**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 aos 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. Neste contexto, os algoritmos *Random Forest* e *XGBoost* foram aplicados para prever o status clínico de pacientes oncológicos com base em dados de 2017 a 2021 Registros Hospitalares de Câncer (RHC). Os modelos foram treinados e validados com dados de mundo real, e a performance foi avaliada por meio de métricas de acurácia, precisão, *recall* e F1-score. Os valores SHAP foram utilizados para interpretar a contribuição de cada variável para os resultados preditivos. O modelo *Random Forest* atingiu uma acurácia de 59,52%, prevendo corretamente o status clínico de 17.040 pacientes, enquanto o *XGBoost* alcançou uma acurácia de 58,99%. Ambos os modelos destacaram o estadiamento clínico como a variável mais influente na predição de desfechos clínicos desfavoráveis. Fatores regionais e o tipo de instituição de tratamento também tiveram um papel significativo nas predições. A análise SHAP confirmou a influência crítica das variáveis clínicas e regionais nos resultados preditivos. A aplicação de modelos de aprendizado de máquina a dados de mundo real demonstrou poder preditivo moderado para o status clínico pós-tratamento de pacientes oncológicos. Este estudo enfatiza a importância das disparidades regionais na infraestrutura de saúde, sugerindo que esses fatores devem ser considerados no planejamento de políticas públicas e na alocação de recursos para o tratamento oncológico.

**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 treatment is essential to achieving better clinical outcomes. Continuous investments in data infrastructure and model refinement are crucial to maximizing their utility in health policy planning and cancer treatment. The Random Forest and XGBoost algorithms were applied to predict the clinical status of cancer patients based on data from the 2017 to 2021 from Cancer Hospital Registries (RHC). The models were trained and validated using real-world data, and their performance was assessed through accuracy, precision, recall, and F1-score metrics. SHAP values were used to interpret the contribution of each variable to the predictive outcomes. The Random Forest model achieved an accuracy of 59.52%, correctly predicting the clinical status of 17,040 patients, while XGBoost reached an accuracy of 58.99%. Both models highlighted clinical staging as the most influential variable in predicting unfavorable outcomes. Regional factors and the type of treatment institution also played a significant role in these predictions. The SHAP analysis confirmed the critical influence of clinical and regional variables on the predictive outcomes. The application of machine learning models to real-world data demonstrated moderate predictive power for post-treatment clinical status in cancer patients. This study emphasizes the importance of regional disparities in healthcare infrastructure, suggesting that these factors should be considered in public policy planning and resource allocation for cancer treatment.

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

**Introdução**

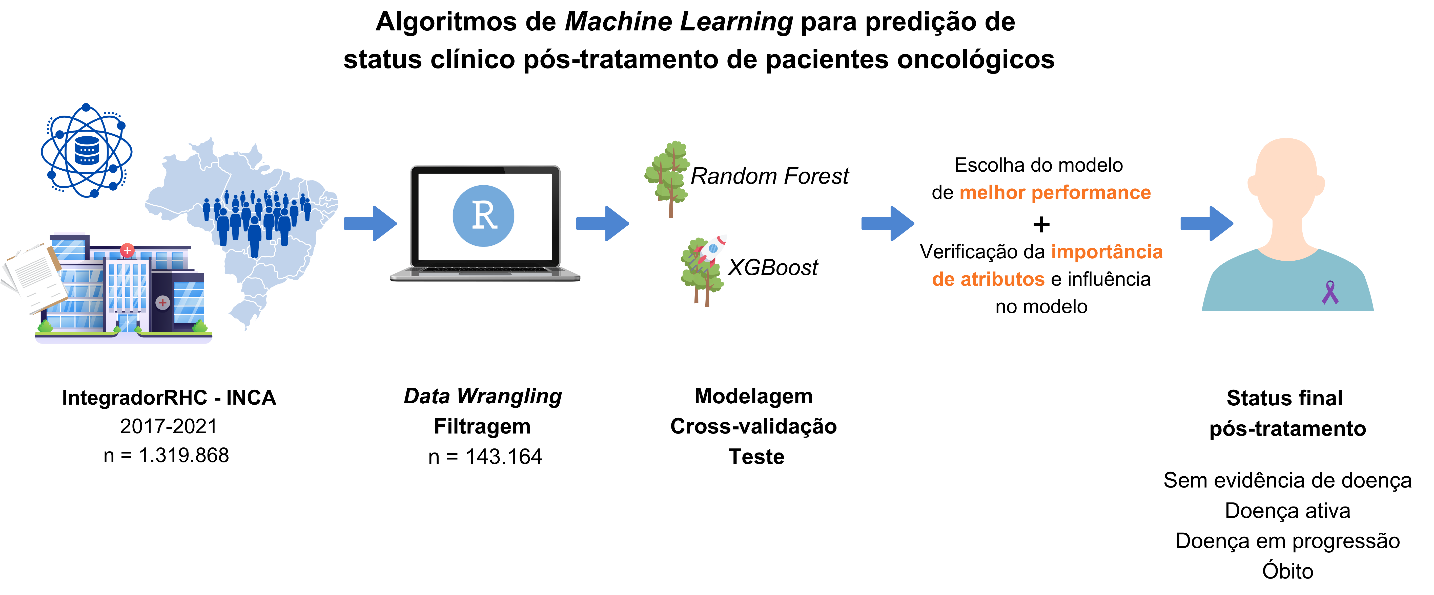
O câncer é uma doença multifatorial e uma das principais questões de saúde pública global, apresentando um aumento significativo nas taxas de incidência e mortalidade nas últimas décadas(INCA - Instituto Nacional de Câncer, [s.d.]; Sung et al., 2021). Esse cenário impõe desafios aos sistemas de saúde 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.

Este trabalho tem como objetivo principal utilizar dados de mundo real de RHCs para treinar algoritmos de aprendizado de máquina e verificar sua performance na predição de status clínico de pacientes oncológicos após a conclusão do 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, diagnóstico precoce e disparidades regionais (Figura 1).



