**Predição de Status Clínico Pós-Tratamento Oncológico: Uma Abordagem de *Machine Learning* em Dados de Mundo Real**

Aléxia Polo Siqueira¹\*;Pedro Ramon Almeida Oiticica2

1 Hospital de Câncer de Barretos. Analista de Dados, Núcleo de Inteligência e Análise de Dados. Rua Antenor Duarte Viléla, 1331 - Dr. Paulo Prata14784-400. Barretos. São Paulo. Brasil.

2 Laboratório Nacional de Nanotecnologia para o Agronegócio (LNNA) Embrapa Instrumentação; 13560-970 São Carlos. SP. Brasil

\*autor correspondente: alexiapolo@usp.br

**Predição de Status Clínico Pós-Tratamento Oncológico: Uma Abordagem de *Machine Learning em Dados de Mundo Real***

**Resumo**

O câncer é uma das principais causas de mortalidade global, e a melhoria da infraestrutura de saúde e do tratamento oncológico é fundamental para trazer benefícios os desfechos clínicos. Investimentos contínuos em infraestrutura de dados e aprimoramento dos modelos são essenciais para maximizar sua utilidade no planejamento de políticas de saúde e no tratamento oncológico. Os algoritmos *Random Forest* e *XGBoost* foram aplicados para prever o status clínico de pacientes oncológicos com base em dados dos Registros Hospitalares de Câncer (RHC). O desempenho dos modelos foi avaliado por métricas como acurácia, precisão, revocação e F1-score. Foi aplicada a técnica de Eliminação Recursiva de Atributos para otimizar a seleção de variáveis. Valores de *SHAP* foram empregados para identificar a influência de cada variável no resultado da predição. O modelo *Random Forest* apresentou uma acurácia de 76,2%, e se mostrou mais eficiente na classificação da classe 'Vivo'. O *XGBoost*, por sua vez, demonstrou uma acurácia de 71,4%, e se destacou na detecção de óbitos apresentando uma revocação superior (79,4%). Esses resultados indicam que ambos os modelos podem ser utilizados de maneira complementar para predizer desfechos clínicos em pacientes oncológicos, especialmente em contextos de saúde pública.

**Palavras-chave:** registros hospitalares de câncer; aprendizado de máquina; dados de mundo real

**Prediction of Clinical Status Post-Oncological Treatment: A Machine Learning Approach Using Real-World Data**

**Abstract**

Cancer is one of the leading causes of global mortality, and improving healthcare infrastructure and cancer treatment is essential to achieve better clinical outcomes. The use of real-world data in machine learning models presents a promising approach to predict post-treatment clinical status and identify the characteristics influencing it. The Random Forest and XGBoost algorithms were applied to predict the clinical status of cancer patients based on data from Cancer Hospital Registries (CHR). The models' performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. Recursive Feature Elimination was employed to optimize variable selection, and SHAP values were used to identify the contribution of each variable to the prediction outcome. The Random Forest model showed an accuracy of 76.2% and proved to be more efficient in classifying the 'Alive' class. *XGBoost*, on the other hand, showed an accuracy of 71.4%, and stood out in the detection of deaths with a higher revocation (79.4%). These results indicate that both models can be used in a complementary way to predict clinical outcomes in cancer patients, especially in public health contexts.

**Keywords:** data quality, hospital registries, machine learning, real-world data

**Introdução**

O câncer é uma doença multifatorial e uma das principais questões de saúde pública global, apresentando um aumento significativo nas taxas de incidência e mortalidade nas últimas décadas(INCA - Instituto Nacional de Câncer, [s.d.]; Sung et al., 2021). Esse cenário impõe desafios aos sistemas de saúde e políticas públicas, especialmente em países com disparidades socioeconômicas e de infraestrutura como o Brasil. A detecção e tratamento precoce são essenciais para melhorar o desfecho e qualidade de vida do paciente(Raoof et al., 2022). No entanto, o acesso desigual aos serviços de saúde especializados no tratamento oncológico agrava ainda mais o impacto da doença em populações heterogêneas(Cardone et al., 2023).

Os Registros Hospitalares de Câncer (RHC) desempenham um papel crucial no planejamento e formulação de políticas públicas de saúde, fornecendo dados de mundo real que documentam a incidência, os padrões de tratamento e a eficácia terapêutica(Jacob Kligerman, 2001). Esses dados, são essenciais para o desenvolvimento de novos indicadores de saúde e para a otimização dos tratamentos médicos (Rudrapatna et al., 2020). Provenientes de várias fontes, como prontuários eletrônicos, dispositivos de monitoramento e estudos retrospectivos, esses dados oferecem uma visão abrangente e detalhada da prática clínica. Essa perspectiva é especialmente relevante na oncologia, onde variáveis clínicas, sociodemográficas e geográficas podem influenciar significativamente os desfechos terapêuticos(Sherman et al., 2016; Saesen et al., 2023).

A concentração de centros especializados, como os CACONs (Centros de Alta Complexidade em Oncologia) e os UNACONs (Unidades de Assistência de Alta Complexidade em Oncologia), em grandes centros urbanos é um reflexo das disparidades regionais e socioeconômicas do país (Portaria nº 140 de 27 de fevereiro de 2014). Em regiões remotas, a infraestrutura insuficiente e a escassez de recursos médicos resultam em atrasos no diagnóstico e no início do tratamento, prejudicando diretamente o prognóstico dos pacientes. Estudos mostram que o tempo entre o diagnóstico e o início do tratamento é um fator determinante para a sobrevida e a eficácia dos tratamentos oncológicos(Jomar et al., 2023; Tope et al., 2023). Neste contexto, a Lei dos 60 Dias (Lei nº 12.732/2012) foi implementada no Brasil em 2012, exigindo que os pacientes com câncer tratados no Sistema Único de Saúde (SUS) comecem o tratamento em até 60 dias após o diagnóstico confirmado. Embora represente um avanço importante, sua implementação enfrenta obstáculos, particularmente em áreas com menor desenvolvimento de infraestrutura de saúde. Diante desse cenário, o uso de ferramentas de análise de dados e aprendizado de máquina podem ajudar na busca por novas soluções de otimização para o tratamento oncológico.

Neste trabalho foram empregando dados de mundo real de RHCs para treinar algoritmos de aprendizado de máquina para previsão do status clínico de pacientes oncológicos, após o término do tratamento. Investigou-se influência dos atributos clínicos, sociodemográficos, e o cumprimento da Lei dos 60 Dias na resposta ao tratamento.

Os resultados obtidos demonstram que a aplicação de algoritmos de aprendizado de máquina a dados de mundo real pode proporcionar insights relevantes para otimizar o tratamento oncológico no SUS, sobretudo acerca do planejamento de políticas públicas no que tange ao rastreio de indivíduos em risco e diagnóstico precoce.

**Material e Métodos**

**Fonte de Dados**

Este estudo utilizou o conjunto de dados proveniente dos Registros Hospitalares de Câncer (RHC), acessado através do tabulador de dados IntegradorRHC (https://irhc.inca.gov.br/RHCNet/, acessado em 07 de julho de 2024). Os dados foram coletados em formato estruturado (.dbf), compreendendo 313.497 observações com 49 atributos (Anexo I – Dicionário de Banco de Dados). Todas as etapas de pré-processamento, análise exploratória, modelagem e avaliação de performance foram realizadas com programas em linguagem R.

*Pré-processamento de dados*

O banco de dados, foi importado usando o pacote *foreign (v0.8.87)*. A limpeza e tratamento dos dados foram executados com o pacote *tidyverse (v2.0.0) (Wickham et al., 2019).*

Inconsistências como idades negativas, caracteres especiais, datas e intervalos de datas contendo erros foram tratados como dados faltantes. Foram considerados apenas as instâncias com o atributo idade superior a 30 anos, uma vez que os padrões biológicos, de tratamento e sobrevida dos tumores em adultos diferem do observados em crianças e jovens, de modo a evitar a introdução de uma variabilidade adicional no modelo. As idades categorizas em faixas etárias com intervalos de 10 anos.

