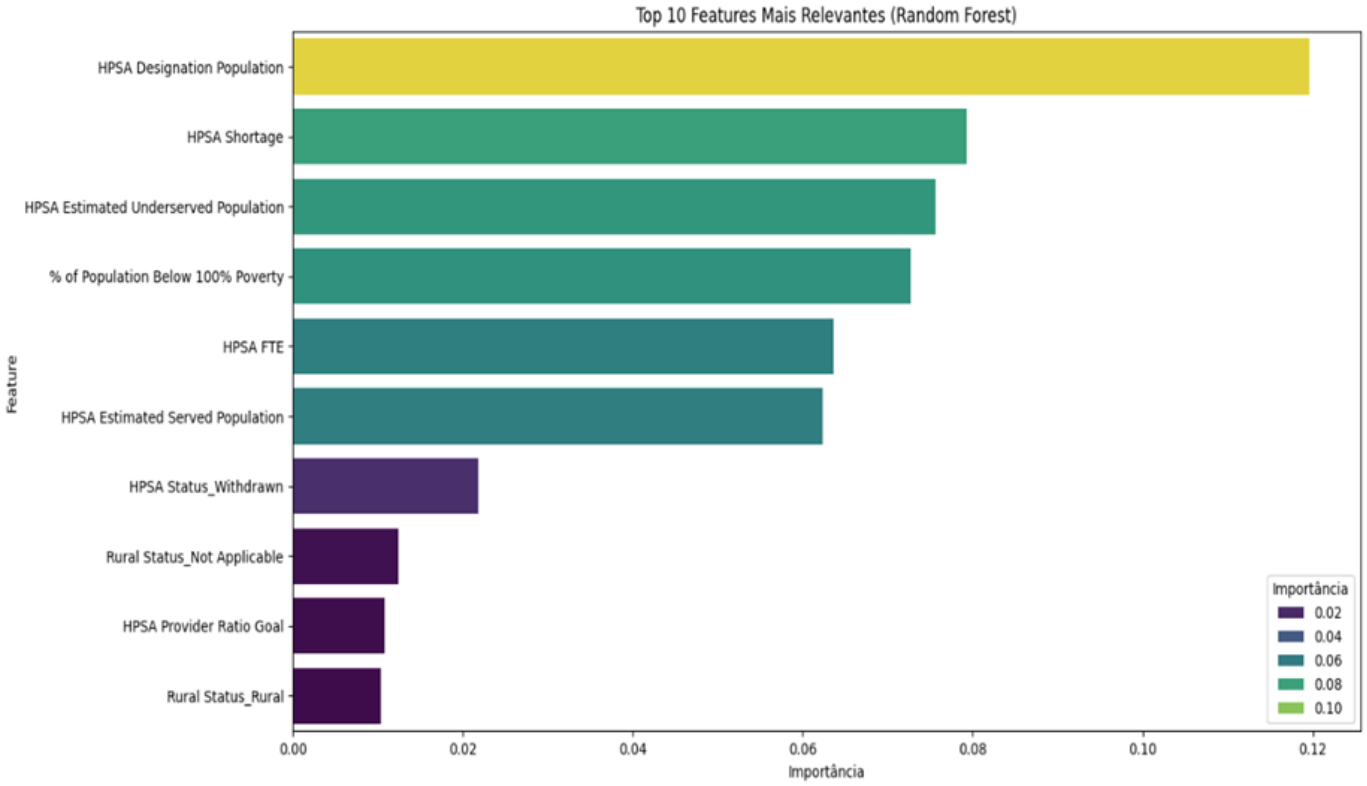

Figura 5. Top 10 Features mais relevantes através do Random Forest.

oferece uma ferramenta potencial para que gestores e formuladores de políticas de saúde direcionem recursos de maneira mais estratégica, priorizando áreas de maior necessidade.

- **Importância das Características:** A análise de importância das características revelou que variáveis como HPSA Designation Population, HPSA Shortage e HPSA Estimated Underserved Population desempenham papéis cruciais na predição da pontuação HPSA. Indicadores associados ao nível de pobreza e localidade rural também aparecem como *features* importantes, indo ao encontro do que foi visto na revisão de literatura. Isso corrobora a ideia de que a escassez de profissionais não é distribuída aleatoriamente pelos EUA, mas decorre de fatores sociais e estruturais rastreáveis. A relevância desses indicadores demográficos e de atendimento confirma a necessidade de considerar múltiplos aspectos ao analisar áreas HPSA.

Em síntese, os resultados obtidos indicam que a abordagem preditiva proposta é promissora para orientar a alocação estratégica de recursos e a formulação de políticas públicas. Para trabalhos futuros, o modelo pode ser empregado para estimar a pontuação HPSA em novas áreas, permitindo que os recursos sejam direcionados de maneira adequada. Em níveis estadual, municipal ou local, as administrações podem desenvolver estratégias para atrair profissionais por meio de incentivos, bem como, um médico que está inaugurando um consultório pode usar a pontuação HPSA de uma região para determinar o impacto de sua atuação.

Referências

[1] Health Resources and Services Administration (HRSA), "HRSA Data Warehouse – Health Professional Shortage Areas," [Online]. Disponível em: https://data.hrsa.gov/data/download, acesso em 17 julho de 2025.

[2] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, New York, NY: Springer, 2001.

[4] Ramezani, N., Hailemariam, M., Breno, A.J. et al. Impact of County-level health infrastructure on participation in a reform effort to reduce the use of jail for individuals with mental health disorders. Health Justice 11, 27, Jul. 04, 2023, https://doi.org/10.1186/s40352-023-00226-9

[5] B. Rochford, S. Pendse, N. Kumar, e M. De Choudhury, "Leveraging Symptom Search Data to Understand Disparities in US Mental Health Care: Demographic Analysis of Search Engine Trace Data," *JMIR Ment. Health*, vol. 10, p. e43253, Jan. 30, 2023, doi:10.2196/43253.

[6] A. Jamal, "Effect of Telemedicine Use on Medical Spending and Health Care Utilization: A Machine Learning Approach," *AJPM Focus*, vol.2, no.3, p.100127, Jun. 15, 2023, doi: 10.1016/j.focus.2023.100127.