**Figura 1.** Graphical Abstract do estudo.

**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 coletados compreenderam os anos de 2017 a 2021, em formato estruturado (.dbf), compreendendo 1.319.868 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 e disponibilizadas em repositório público *GitHub (*https://github.com/alexiapolos/predicao\_status\_rhc.git*)*.

*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 quando comparados 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.

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 (N = 1.039.245), com agrupamento e balanceamento de categorias para as análises subsequentes. A lista completa de agrupamentos está descrita no código disponível no *GitHub*.

Os atributos selecionados para análises e suas respectivas categorias se encontram descritas na Tabela 1. As instâncias contendo dados faltantes foram removidas da base de análise, resultando em um banco de dados de 143.164 observações.

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

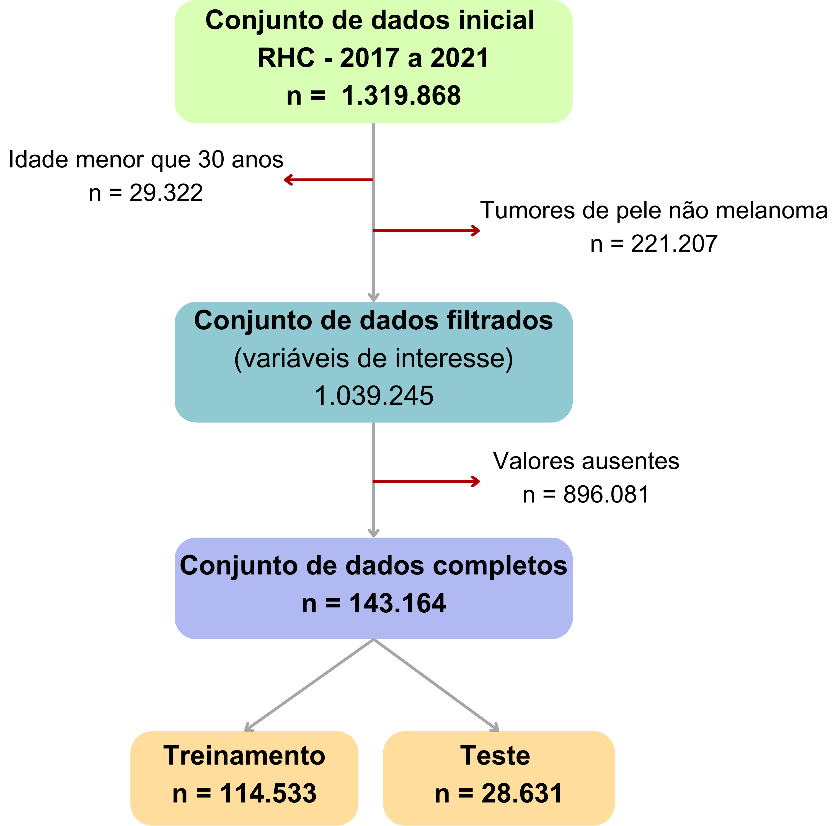
|  |  |  |
| --- | --- | --- |
| Variável | Tipo | Níveis |
| faixa\_etaria | Fator | "30-40 anos", "41-50 anos", "51-60 anos", "61-70 anos", "Mais de 70 anos" |
| sexo | Fator | "Masc", "Fem" |
| tabagismo | Fator | "Ex-tabagista", "Nunca tabagista", "Tabagista ativo" |
| etnia | Fator | "Amarelos e indígena", "Branca", "Pardos e pretos" |
| hist\_familiar | Fator | “Não”, “Sim” |
| 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\_final\_trat | Fator | "Doença ativa", "Em progressão/paliativo", "Óbito", "Sem evidência de doença/remissão completa" |
| estadiamento | Fator | "0 e I", "II", "III", "IV" |
| regiao | Fator | "Centro-Oeste", "Nordeste", "Norte", "Sudeste", "Sul" |
| tumores\_agrupado | Fator | "Sistema Digestivo", "Sistema respiratório", "Mama e Ginecológicos", "Sistema urológico", "Melanomas, Sarcomas e SNC", "Sistema hematopoético”, “Outros". |

*Análise exploratória e descritiva*

Foi realizada uma análise exploratória destes dados filtrados, na qual foi avaliada a frequência das variáveis categóricas, bem como aplicado o teste de *Crámer’s V* para quantificação da força da associação entre cada variável categórica e a variável resposta (status\_doença\_final\_trat). O coeficiente de *Cramér’s V* é uma medida estatística que varia de 0 a 1, onde valores próximos a 0 indicam nenhuma associação e valores próximos a 1 indicam uma associação forte (Harald Cramér, 1946).

*Treinamento do modelo, ~~e~~ validação cruzada e teste `*

Para avaliação da capacidade preditiva dos modelos a modelagem consistiu em 3 etapas: treinamento, validação cruzada e teste. O particionamento do banco de dados foi realizado na proporção de 80% (n = 114.533) para o conjunto de treinamento e validação cruzada e 20% (n = 28.631) para o conjunto de teste (Figura 2). Esta etapa foi feita de maneira estratificada com relação a variável resposta (status\_doença\_final\_trat), de maneira a minimizar os efeitos do desbalanceamento das categorias no modelo.



**Figura 2.** 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 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 |
| *ntree* | Número de árvores | 100, 200, 500 |
| *mtry* | Número de variáveis consideradas para divisão em cada nó da árvore | 3 ,4, 5 |
| *splitrule* | Critério utilizado para divisão dos nós | Índice de Gini |
| *min.node.size* | Número mínimo de observações que cada nó deve conter | 1 e 4 |

**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 4 *folds* em ambos os modelos. 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. *Recall* (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.

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*.

*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 “S*Hapley 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* (v2.4.3) em R.

Em seguida, foi realizada uma Análise de Correspondência Múltipla (MCA) para explorar as relações entre variáveis categóricas de cunho regional e assistencial, respeito a Lei dos 60 dias e os desfechos clínicos dos pacientes após o tratamento. A análise foi realizada utilizando o pacote *FactoMineR* (v2.11) em R.