A topografia tumoral, um atributo de extrema importância para a análise, foi codificado de acordo com a Classificação Internacional de Doenças para Oncologia (CID-O) (Organização Mundial da Saúde, [s.d.]). Os tumores da topografia pele com subtipo não-melanoma são os mais incidentes dentre os tipos tumorais, e se diferenciam por terem menor agressividade, e menores taxas de mortalidade, comparado aos tumores do tipo melanoma. Por este motivo este subtipo é geralmente tratado separadamente a outras topografias. Para o presente trabalho foram filtrados apenas os subtipos de tumores de pele melanoma junto a outras topografias. As topografias foram agrupadas em categorias relativas à localização, como "Cabeça e Pescoço", "Tórax", "Digestivo",” Mama”,” Urologia” dentre outras.

Foi criada a variável referente ao cumprimento da Lei dos 60 Dias (lei\_60\_respeitou) com base na diferença entre a data do diagnóstico e a data de início do tratamento. Diferenças de datas negativas ou inconsistentes foram tratadas como nulas.

Em seguida, foi realizada a seleção de atributos, com agrupamento e balanceamento de categorias para as análises subsequentes. A lista completa de agrupamentos está descrita no código (Anexo II).

Os atributos selecionados para análises subsequentes incluem: faixa etária, sexo, tipo de caso (se é virgem de tratamento ou não), unidade federativa de residência, tratamento oncológico inicial, respeito à Lei dos 60 Dias, se o tratamento foi realizado em unidades ou centros de alta complexidade em oncologia, status da doença ao final do tratamento, estadiamento clínico ao diagnóstico, região de residência, e localização do tumor (N = 247.740 observações). As instâncias contendo dados faltantes foram removidas da base de análise, resultando em um banco de dados de 51.487 observações.

A relação de variáveis categóricas e suas categorias é apresentada na Tabela 1.

**Tabela 1.** Relação de variáveis categóricas e suas respectivas categorias.

|  |  |  |
| --- | --- | --- |
| Variável | Tipo | Níveis |
| faixa\_etaria | Fator | "30-40 anos", "41-50 anos", "51-60 anos", "61-70 anos", "Mais de 70 anos" |
| sexo | Fator | "Masc", "Fem" |
| tipo\_caso | Fator | "Analítico", "Não analítico" |
| estado\_residencia | Fator | "AC", "AL", "AM", "AP", "BA", "CE", "DF", "ES", "GO", "MA", "MG", "MS", "MT", "PA", "PB", "PE", "PI", "PR", "RJ", "RN", "RO", "RR", "RS", "SC", "SE", "SP", "TO" |
| lei60\_respeitou | Fator | "Não", "Sim", |
| tratamento\_inicial | Fator | "Nenhum", "Cirurgia", "Radioterapia", "Quimioterapia", "Outro" |
| cacon\_unacon | Fator | "CACON", "Outro", "UNACON", |
| status\_doenca | Fator | "Vivo", "Óbito" |
| estadiamento | Fator | "0 e I", "II", "III", "IV" |
| regiao | Fator | "Centro-Oeste", "Nordeste", "Norte", "Sudeste", "Sul" |
| tumores\_agrupado | Fator | "Cabeça e pescoço", "Digestivo", "Gineco", "Mama", "Outros", "Sarcomas", "Sistema endócrino e glândulas", "Sistema hematopoético", "Sistema nervoso", "Tórax", "Urologia" |

**Treinamento do modelo e validação cruzada**

Para avaliação da a capacidade preditiva dos modelos e garantir robustez, a modelagem consistiu em 3 etapas: treinamento, validação cruzada e teste.

O particionamento do banco de dados foi realizado na proporção de 80% (n = 41.486) para o conjunto de treinamento e 20% (n = 10.371) para o conjunto de teste (Figura 1). Esta etapa foi feita de maneira estratificada com relação a variável resposta (status\_doença), de maneira a minimizar os efeitos do desbalanceamento das categorias no modelo.

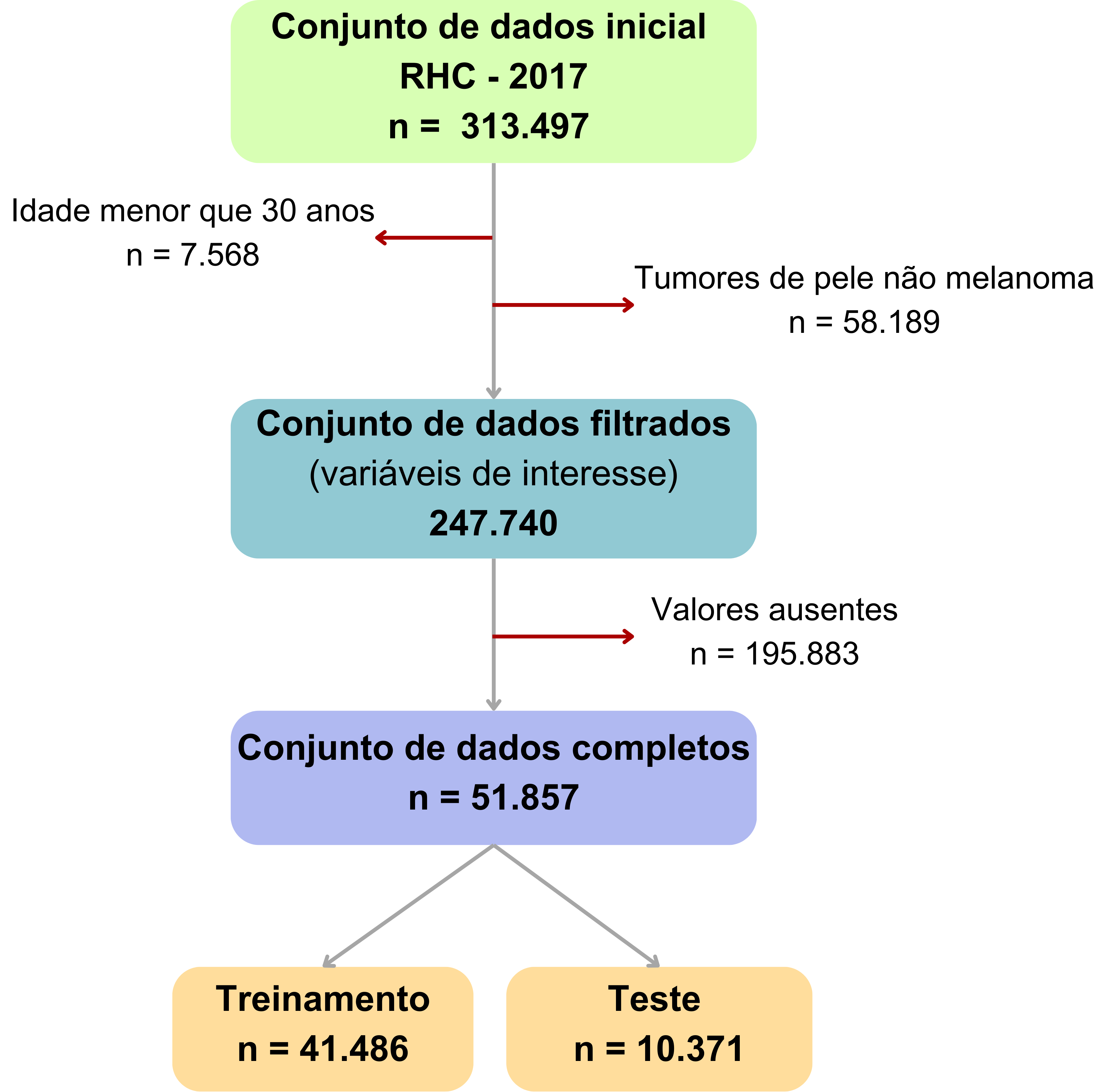


Figura 1 - Fluxograma do processo de filtragem e particionamento dos dados.

O conjunto de treinamento foi utilizado para ajustar os modelos iniciais, enquanto o conjunto de validação serviu para otimizar os hiperparâmetros dos algoritmos. O desempenho final dos modelos foi então avaliado no conjunto de teste.

Os modelos preditivos foram treinados utilizando as técnicas de aprendizado supervisionado *Random Forest* e *XGBoost*, implementadas em R utilizando o pacote *caret* (v6.0.94). Para a escolha dos melhores hiperparâmetros, foi implementada a técnica de busca aleatória *Random Search,* visando encontrar a melhor configuração para a acurácia do modelo.