**Resultados e Discussão**

*Análise exploratória e descritiva*

A análise descritiva dos dados do RHC (2017-2021) acerca dos atributos selecionados se encontra apresentada na Tabela 4. No registro de 143.164 pacientes, foi observado que a maioria dos casos de câncer ocorrem nas faixas etárias a partir de 60 anos (52%), com predominância do sexo feminino (59%) e casos virgens de tratamento oncológico (87%). Cerca de 63% dos pacientes são de origem da região sudeste e sul do país. Com relação a etnia, eram em sua maioria pardos e pretos (56%), nunca tabagistas (57%), e 54% com histórico familiar de câncer. Foram mais prevalentes na casuística casos de tumores de mama e ginecológicos (43%) e do trato digestivo (23%), diagnosticados com doença locorregional ou avançada (53%), em estadiamentos clínicos III ou IV. Cerca de 40% receberam abordagem de tratamento inicial cirúrgica, e 45% se apresentam com status de doença ativa após tratamento.

**Tabela 4.** Tabela de frequências dos casos dos anos de 2017 a 2021.

|  |  |  |
| --- | --- | --- |
| Variável | Categoria | N (%) |
| Faixa etária | 31-40 anos | 12.185 (8,5%) |
| 41-50 anos | 21.954 (15%) |
| 51-60 anos | 34.238 (24%) |
| 61-70 anos | 40.193 (28%) |
| Mais de 70 anos | 34.594 (24%) |
| Sexo | Masc | 58.201 (41%) |
| Fem | 84.963 (59%) |
| Tipo de caso | Analítico | 123.964 (87%) |
| Não analítico | 19.200 (13%) |
| Etnia | Amarelos e indígena | 1.239 (0,9%) |
| Branca | 61.728 (43%) |
| Pardos e pretos | 80.197 (56%) |
| Tabagismo | Ex-tabagista | 34.401 (24%) |
| Nunca tabagista | 82.016 (57%) |
| Tabagista ativo | 26.747 (19%) |
| Histórico familiar | Não | 65.853 (46%) |
| Sim | 77.311 (54%) |
| Estado de residência | AC | 190 (0,1%) |
| AL | 2.967 (2,1%) |
| AM | 2.543 (1,8%) |
| AP | 164 (0,1%) |
| BA | 11.972 (8,4%) |
| CE | 3.649 (2,5%) |
| DF | 1.506 (1,1%) |
| ES | 4.027 (2,8%) |
| GO | 302 (0.2%) |
| MA | 3.373 (2,4%) |
| MG | 46.701 (33%) |
| MS | 978 (0,7%) |
| MT | 883 (0.6%) |
| PA | 2.157 (1,5%) |
| PB | 684 (0,5%) |
| PE | 10.462 (7,3%) |
| PI | 628 (0,4%) |
| PR | 21.235 (15%) |
| RJ | 5.630 (3,9%) |
| RN | 3.080 (2,2%) |
| RO | 3.236 (2,3%) |
| RR | 117 (<0,1%) |
| RS | 9.508 (6,6%) |
| SC | 3.667 (2,6%) |
| SE | 190 (0,1%) |
| SP | 165 (0,1%) |
| TO | 3.150 (2,2%) |
| Região | Centro-Oeste | 3.669 (2,6%) |
| Nordeste | 37.005 (26%) |
| Norte | 11.557 (8,1%) |
| Sudeste | 56.523 (39%) |
| Sul | 34.410 (24%) |
| Respeito a Lei dos 60 dias | Não | 74.145 (52%) |
| Sim | 69.019 (48%) |
| Tratamento inicial | Cirurgia | 56.857 (40%) |
| Outro | 13.743 (9,6%) |
| Quimioterapia | 48.084 (34%) |
| Radioterapia | 24.480 (17%) |
| Unidade de tratamento | CACON | 52.169 (36%) |
| Outra | 36 (<0,1%) |
| UNACON | 90.959 (64%) |
| Estadiamento clínico ao diagnóstico | 0 | 5.721 (4,0%) |
| I | 22.517 (16%) |
| II | 39.476 (28%) |
| III | 38.482 (27%) |
| IV | 36.968 (26%) |
| Localização do tumor | Mama e Ginecológicos | 61.232 (43%) |
| Melanomas, Sarcomas e SNC | 1.747 (1,2%) |
| Sistema Digestivo | 32.271 (23%) |
| Sistema Hematopoético | 1.336 (0,9%) |
| Sistema Respiratório | 17.830 (12%) |
| Sistema Urológico | 28.748 (20%) |
| Status doença após tratamento | Doença ativa | 63.923 (45%) |
| Em progressão/paliativo | 21.351 (15%) |
| Óbito | 15.862 (11%) |
| Sem evidência/remissão1 | 42.028 (29%) |

N - número de casos., 1Sem evidência de doença/remissão completa

Foi realizado o teste de Cramér’s V para verificar a associação entre os atributos e a variável resposta a ser predita pelo modelo (status\_doenca\_final\_trat). Os resultados indicaram uma forte associação com a variável tipo de caso (se virgem ou não de tratamento), coeficiente 0,887), tabagismo (coeficiente = 0,866) e moderada com a variável etnia (coeficiente = 0,442). As variáveis respeito à Lei dos 60 dias e estadiamento clínico também mostraram associações moderadas, com coeficientes de 0,317 e 0,326, respectivamente. As demais variáveis apresentaram fracas associações, conforme apresentado na Tabela 5.

**Tabela 5.** Coeficientes de Cramér's V para as variáveis categóricas em relação ao status da doença após o tratamento oncológico.

|  |  |
| --- | --- |
| Variável | Coeficiente de Cramer’s V |
| Tipo de caso | 0,877 |
| Tabagismo | 0,866 |
| Etnia | 0,442 |
| Estadiamento clínico | 0,326 |
| Histórico familiar | 0,311 |
| Localização do tumor | 0,180 |
| Tratamento inicial | 0,152 |
| Estado de residência | 0,215 |
| Respeito a Lei dos 60 dias | 0,117 |
| Sexo | 0,117 |
| Unidade de tratamento | 0,097 |
| Região | 0,097 |
| Faixa etária | 0,063 |

*Modelagem Random Forest*

O modelo Random Forest foi ajustado com 114.533 amostras e 13 variáveis preditoras. O processo de *resampling* foi realizado por meio de validação cruzada em 4 *folds,* com reposição. O modelo de melhor acurácia média entre as predições foi treinado em 500 árvores, utilizando 5 variáveis na divisão dos nós das árvores, e com no mínimo 1 observação necessárias para cada nó.