Os hiperparâmetros testados para *Random Forest* estão discriminados na Tabela 2, e para *XGBoost*  na Tabela 3.

**Tabela 2 -** Hiperparâmetros utilizados no ajuste do modelo *Random Forest*.

|  |  |  |
| --- | --- | --- |
| **Hiperparâmetro** | **Descrição** | **Valores Testados** |
| *mtry* | Número de variáveis consideradas para divisão em cada nó da árvore | Aleatórias de 2 a 10 |
| *splitrule* | Critério utilizado para divisão dos nós | Gini e extratrees |
| *min.node.size* | Número mínimo de observações que cada nó deve conter | Aleatórias de 1 a 10 |

**Tabela 3 -** Hiperparâmetros utilizados no ajuste do modelo *XGBoost* .

|  |  |  |
| --- | --- | --- |
| **Hiperparâmetro** | **Descrição** | **Valores Testados1** |
| *nrounds* | Número de rodadas de *boosting* | 50 e 100 |
| *max\_depth* | Profundidade máxima das árvores | 3 e 4 |
| *eta* | Taxa de aprendizado | Aleatórias entre 0.05 e 0.1 |
| *gamma* | Redução mínima de perda para realizar divisão dos nós | Amostras aleatórias entre 0 e 2 |
| *colsample\_bytree* | Proporção de colunas amostradas em cada árvore | Aleatórias entre 0.6 e 0.8 |
| *min\_child\_weight* | Peso mínimo de instâncias de cada nó | Aleatórias entre 1 e 4 |
| *subsample* | Proporção de amostras utilizada para construir cada árvore | Aleatórias entre 0.5 e 0.8 |

1- A quantidade e variedade de hiperparâmetros testados para o modelo XGBoost foram limitados pelo poder computacional das ferramentas disponíveis para avaliação.

A técnica de validação cruzada foi otimizada para 5 *folds* no modelo *Random Forest* e 3 *folds* para o modelo *XGBoost*. Após a execução da validação cruzada, o melhor conjunto de hiperparâmetros foi selecionado para cada modelo com base na acurácia média obtida nos diferentes *folds.* Em seguida, os modelos foram aplicados ao conjunto de teste independente, para avaliação de performance e importância das variáveis.

*Avaliação dos modelos*

Para avaliação de performance, foram verificadas as métricas:

1. Acurácia: avalia a proporção de previsões corretas com relação ao total de previsões feitas pelo modelo, representada pela fórmula:

*A = VP + VN / (VP+VN + FP + FN),*

na qual VP corresponde aos verdadeiros positivos, VN corresponde aos verdadeiros negativos, FP corresponde aos falsos positivos e FN aos falsos negativos;

1. Precisão: avalia a proporção de verdadeiros positivos nas previsões feitas pelo modelo, representada pela fórmula:

*P = VP / (VP+VN),*

na qual VP corresponde aos verdadeiros positivos, FP corresponde aos falsos positivos;

1. Especificidade: avalia a proporção de verdadeiros negativos em meio a todos que deveriam ter sido classificados como positivos, representada pela fórmula:

*E = VN / (VN + FP)​,*

na qual VN corresponde aos verdadeiros negativos, e FP corresponde aos falsos positivos;

1. Revocação (sensibilidade): avalia a proporção de verdadeiros positivos em meio a todos que deveriam ter sido classificados como positivos, representada pela fórmula:

*R = VN / (VP + FN),*

na qual VP corresponde aos verdadeiros positivos, VP aos verdadeiros positivos e FN corresponde aos falsos negativos;

1. F1-score: compreende a média harmônica entre precisão e revocação, considerando a performance do modelo entre falsos positivos e falsos negativos, representada pela fórmula:

*F1 = 2 x (P x R / P+R)*

Nesta avaliação, para escolha do modelo de melhor performance priorizou-se a métrica de acurácia, seguida de F1-score, e especificidade.

Em seguida, a avaliação da importância das variáveis foi realizada utilizando o método de importância por impureza (*impurity importance,* índice de Gini) em ambos os modelos ajustados de *Random Forest* e *XGBoost*.

*Otimização do modelo de melhor performance*

A técnica de Eliminação Recursiva de Atributos (do ingês *Recursive Feature Elimination*, RFE) foi aplicada no modelo ajustado de melhor performance, de modo a selecionar as variáveis mais relevantes na modelagem e como uma estratégia de otimização e aprimoramento do mesmo. O RFE é um método iterativo que, a cada iteração, treina o modelo com diferentes subconjuntos de variáveis. A importância de cada uma é avaliada, além de considerado o índice de Gini, por meio da remoção iterativa das variáveis de menor importância e comparação da avaliação de desempenho do modelo nestes contextos. O modelo é re-treinado até que um número desejado de características seja atingido ou a performance do modelo se estabilize(James et al., 2021; Awad et al., 2023). Essa abordagem ajuda a reduzir a dimensionalidade dos dados, mitigando o risco de sobreajuste e melhorando a interpretabilidade e eficiência do modelo

Neste estudo foram testados conjuntos de 3, 5, 7 e 10 variáveis, e a avaliação do modelo foi conduzida com validação cruzada repetida (5 repetições de 5 *folds*). As variáveis selecionadas no processo foram utilizadas para construir um modelo Random Forest otimizado, tendo como base o modelo de melhor performance da etapa anterior.

Após esta etapa, o modelo de melhor desempenho dentre o que incluiu todas as variáveis do banco de dados para a modelagem e pelo método de seleção de atributos foi escolhido como modelo final.

*Investigação da contribuição individual de cada variável no modelo final*

Após a definição do modelo final, foi realizada uma análise detalhada para entender a contribuição individual de cada variável no resultado predito, utilizando o método SHAP (do inglês “SHapley Additive exPlanations”). Esse método permite quantificar de forma clara e interpretável a influência de cada variável nas predições, fornecendo uma explicação robusta e equilibrada sobre o impacto de cada atributo na performance do modelo (Lundberg & Lee, 2017). A análise foi realizada utilizando o pacote DALEX em R (v2.4.3).

**Resultados e Discussão**

*Análise exploratória e descritiva*

A análise descritiva dos dados do RHC do ano de 2017 acerca dos atributos selecionados se encontra apresentada na Tabela 4.

Neste registro de 51.857 pacientes, foi observado que a maioria dos casos de câncer ocorrem nas faixas etárias a partir de 60 anos (53%), com predominância do sexo feminino (59%) e casos virgens de tratamento oncológico (88%). Cerca de 60% dos pacientes são de origem da região sudeste e sul do país. Foram mais prevalentes na casuística casos de tumores de mama e ginecologicos (42%) e do trato digestivo (22%), diagnosticados com doença locorregional ou avançada (52%), em estadiamentos clínicos III ou IV. Cerca de 40% receberam abordagem de tratamento inicial cirúrgica, e 88% se apresentam com status vivo após o tratamento.

Tabela 4 - Tabela de frequências dos casos do ano de 2017.

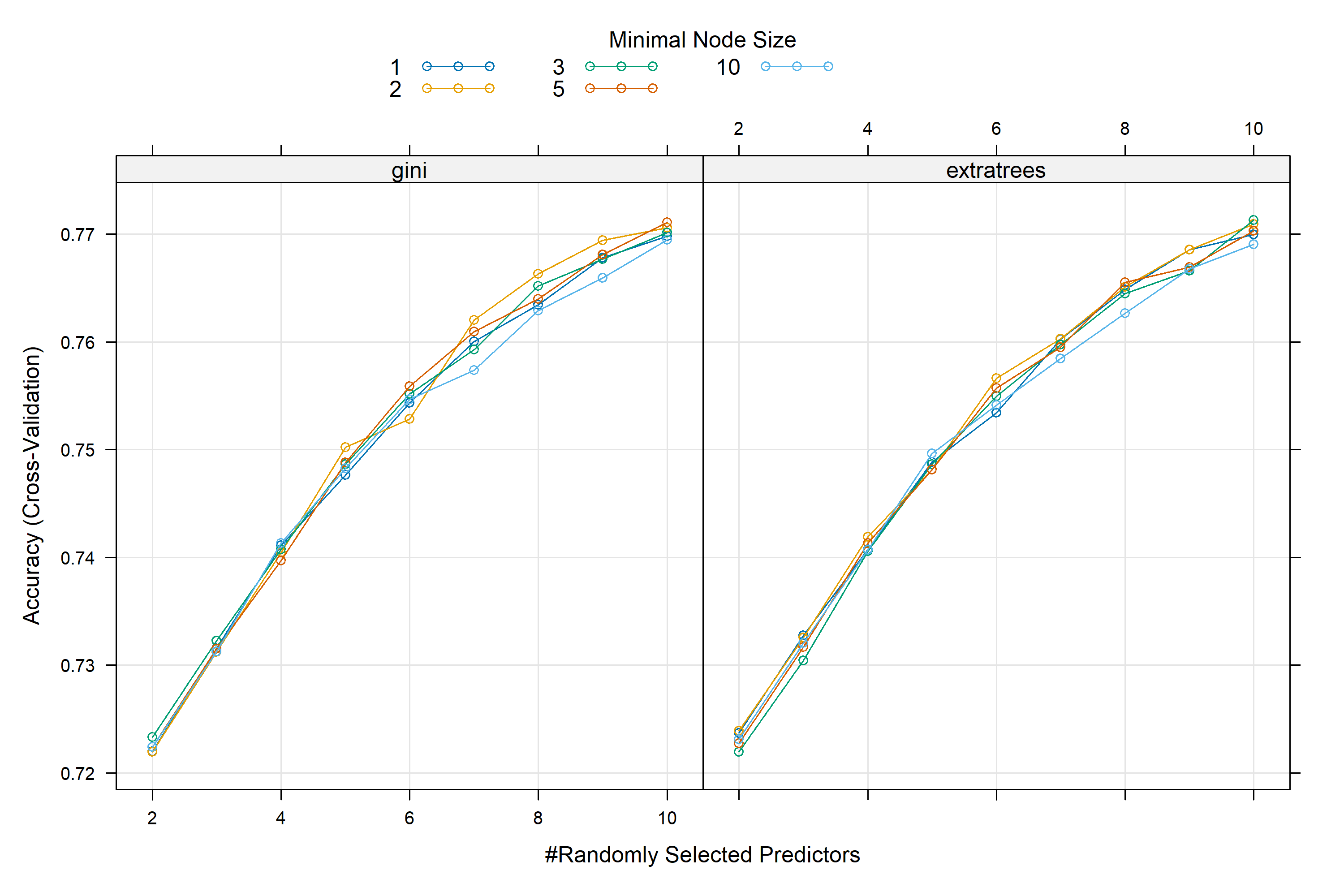
|  |  |  |
| --- | --- | --- |
| **Variável** | **Categoria** | **N (%)** |
| Faixa etária | 31-40 anos | 4.422 (8,5%) |
| 41-50 anos | 7.896 (15%) |
| 51-60 anos | 12.296 (24%) |
| 61-70 anos | 14.410 (28%) |
| Mais de 70 anos | 12.833 (25%) |
| Sexo | Masculino | 21.281 (41%) |
| Feminino | 30.576 (59%) |
| Tipo de caso | Analítico | 45.434 (88%) |
| Não analítico | 6.423 (12%) |
| Estado de residência | AC | 95 (0,2%) |
| AL | 718 (1,4%) |
| AM | 681 (1,3%) |
| AP | 114 (0,2%) |
| BA | 4.151 (8,0%) |
| CE | 1.610 (3,1%) |
| DF | 579 (1,1%) |
| ES | 1.943 (3,7%) |
| GO | 415 (0,8%) |
| MA | 1.809 (3,5%) |
| MG | 10.199 (20%) |
| MS | 616 (1,2%) |
| MT | 586 (1,1%) |
| PA | 1.329 (2,6%) |
| PB | 219 (0,4%) |
| PE | 3.606 (7,0%) |
| PI | 382 (0,7%) |
| PR | 7.516 (14%) |
| RJ | 2.986 (5,8%) |
| RN | 2.700 (5,2%) |
| RO | 611 (1,2%) |
| RR | 97 (0,2%) |
| RS | 5.943 (11%) |
| SC | 2.014 (309%) |
| SE | 166 (0,3%) |
| SP | 31 (<0,1%) |
| TO | 741 (1,4%) |
| Respeito a Lei dos 60 dias | Não | 25.626 (49%) |
| Sim | 26.231 (51%) |
| Tratamento inicial | Cirurgia | 20.710 (40%) |
| Outro | 4.582 (8.8%) |
| Quimioterapia | 17.201 (33%) |
| Radioterapia | 9.364 (18%) |
| Unidade de tratamento | CACON | 19.830 (38%) |
| Outra | 99 (0,2%) |
| UNACON | 31.928 (62%) |
| Status após tratamento | Óbito | 6.342 (12%) |
| Vivo | 45.515 (88%) |
| Estadiamento clínico ao diagnóstico | 0 | 2.267 (4,4%) |
| I | 8.510 (16%) |
| II | 13.996 (27%) |
| III | 13.670 (26%) |
| IV | 13.414 (26%) |
| Região | Centro-Oeste | 2.196 (4,2%) |
| Nordeste | 15.361 (30%) |
| Norte | 3.668 (7,1%) |
| Sudeste | 15.159 (29%) |
| Sul | 15.473 (30%) |
| Localização do tumor | Mama e Ginecológicos | 21.590 (42%) |
| Melanomas, Sarcomas e SNC | 767 (1,5%) |
| Sistema Digestivo | 11.215 (22%) |
| Sistema hematopoético | 869 (1,7%) |
| Sistema Respiratório | 6.933 (13%) |
| Sistema Urológico | 10.483 (20%) |

N - número de casos.

*Modelagem Random Forest*

O modelo Random Forest foi ajustado com 41.486 amostras e 10 variáveis preditoras. O processo de *resampling* foi realizado por meio de validação cruzada com 5 *folds* e reamostragem utilizando a técnica de *up-sampling*, para balanceamento das classes. A amostragem entre os *folds* foi feita com reposição.

O modelo de melhor acurácia média entre as predições foi treinado em 500 árvores, utilizando todas as 10 variáveis na divisão dos nós das árvores, por meio de divisão aleatória (hiperparâmetro *extratrees, do inglês “Extremely Randomized Trees”),* com 1 divisão aleatória testada em cada nó, e com no mínimo 3 observações necessárias para cada nó (Figura 2).

**Figura 2 -** Comparação do desempenho do modelo *Random Forest* ajustado com diferentes combinações de hiperparâmetros.

O tipo de árvore empregado foi para estimativa de probabilidade, na qual ao invés de apenas classificar as observações, o modelo fornece probabilidades para cada classe, de maneira a se possível se verificar o grau de certeza das predições