*Modelagem XGBoost*

O modelo *XGBoost* também foi ajustado com 114.533 amostras e 13 variáveis preditoras. Assim como no *Random Fores*t, o processo de *resampling* foi realizado por meio de validação cruzada em 4 *folds,* também com reposição. O modelo com melhor desempenho foi ajustado utilizando 99 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 80,14%, controlando o impacto de cada árvore adicionada no modelo final, enquanto o parâmetro *gamma* foi configurado para 0,331, restringindo a divisão dos nós a casos em que há um ganho significativo. Ainda, considerou-se 2 como um peso mínimo de instâncias, o que ajuda a evitar divisões de nós com pouca representatividade. 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 79,42% das variáveis em cada árvore e utilizar 64,67% das amostras disponíveis para cada iteração, o que contribuiu para a robustez e estabilidade.

Foi calculada a acurácia média e desvio padrão entre os *folds* de cross-validação, que se encontram descritos na Tabela 6.

**Tabela 6.** Valores de acurácia média e desvio padrão entre os folds para ambos os modelos Random Forest e XGBoost.

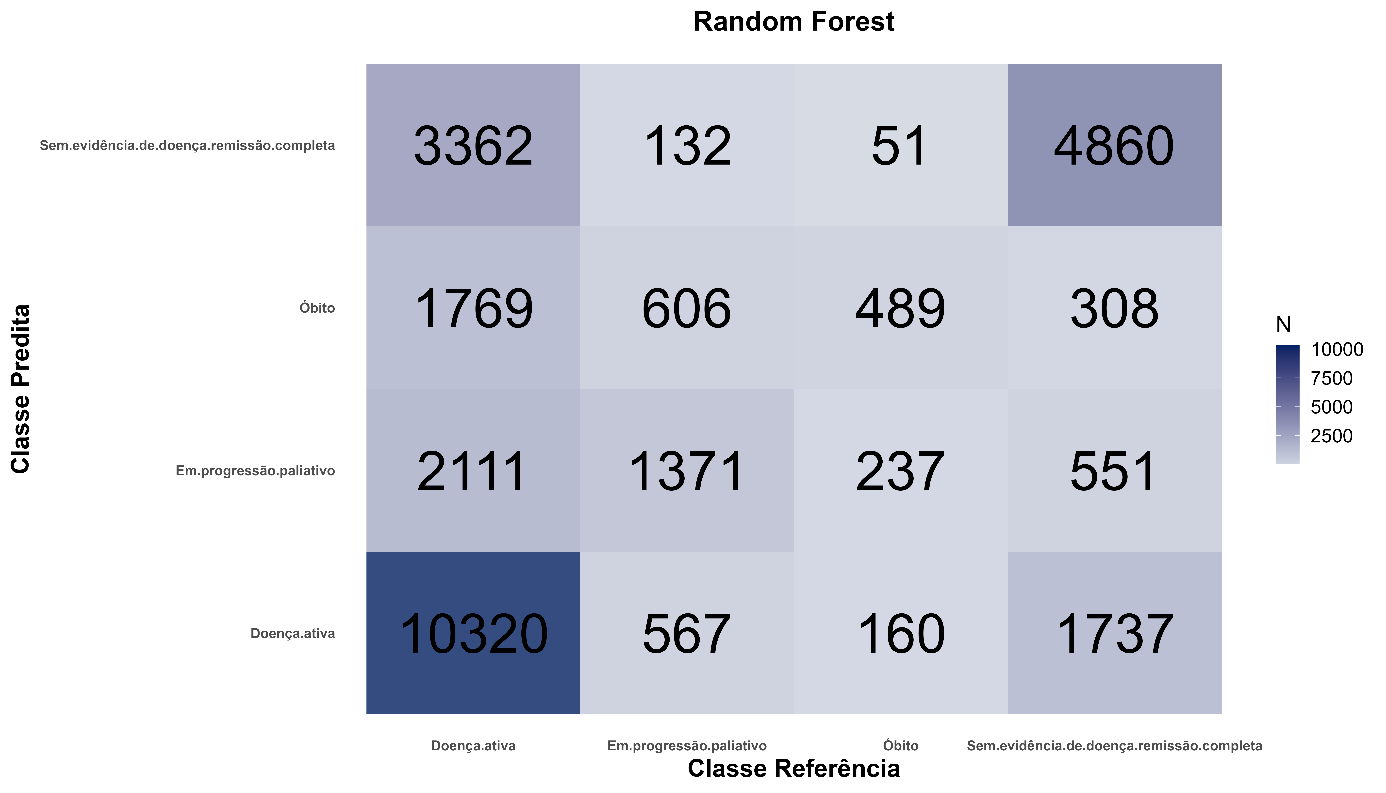
|  |  |
| --- | --- |
| Modelo | Acurácia média ± Desvio padrão |
| *Random Forest* | 59,72 ± 0,36 |
| *XGBoost* | 59,31 ± 0,22 |

*Teste e avaliação de performance*

Após o treinamento, ambos os modelos, foram aplicados a um conjunto de dados de teste independente para avaliar sua performance final na predição do status clínico de pacientes oncológicos. As variáveis críticas de pior desfecho após tratamento são “Em progressão/paliativo” e “Óbito”, e são estas as classes de maior interesse em predição assertiva do status do paciente.

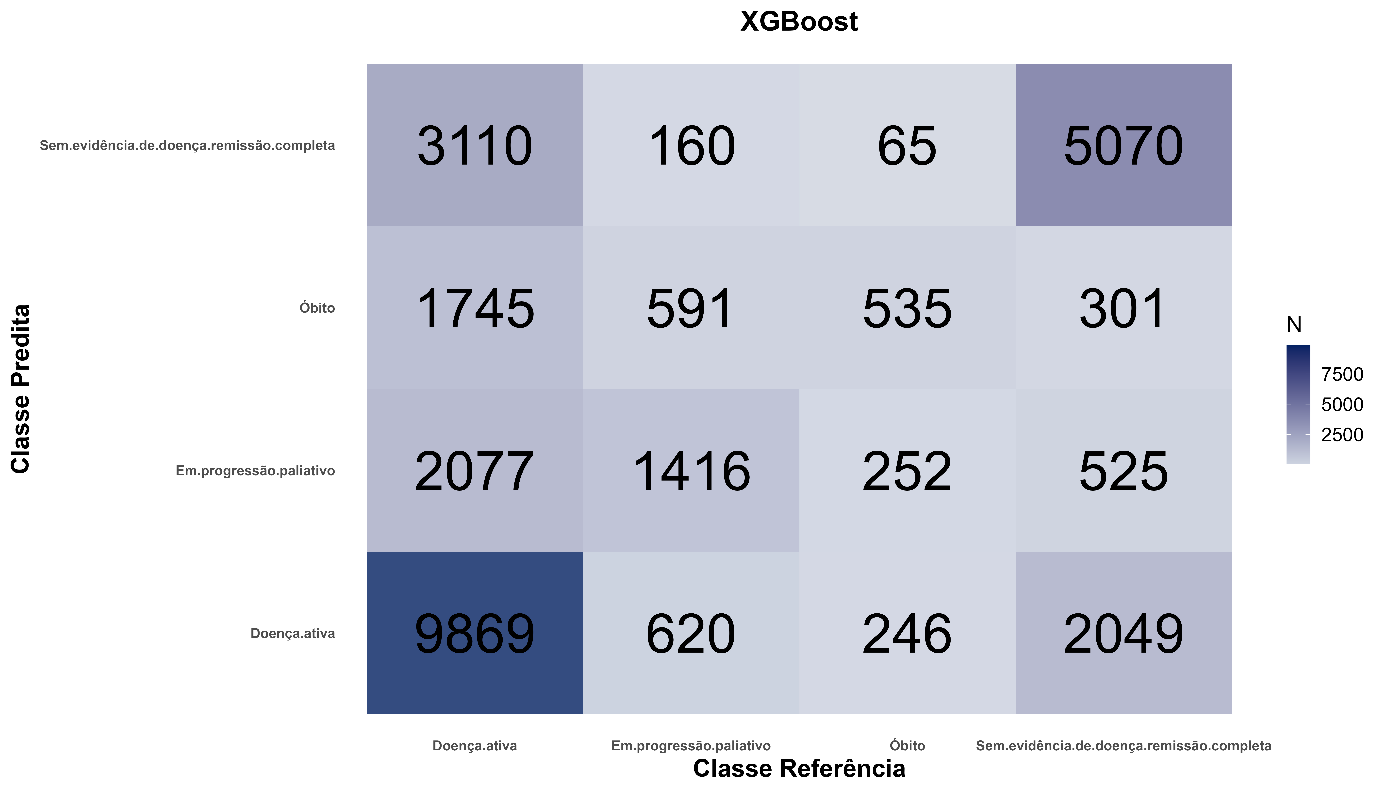
As métricas de desempenho incluem acurácia geral, precisão, *recall*, especificidade e *F1-score* para as quatro classes da variável de interesse: “Doença ativa”, “Em progressão/paliativo”, “Óbito” e “Sem evidência de doença/remissão completa”.

O modelo *Random Forest* (Figura 4) previu corretamente o status de 17.040 pacientes, demonstrando uma acurácia média de 59,52%. No entanto, erros de classificação ocorreram, principalmente entre as classes “Doença ativa” e “Sem evidência de doença/remissão completa” (Figura 3).



**Figura 3 .** Matriz de confusão da predição de desfechos clínicos no conjunto de teste para o modelo Random Forest de melhor performance. As cores refletem o número de observações classificadas em cada classe, conforme o gradiente de cores.

O modelo *XGBoost* previu corretamente o status de 16.890 casos, demonstrando uma acurácia média de 58,99%. Assim como no modelo *Random Forest*, ocorreram erros de classificação principalmente entre as classes “Doença ativa” e Sem evidência de doença/remissão completa” (Figura 4).



**Figura 4.** Matriz de confusão da predição de desfechos clínicos no conjunto de teste para o modelo XGBoost de melhor performance. As cores refletem o número de observações classificadas em cada classe, com o gradiente de cores.

De maneira geral, as análises sugerem que o modelo *Random Forest* demonstrou uma sensível vantagem em termos de acurácia global. Esta vantagem, apesar de pequena em termos percentuais, representou a classificação assertiva de 150 pacientes em tratamento de câncer.

Ao avaliar as métricas para predição por classe (Tabela 7), considerando que temos um maior interesse na predição assertiva de classes de pior desfecho, o modelo *Random Forest* teve uma especificidade de 98,24% para classe “Óbito”, indicando que foi eficaz em evitar falsos positivos. No entanto, o *recall* de 15,42% indica que uma proporção considerável casos de óbito não foi identificado corretamente. Já o modelo *XGBoost* apresentou uma especificidade ligeiramente inferior para a classe “Óbito” (97,79%), mas seu recall foi ligeiramente superior (16,87%).

**Tabela 7.** Métricas de desempenho dos modelos ajustados Random Forest e XGBoost na predição do status clínico do paciente após o tratamento oncológico.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Modelo | Especificidade | Precisão | *Recall* | *F1-score* |
| Doença ativa | *Random Forest* | 54,30% | 58,76% | 80,73% | 68,02% |
|  | *XGBoost* | 56,26% | 58,74% | 77,20% | 66,72% |
| Em progressão/ paliativa | *Random Forest* | 94,64% | 51,23% | 32,11% | 39,48% |
|  | *XGBoost* | 94,37% | 50,81% | 33,16% | 40,13% |
| Óbito | *Random Forest* | 98,24% | 52,19% | 15,42% | 23,80% |
|  | *XGBoost* | 97,79% | 48,72% | 16,87% | 25,06% |
| Sem evidência/  remissão1 | *Random Forest* | 87,17% | 65,18% | 57,82% | 61,28% |
|  | *XGBoost* | 85,79% | 63,81% | 60,32% | 62,02% |

1Sem evidência de doença/ em remissão completa.