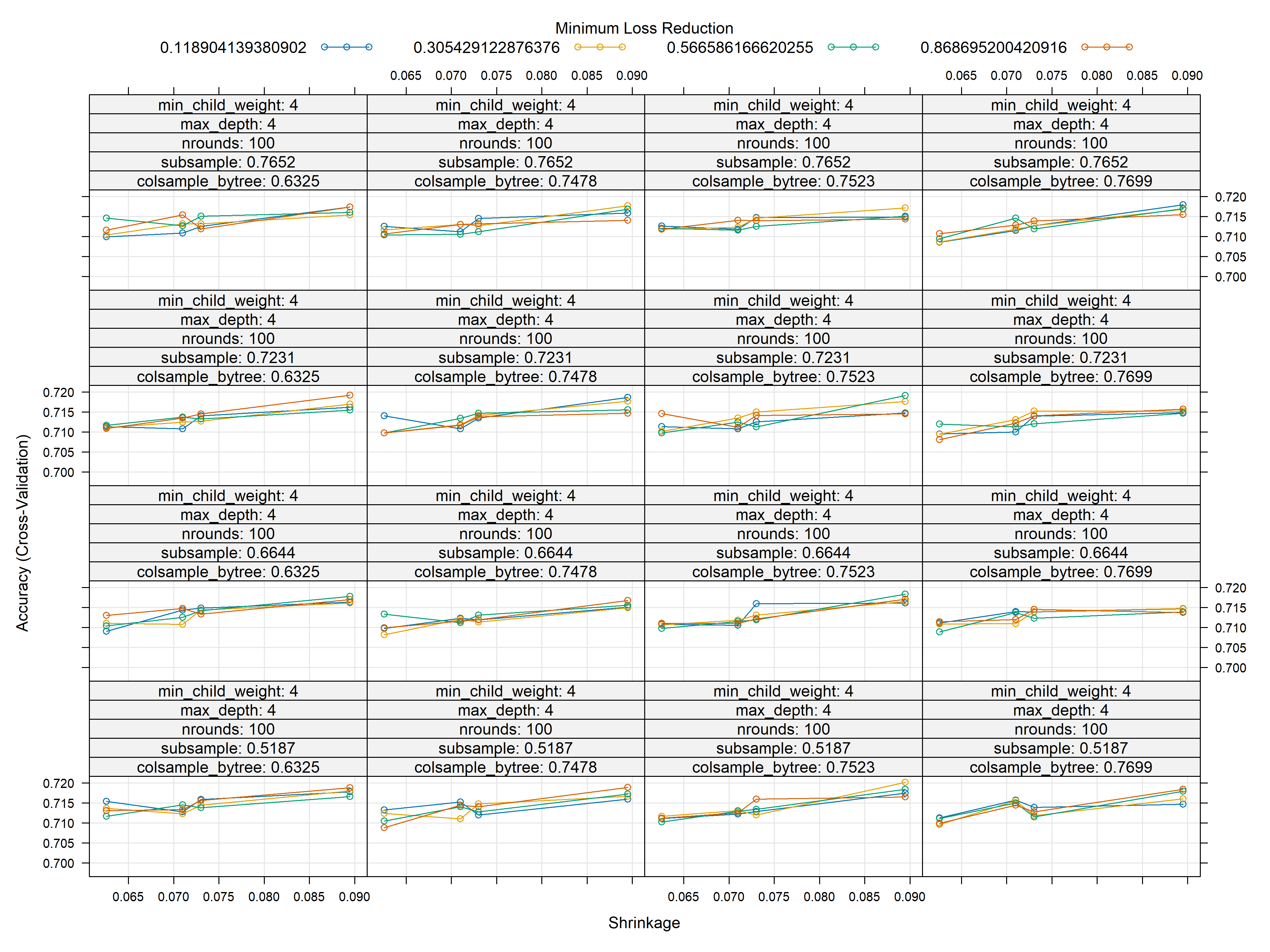
Este modelo apresentou um a taxa de 11,78%, indicando uma boa performance geral.

*Modelagem XGBoost*

O modelo *XGBoost* foi ajustado com as mesmas 41.486 amostras e 10 variáveis preditoras. Assim como o *Random Forest*, o processo de *resampling* foi realizado por meio de validação cruzada com 3 *folds* e reamostragem utilizando a técnica de *up-sampling*, para balanceamento das classes. A amostragem entre os *folds* também foi feita com reposição.

O modelo com melhor desempenho foi ajustado utilizando 100 iterações de *boosting* cada uma construindo uma nova árvore de decisão com o objetivo de corrigir os erros cometidos pelas árvores anteriores. O valor de taxa de aprendizado foi ajustado para 0,089, controlando o impacto de cada árvore adicionada no modelo final, enquanto o parâmetro gamma foi configurado para 0,305, restringindo a divisão dos nós a casos em que há um ganho significativo. Ainda, considerou 4 como um peso mínimo de instâncias, o que ajuda a evitar divisões com pouca representatividade (Figura 3).

A profundidade máxima das árvores foi 4, garantindo que o modelo pudesse capturar interações mais complexas entre as variáveis, e foi ajustado para selecionar aleatoriamente 75,2% das variáveis em cada árvore e utilizar 51,8% das amostras disponíveis para cada iteração, o que contribuiu para a robustez e estabilidade.

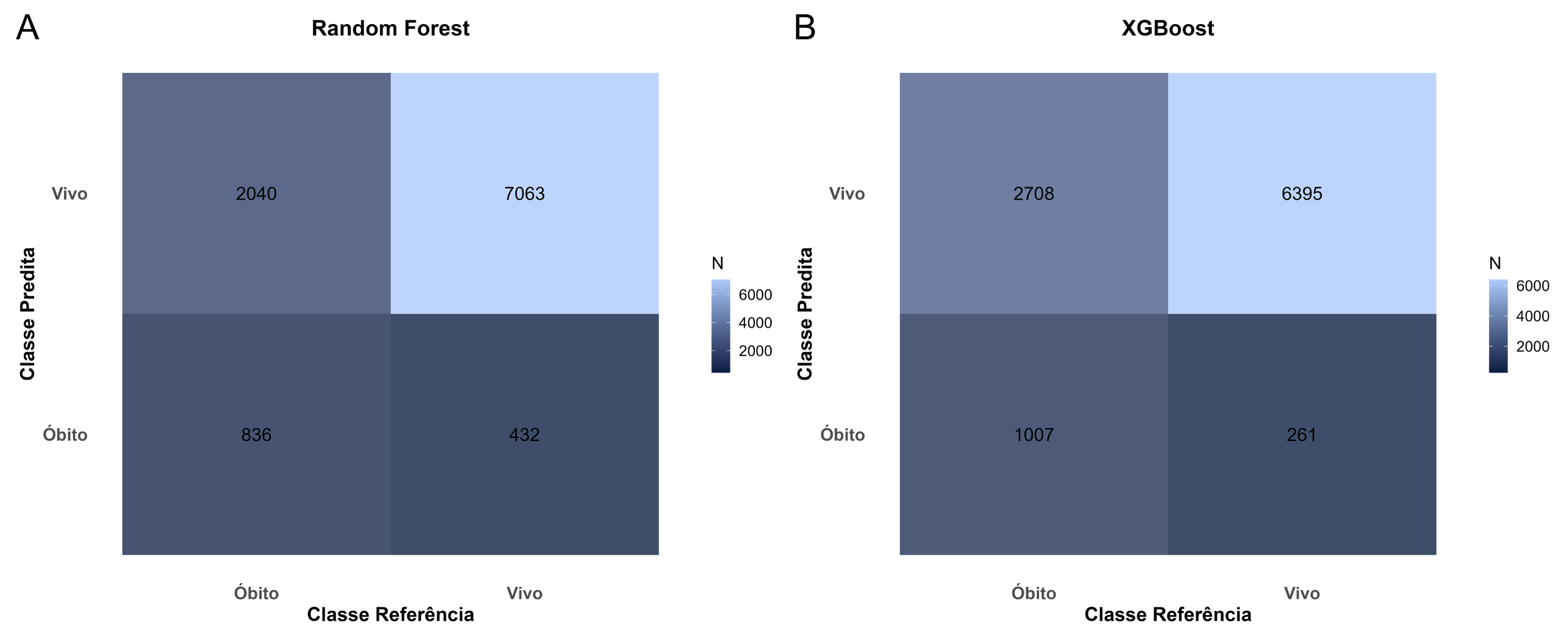
**Figura 3 -** Comparação do desempenho do modelo *XGBoost* ajustado com diferentes combinações de hiperparâmetros

Este modelo apresentou uma taxa de erro de 10,47%, indicando um desempenho levemente superior ao *Random Forest*.

*Teste e avaliação de performance*

Em seguida, ambos os modelos foram aplicados ao conjunto de teste independente para avaliação da performance final.

O modelo Random Forest (Figura 4A) obteve uma melhor classificação entre as classes "Vivo" e "Óbito" quando comparado ao *XGBoost* (Figura 4B). No entanto, o *XGBoost* apresentou maior sensibilidade para a classe "Óbito", o que sugere uma melhor capacidade de identificar corretamente os casos verdadeiros de óbito.

**Figura 4 -** Matrizes de confusão da predição de desfechos clínicos no conjunto de teste para os modelos de melhor peformance *Random Fores*t (A) e *XGBoost* (B). As cores refletem o número de observações classificadas em cada classe, com o gradiente de cores.

O *Random Forest* apresentou uma acurácia de 76,2%, com uma precisão de 29,1% e F1-score de 40,3%. Em termos de especificidade, atingiu 77,6%, enquanto a revocação foi de 65,9%. Por outro lado, o *XGBoost* apresentou uma acurácia de 71,4%, uma precisão de 27,1% e F1-score de 40,4%. A especificidade foi um pouco inferior, com 70,3%, mas a revocação foi superior à do *Random* Forest, com 79,4%. Isso indica que o modelo *XGBoost* teve um desempenho melhor na identificação de verdadeiros positivos, enquanto o *Random Forest* apresentou um melhor balanço entre precisão e especificidade (Tabela 5) .