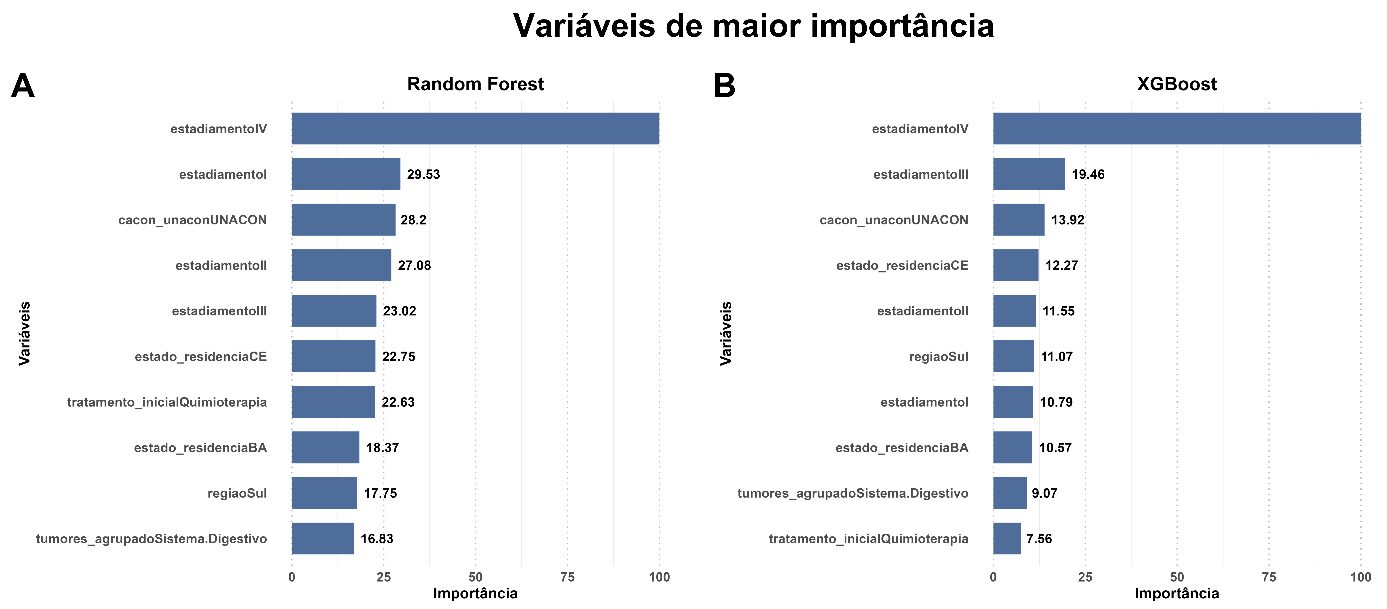
Para predição da classe “Em progressão/paliativo”, o modelo *Random Forest* apresentou uma especificidade de 94,64% e um *recall* de 32,11%. Já o modelo *XGBoost*, apresentou especificidade de 94,37% e um *recall* de 33,16%.

No que tange às predições por classe, o modelo *Random Forest* demonstrou uma melhor capacidade de evitar falsos positivos, enquanto o *XGBoost* se destacou na identificação de casos verdadeiros positivos para a classe “Óbito”.

*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, podem-se destacar também a unidade de tratamento do paciente, estado de residência e região, e tratamento inicial. A Figura 6 ilustra as 10 variáveis que apresentaram uma maior importância relativa em ambos os modelos.



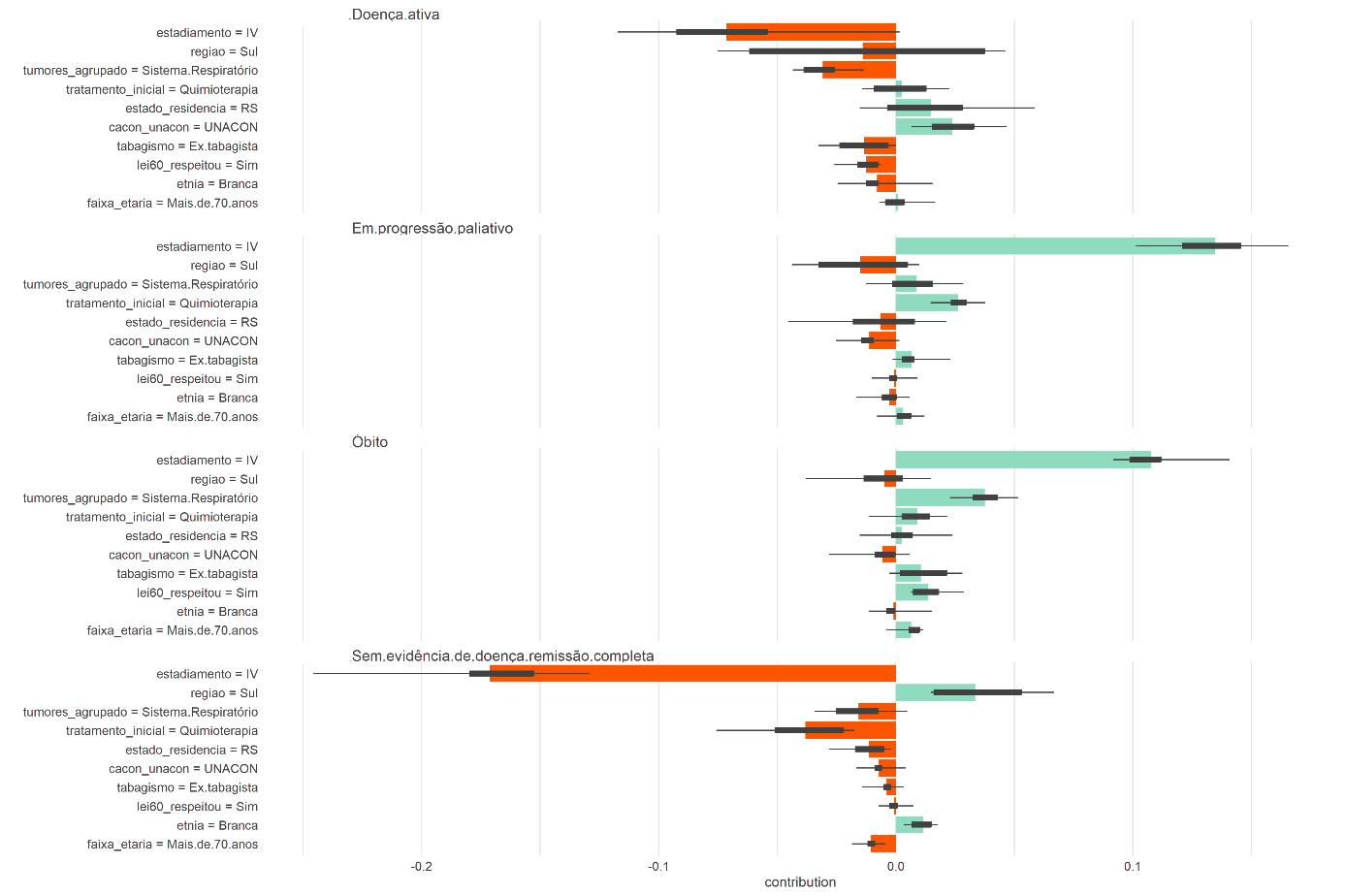
**Figura 5.** Gráfico de importância das variáveis para os modelos Random Forest (A) e XGBoost (B) com valores relativos.