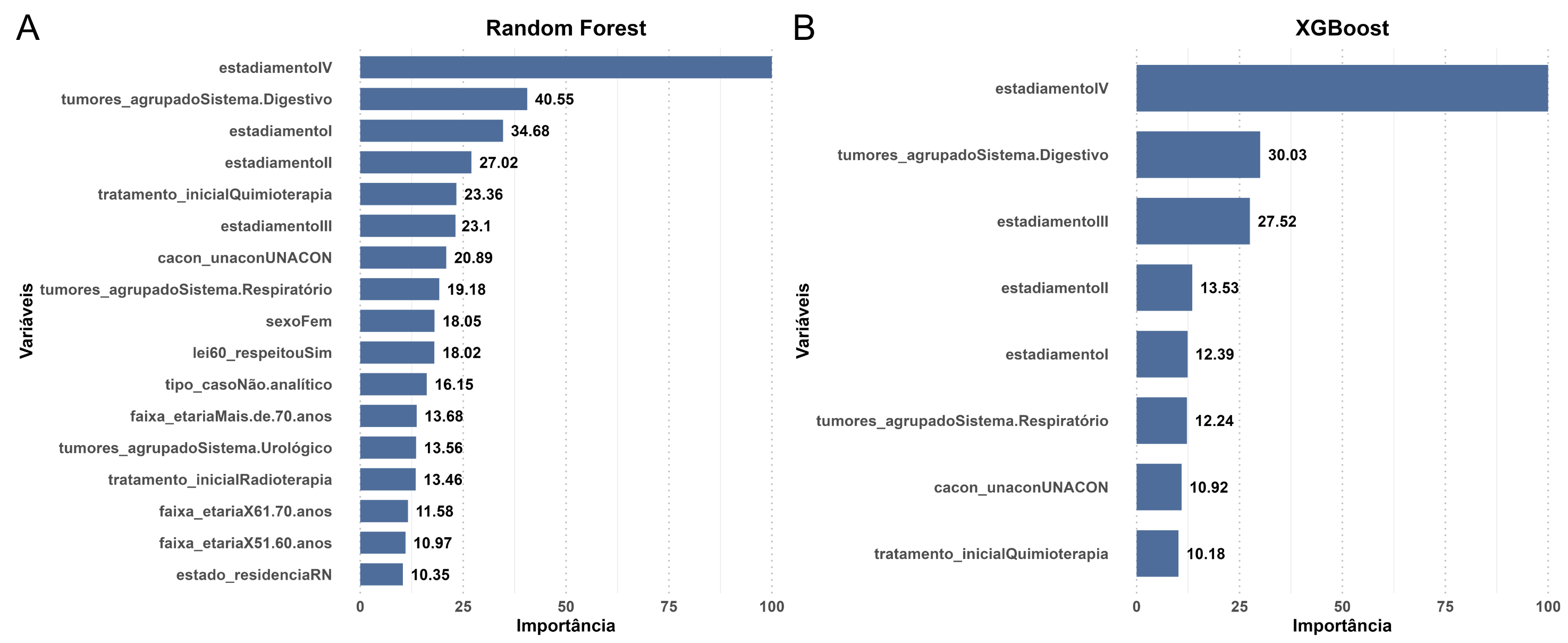
**Tabela 5 -** Métricas de avaliação de desempenho para os modelos *Random Forest* e *XGBoost*.

|  |  |  |
| --- | --- | --- |
| Métrica | *Random Forest* | *XGBoost* |
| Acurácia | 0,762 | 0,714 |
| Precisão | 0,291 | 0,271 |
| Especificidade | 0,776 | 0,703 |
| Revocação | 0,659 | 0,794 |
| F1-Score | 0,403 | 0,404 |

*Avaliação da importância das variáveis*

A análise de importância das variáveis para ambos os modelos foi realizada utilizando o método de importância por impureza, conhecido como índice de Gini.

A variável estadiamento clínico (categoria IV) foi a mais relevante em ambos os modelos, com importância máxima (100%), indicando sua influência decisiva nas predições dos desfechos clínicos dos pacientes. Além de outras categorias da variáveis estadiamento se mostrarem decisivas quanto a desfecho em ambos os modelos, pode-se destacar também a localização de tumores no sistema digestivo, tratamento inicial com quimioterapia, dentre outros. A Figura 5 ilustra as variáveis que apresentaram uma importância relativa maior que 10 em pelo menos um dos modelos.

**Figura 5** - Gráfico de importância das variáveis para os modelos *Random Forest* (A) e *XGBoost* (B) com valores relativos.

*Otimização do modelo de melhor performance*

A aplicação da técnica de RFE permitiu a seleção dos atributos de estadiamento clínico, estado de residência e tipo de caso para ajuste do modelo *Random Forest*. A matriz de confusão (Figura 6) demonstra que o modelo apresentou dificuldades em classificar corretamente os casos de "Óbito", com um número significativo de falsos negativos (943) e falsos positivos (6.112).

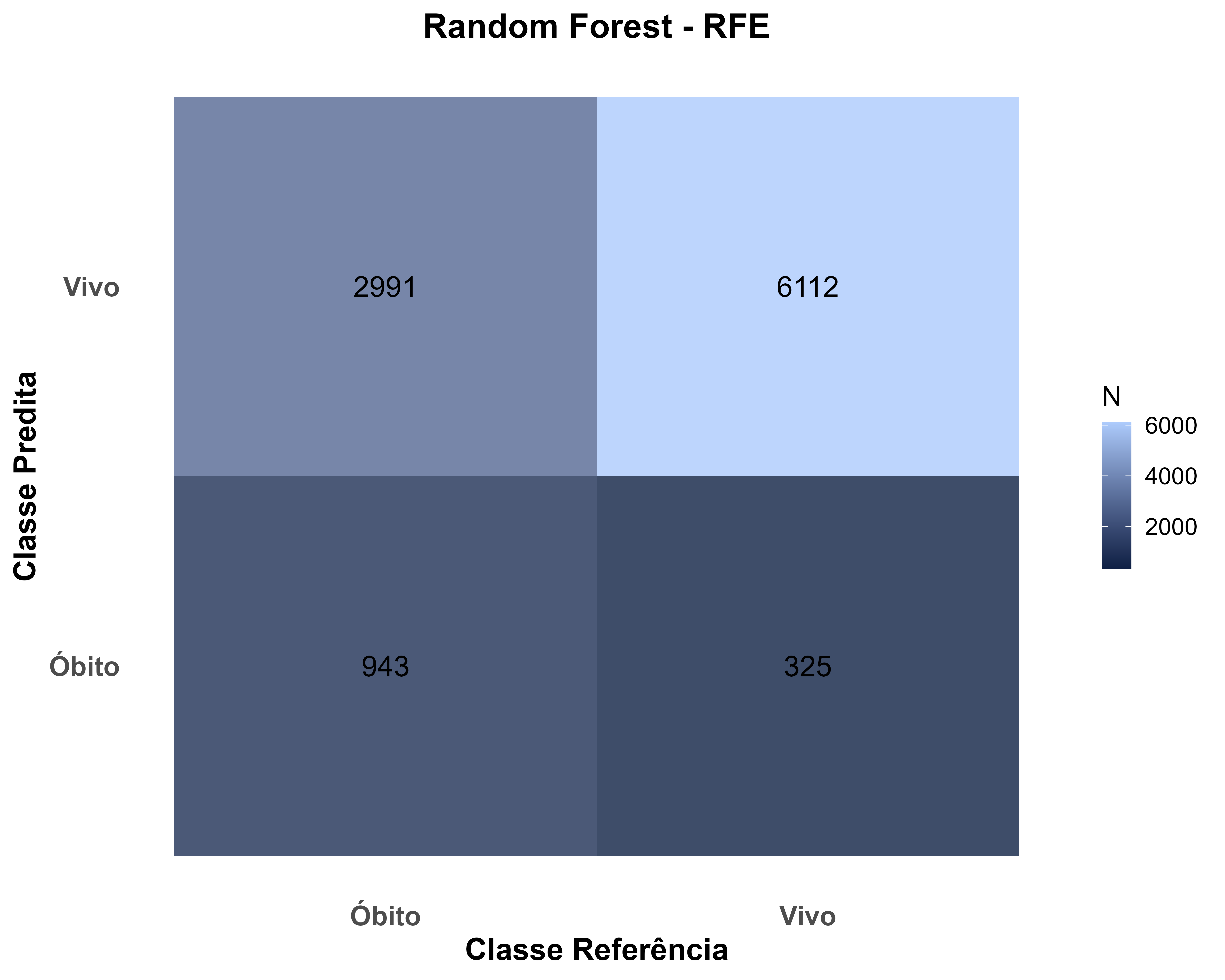


Figura 6 - Matriz de confusão do modelo *Random Forest* otimizado com a técnica de Eliminação Recursiva de Atributos (RFE).