A análise de importância das variáveis revela que o modelo *Random Forest* apresentou um melhor aproveitamento das variáveis disponíveis, distribuindo de forma mais equilibrada a relevância entre um maior número de preditores. Assim, este modelo se propõe a explorar de maneira mais eficaz as variáveis presentes atribuindo-lhes importância significativa, o que lhe confere uma maior flexibilidade e robustez para identificar padrões complexos e sutis nos dados. Por este motivo, o *Random Forest* foi escolhido como o modelo final.

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

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

Em consonância com a análise prévia de importância das variáveis, o estadiamento clínico destacou-se como a variável mais influente na predição de todas as classes, indicando que diferentes níveis de estadiamento têm um impacto substancial na previsão do desfecho clínico dos pacientes. De modo geral, o estadiamento avançado (especialmente o estágio IV) apresentou uma contribuição expressiva para a probabilidade de piora no estado clínico, particularmente nas classes "Em progressão/paliativo" e "Óbito", conforme denota a Figura 7.



**Figura 6.** 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.

Na classe "Em progressão/paliativo", o estadiamento IV emergiu como a variável de maior influência, seguido pela região Sul e pelo tratamento inicial com quimioterapia, sugerindo uma associação entre essas variáveis e a progressão da doença ou a necessidade de cuidados paliativos. Para a classe "Óbito", o estadiamento IV também foi a principal variável, aumentando significativamente a probabilidade de predição de mortalidade. Outros fatores, como o estado de residência na região Sul e o tratamento em unidades de alta complexidade (UNACON), também contribuíram positivamente para a predição de óbito, o que pode refletir diferenças regionais nos cuidados de saúde e características clínicas específicas.

Esses resultados reforçam a importância do estadiamento clínico, bem como de fatores regionais e relacionados ao tratamento, na determinação do prognóstico de pacientes oncológicos. Além disso, evidenciam a capacidade do modelo Random Forest, interpretado via SHAP, de identificar variáveis-chave que influenciam negativamente os desfechos clínicos, fornecendo uma base sólida para o desenvolvimento de estratégias mais direcionadas na oncologia.

*Análise de Correspondência Múltipla*

De maneira complementar a análise dos valores SHAP, foi realizada uma MCA para explorar as associações entre as variáveis regionais, como região de residência, tipo de instituição de tratamento e o cumprimento da Lei dos 60 dias para com o status dos pacientes após o tratamento (Figura 7).

*Gráfico, Gráfico de dispersão

Descrição gerada automaticamente com confiança média*

Figura 7. Análise de Correspondência Múltipla, ilustrando o agrupamento de pacientes oncológicos com base nas variáveis regionais e tratamento.

Nesta análise, observou-se que os pacientes da região Sul, tratados em unidades de menor complexidade (excluindo-se CACONs e UNACONs), tendem a se agrupar na categoria "Óbito". O cumprimento da Lei dos 60 Dias (categoria “Sim”) aparece entre os desfechos de "Óbito" e "Sem evidência de doença", sugerindo que, embora o respeito à legislação exerça um impacto positivo no prognóstico, esse efeito é limitado e influenciado por outras variáveis.

Por outro lado, o não cumprimento da Lei dos 60 Dias está posicionado distante dos desfechos de "Óbito" e "Sem evidência de doença", associando-se mais fortemente à categoria "Doença ativa". Isso indica que a falta de cumprimento da legislação tende a estar ligada a um prognóstico intermediário, no qual os pacientes têm menos chances de alcançar a remissão completa, com possíveis influências adicionais de fatores clínicos e regionais.

Além disso, as regiões Nordeste e as unidades CACON aparecem distantes tanto das categorias "Sim" quanto "Não" em relação à Lei dos 60 Dias. Esse distanciamento sugere que, nesses contextos, o impacto direto da legislação é menos relevante, possivelmente porque a infraestrutura e os recursos especializados dessas unidades de alta complexidade compensam atrasos no início do tratamento. Essa observação ressalta a importância de outros fatores, como a qualidade do atendimento e a disponibilidade de recursos, que podem influenciar os desfechos clínicos independentemente do tempo de espera.

Esses achados reforçam que, além do estadiamento clínico, fatores regionais e a complexidade do centro de tratamento desempenham um papel crucial na definição dos desfechos clínicos, ressaltando a importância de políticas públicas que garantam o acesso a tratamentos de alta complexidade em todo o território brasileiro.