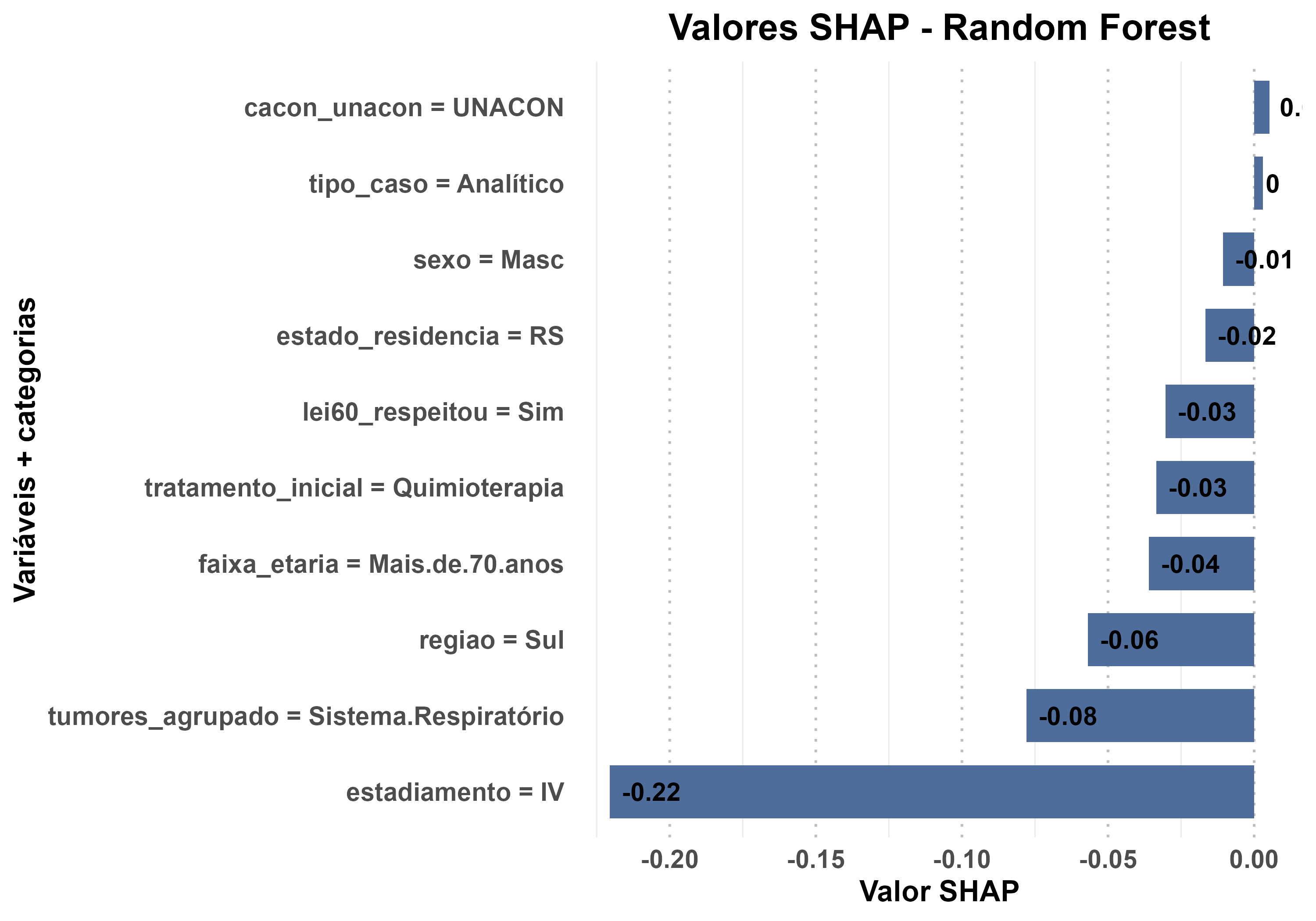
Com relação às métricas, o modelo apresentou um desempenho inferior em comparação ao modelo ajustado sem a seleção de variáveis. Apresentou uma acurácia de 68%, enquanto o modelo completo obteve 76,2%. Além disso, outras métricas como precisão (24,0%), especificidade (67,1%) e F1-score (36,2%) também apresentaram valores mais baixos.

Portanto, neste modelo houve uma maior taxa de erro na predição da classe "Óbito", indicando que a tentativa de otimização com a técnica de RFE não trouxe melhorias significativas. Dessa forma, optou-se por adotar o modelo *Random Forest* sem a aplicação do RFE como o modelo final, devido à sua superioridade em termos de acurácia e desempenho.

*Investigação da contribuição individual de cada variável no modelo final*

Após a escolha do modelo final, de maneira complementar foi realizada uma análise de interpretação das variáveis mais influentes no desempenho do modelo utilizando o método SHAP.

A variável estadiamento clínico se destacou como a mais influente, indicando que diferentes níveis de estadiamento têm um impacto substancial na predição do desfecho clínico. Nest caso, um estadiamento mais avançado (IV) indicou um impacto negativo quando o desfecho de referência é “Vivo”. Este comportamento tambem foi observado, mas com menor magnitude, em tumores do sistema respiraório, região de residência sul e em pacientes com idades maiores de 70 anos. O impacto das variáveis, de maneira mais detalhada, esta ilustrado na Figura 7.



**Figura 7.** Impacto das variáveis explicativas no modelo *Random Forest* utilizando o método SHAP. As barras representam o valor SHAP médio para cada variável, indicando a magnitude e a direção do impacto nas predições.

*Discussão*

Os resultados deste estudo evidenciam o potencial dos algoritmos de *machine learning* na predição de desfechos clínicos em pacientes oncológicos, especialmente no contexto de sistemas de saúde pública, como o SUS. A aplicação de *Random Forest* e *XGBoost* para prever o status clínico após o tratamento proporcionou *insights* valiosos sobre os padrões entre variáveis clínicas e sociodemográficas que influenciam significativamente os resultados do tratamento.

O modelo *Random Forest* se destacou esse modelo se destacou por proporcionar um bom equilíbrio entre precisão e especificidade, sendo mais eficaz na classificação de pacientes como "Vivo", enquanto o modelo *XGBoost*, que apresentou uma acurácia ligeiramente inferior, demonstrou uma capacidade superior de detectar pacientes com maior risco de mortalidade (status óbito). Apesar deste achado, verificou-se que o modelo *Random Forest* apresentou mértricas superiores, com exceção apenas da revocação, e maior sucesso na separação entre as classes "Vivo" e "Óbito", e por este motivo foi escolhido para as análises posteriores.

A análise de importância das variáveis utilizando o método de impureza evidenciou o estadiamento clínico como o principal preditor de desfecho nos dois modelos, evidenciando que pacientes diagnosticados em estágios mais avançados (III e IV) possuem um pior prognóstico em comparação àqueles diagnosticados precocemente. Essa análise sugere que o momento de diagnóstico, ou seja, se o paciente apresenta ou não diagnóstico precoce, é um fator indubitávelmente determinante para a predição dos desfechos oncológicos, e reforça a necessidade e investimento em estratégias de prevenção e monitoramento de indivíduos em risco

No modelo *Random Forest*, a localização do tumor também se mostrou uma variável influente, especialmente para tumores do sistema digestivo e do sistema respiratório, indicando que características biológicas associadas a esses tipos de cânceres impactam diretamente o desfecho clínico. Destaca-se também a variável de tratamento inicial e unidade de tratamento. Estes achados foram confirmados também pela análise de valores SHAP. No modelo *XGBoost*, além do próprio estadiamento clínico, variáveis como unidade de tratamento (se CACON ou UNACON) e tipo de caso (se virgem ou não de tratamento) também mostraram influência considerável, sugerindo que o acesso ao tratamento especializado também é determinante para os resultados observados.

Embora a técnica RFE tenha sido aplicada ao modelo *Random Forest* com o intuito de otimizar o conjunto de variáveis e melhorar a performance, o resultado não foi satisfatório. O modelo ajustado com a RFE apresentou uma acurácia de 68%, inferior ao modelo inicial (76,2%). A redução de desempenho pode ser atribuída à exclusão de variáveis que, embora com importância relativa baixa, podem contribuir para o equilíbrio das classes na predição. Esse resultado ressalta a importância de se considerar o impacto global das variáveis no modelo.

Em suma, os achados corroboram a necessidade de se implementar estratégias de diagnóstico precoce e ampliar o acesso ao tratamento especializado, principalmente em regiões menos assistidas. No contexto do SUS, onde disparidades regionais e limitações de recursos são desafios frequentes, a aplicação de modelos preditivos pode auxiliar na identificação de pacientes com risco de desfecho de doença desfavorável, bem como a investigação de seus atributos pode fornecer insights para alvos de intervenção para prevenção deste desfecho, como por exemplo indicar necessidade desenvolvimento de programas de triagem e de estratégias de prevenção de doença.

O algoritmo *Random Forest* apresentou melhor desempenho na predição de status final após tratamento oncológico. Na prática clínica, uma acurácia de 76,2% pode ser insuficiente, pois a falha em identificar corretamente pacientes com alto risco de óbito pode ter consequências no manejo terapêutico. Porém, esta solução se mostra como ponto de partida para novas estratégias. Adicionalmente, o uso de dados multimodais, incluindo informações genômicas, poderia enriquecer ainda mais os modelos desenvolvidos, proporcionando predições mais acuradas e personalizadas. Estudos futuros devem explorar a inclusão de novos atributos, como perfil molecular dos tumores, além de investigar a aplicabilidade dos modelos em coortes separadas por regiões em virtude de seu perfil distinto, de modo a garantir uma melhor generalização e robustez dos resultados para as diferentes populações.

**Considerações finais**

Os modelos de *machine learning* demonstram um potencial significativo para a predição de desfechos clínicos em pacientes oncológicos, oferecendo subsídios importantes para a personalização do tratamento e a formulação de políticas públicas. No entanto, a eficácia desses modelos está intrinsecamente relacionada à qualidade e completude dos dados utilizados, uma vez que informações incompletas ou inconsistentes podem comprometer a performance preditiva e limitar sua aplicabilidade em contextos clínicos.

A implementação desses algoritmos no planejamento de políticas de saúde e no monitoramento de pacientes deve ser conduzida com cautela, devido ao risco de previsões incorretas que podem resultar em decisões inadequadas tanto no manejo terapêutico quanto na alocação de recursos. Assim, é imperativo que essas ferramentas sejam empregadas de forma responsável, atuando como suporte ao processo de decisão clínica, e não como substitutas de avaliação médica direta.

Investimentos contínuos em infraestrutura de dados, incluindo a padronização e a integração de fontes de informação, bem como o aprimoramento dos modelos por meio de abordagens multimodais e ajustes de hiperparâmetros, são fundamentais para maximizar o impacto positivo dos algoritmos no tratamento oncológico. Com uma base sólida de dados e com a devida consideração ética, esses modelos podem desempenhar um papel crucial no desenvolvimento de estratégias mais eficazes e personalizadas para a assistência e tratamento do câncer.

**Agradecimentos**

Gostaria de agradecer ao meu orientador, Pedro Oiticica, pelo suporte constante e orientação valiosa. Ao meu colega André Pinto, cujas influências, inspiração e incentivo foram fundamentais para que eu iniciasse e concluísse este curso e estudo. Ao meu amigo e estatístico Welinton Hirai, por seus conselhos técnicos e insights valiosos. Agradeço também aos meus pais, avós, e à minha bisavó Dona Mariinha, por serem a minha base e inspiração. E, especialmente aos meus companheiros diários, Bruno e Dudu, obrigada pelo amor, apoio e incentivo diário ao meu crescimento.

**Referências**

Awad, M.; & Fraihat, S. 2023. Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. *J. Sens. Actuator Networks* 12(5).

Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.

Cardone, C.; & Arnold, D. 2023. The Cancer Treatment Gap in Lower- to Middle-Income Countries. *Oncology* 101(Suppl. 1): 2–4.

Chen, T.; & Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016: 785–794.

Hastie, T.; Tibshirani, R.; & Friedman, J. 2009. The Elements of Statistical Learning. Springer Series in Statistics .

INCA - Instituto Nacional de Câncer ([s.d.]) *Estimativa 2023: incidência de câncer no Brasil*.

Jacob Kligerman 2001. Registro Hospitalar de Câncer no Brasil. *Revista Brasileira de Cancerologia* 47(4): 357–359.

James, G.; Witten, D.; Hastie, T.; & Tibshirani, R. 2021. An Introduction to Statistical Learning. Springer Texts in Statistics .

Jomar, R. T.; Velasco, N. S.; Mendes, G. L. Q.; Guimaraes, R. M.; de Oliveira Fonseca, V. A.; & Meira, K. C. 2023. Fatores associados ao tempo para submissão ao primeiro tratamento do câncer de mama. *Ciência & Saúde Coletiva* 28(7): 2155–2164.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28(5): 1–26.

Lundberg, S. M.; Allen, P. G.; & Lee, S. I. ([s.d.]) A Unified Approach to Interpreting Model Predictions.

Organização Mundial da Saúde ([s.d.]) *CID-O Classificação Internacional de Doenças para Oncologia*.

Raoof, S.; Lee, R. J.; Jajoo, K.; Mancias, J. D.; Rebbeck, T. R.; & Skates, S. J. 2022. Multicancer Early Detection Technologies: A Review Informed by Past Cancer Screening Studies. *Cancer Epidemiology Biomarkers and Prevention* 31(6): 1139–1145.

Saesen, R.; Van Hemelrijck, M.; Bogaerts, J.; Booth, C. M.; Cornelissen, J. J.; Dekker, A.; Eisenhauer, E. A.; Freitas, A.; Gronchi, A.; Hernán, M. A.; Hulstaert, F.; Ost, P.; Szturz, P.; Verkooijen, H. M.; Weller, M.; Wilson, R.; Lacombe, D.; & van der Graaf, W. T. 2023. Defining the role of real-world data in cancer clinical research: The position of the European Organisation for Research and Treatment of Cancer. *European Journal of Cancer* 186: 52–61.

Schonlau, M.; & Zou, R. Y. 2020. The random forest algorithm for statistical learning. *Stata Journal* 20(1): 3–29.

Sherman, R. E.; Anderson, S. A.; Dal, G. J.; Gray, G. W.; Gross, T.; Hunter, N. L.; Lavange, L.; Marinac-Dabic, D.; Marks, P. W.; Robb, M. A.; Shuren, J.; Temple, R.; Woodcock, J.; Yue, L. Q.; & Califf, R. M. 2016. Real-World Evidence-What Is It and What Can It Tell Us? *NEJM* .

Štrumbelj, E.; & Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41(3): 647–665.

Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; & Bray, F. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians* 71(3): 209–249.

Tope, P.; Farah, E.; Ali, R.; El-Zein, M.; Miller, W. H.; & Franco, E. L. 2023. The impact of lag time to cancer diagnosis and treatment on clinical outcomes prior to the COVID-19 pandemic: a scoping review of systematic reviews and meta-analyses. *eLife* 12.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; D’, L.; Mcgowan, A.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Lin Pedersen, T.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; & Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43): 1686.

Zhang, P.; Jia, Y.; & Shang, Y. 2022. Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks* 18(6).

**Apêndice ou Anexo**

A incluir: o dicionário de dados e código de tratamento de dados categóricos.