*Discussão*

Os resultados deste estudo indicam que os modelos Random Forest e XGBoost possuem uma capacidade moderada de prever os desfechos clínicos de pacientes oncológicos com base em variáveis sociodemográficas e clínicas. Embora a acurácia não tenha sido extremamente alta, os modelos destacaram importantes fatores preditivos que podem auxiliar no planejamento de políticas públicas e intervenções clínicas mais direcionadas.

O modelo *Random Forest* demonstrou um desempenho ligeiramente superior em relação ao *XGBoost* na predição do status clínico dos pacientes. A acurácia média do *Random Forest* foi de 59,52%, comparada a 58,99% do *XGBoost*. Apesar da pequena diferença em termos percentuais, este modelo representou a classificação correta de 150 pacientes adicionais, destacando a robustez do *Random Forest* em cenários clínicos onde a acurácia marginal pode ser decisiva.

Ao avaliar a predição das classes "Óbito" e "Em progressão/paliativo" — classes mais críticas para predição de desfechos desfavoráveis — o *Random Forest* se mostrou mais eficaz na redução de falsos positivos para a classe "Óbito" quando comparada ao *XGBoost*, sugeriando que, em cenários onde a detecção de casos críticos é crucial, o primeiro modelo apresenta maior precisão em evitar diagnósticos incorretos de óbito.

O estadiamento clínico, especialmente o estágio IV, foi consistentemente identificado como o principal preditor dos piores desfechos clínicos, como óbito e progressão da doença. Esses achados estão em consonância com a literatura, que indica que o câncer em estágio avançado está associado a prognósticos mais graves e menor chance de remissão. Estes resultados enfatizam a necessidade de diagnóstico precoce do câncer e a importância de estratégias de rastreamento e prevenção da doença, cruciais para identificação de casos iniciais, em um contexto no qual as chances de tratamento bem-sucedido são mais favoráveis.

A análise dos valores SHAP reforça que, além do estadiamento, fatores regionais, como a região Sul e o tipo de instituição de tratamento, também influenciam significativamente os desfechos clínicos. A associação entre pacientes da região Sul e maiores taxas de óbito sugere possíveis disparidades regionais nos cuidados oncológicos, que podem estar relacionadas tanto a fatores socioeconômicos quanto à estrutura de saúde disponível. Essa conclusão é corroborada pelos resultados da MCA, que mostraram um agrupamento claro desses pacientes na categoria de óbito.

Adicionalmente, os resultados do teste de Cramér’s V fornecem mais evidências sobre a relevância de variáveis clínicas específicas na predição do desfecho, indicando forte associação, principalmente entre a as variáveis "tipo de caso" (se o paciente já havia passado por tratamento ou não) e tabagismo com o desfecho clínico do paciente. Esses resultados sugerem que pacientes que nunca haviam recebido tratamento ou que tinham histórico de tabagismo apresentaram maior risco de piores desfechos.

Além disso, foi avaliado o impacto da Lei dos 60 Dias, que estabelece o prazo máximo para o início do tratamento oncológico após o diagnóstico. No entanto, os resultados sugerem que o cumprimento dessa legislação não foi determinante para alterar significativamente os desfechos clínicos. Sua posição intermediária no mapa perceptual da MCA, indica que, além da agilidade no início do tratamento, fatores como o estadiamento clínico e a qualidade do tratamento recebido nas unidades desempenham um papel mais importante na determinação dos resultados. Esse achado reforça a necessidade de melhorias não apenas na rapidez do início do tratamento após o diagnóstico, mas também na qualidade do atendimento oncológico oferecido no contexto do SUS.

O desempenho do modelo *Random Forest*, embora moderado, está em consonância com estudos da literatura sobre a aplicação de algoritmos de aprendizado de máquina em dados de mundo real, especialmente no contexto da saúde pública (Bisong et al., 2024)(Shahid et al., 2023). Esses conjuntos de dados são frequentemente heterogêneos, desbalanceados e incompletos. Diante dessa realidade, a escolha por modelos baseados em árvores se justifica por sua robustez e capacidade de generalização em contextos complexos, mesmo em cenários onde os recursos e o acesso a dados de alta qualidade são limitados.

Por outro lado, embora o *XGBoost* tenha um desempenho competitivo, frequentemente exige ajustes mais detalhados de hiperparâmetros e maior poder computacional, o que pode ser desafiador em ambientes com infraestrutura limitada, como os sistemas públicos de saúde. Assim, o *Random Forest* se destaca como uma escolha apropriada, oferecendo um equilíbrio entre robustez, performance computacional e aplicabilidade prática, especialmente em contextos de saúde pública onde os recursos são restritos

**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.

Além do estadiamento clínico, fatores regionais, como a região de residência e o tipo de instituição de tratamento, foram associados a piores desfechos, destacando desigualdades no acesso e na qualidade do tratamento oncológico em diferentes partes do Brasil. O cumprimento da Lei dos 60 Dias demonstrou-se um fator importante, embora insuficiente para garantir a remissão completa dos pacientes. No contexto do SUS, estes achados reforçam a necessidade de políticas públicas que garantam não apenas o acesso rápido ao tratamento, mas também a equidade na distribuição de recursos de alta complexidade, além da implementação de estratégias de diagnóstico precoce.

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. 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.

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 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 Ramon 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. Aos meus amigos Welinton Hirai, Letícia Graziela e Patrícia Pedroso Estevam 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 ao meu crescimento.

**Referências**

Bisong, E.; Jibril, N.; Premnath, P.; Buligwa, E.; Oboh, G.; & Chukwuma, A. 2024. Predicting high blood pressure using machine learning models in low- and middle-income countries. *BMC Medical Informatics and Decision Making* 24(1): 1–21.

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

Harald Cramér 1946. *Mathematical Methods Of Statistics*. Princeton University Press.

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.

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.

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.

Shahid, M. U.; & Farooq, M. 2023. Boosted Random Forests for Predicting Treatment Failure of Chemotherapy Regimens. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13897 LNAI: 13–24.

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* .

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.

**Anexo 1 – Dicionário de banco de dados**



**Apêndice 2 – Tempo de predições dos modelos**

*Random Forest:* 2,84 segundos

*XGBoost:* 4,28 segundos