**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. 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). O desempenho dos modelos foi avaliado por métricas como acurácia, precisão, *recall* e F1-score. Valores de *SHAP* foram empregados para identificar a influência de cada variável no resultado da predição. O Random Forest apresentou uma acurácia média de 59,52%, prevendo corretamente o status de 17.040 pacientes, enquanto o XGBoost atingiu 58,99%. A análise de importância de variáveis e SHAP values revelaram que o estadiamento clínico foi a variável mais influente em ambos os modelos. O modelo Random Forest teve vantagem na redução de falsos positivos para a classe "Óbito" (especificidade de 98,24%), embora seu recall fosse menor (15,42%).

.

**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, and SHAP values were used to identify the contribution of each variable to the prediction outcome.

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

**Introdução**

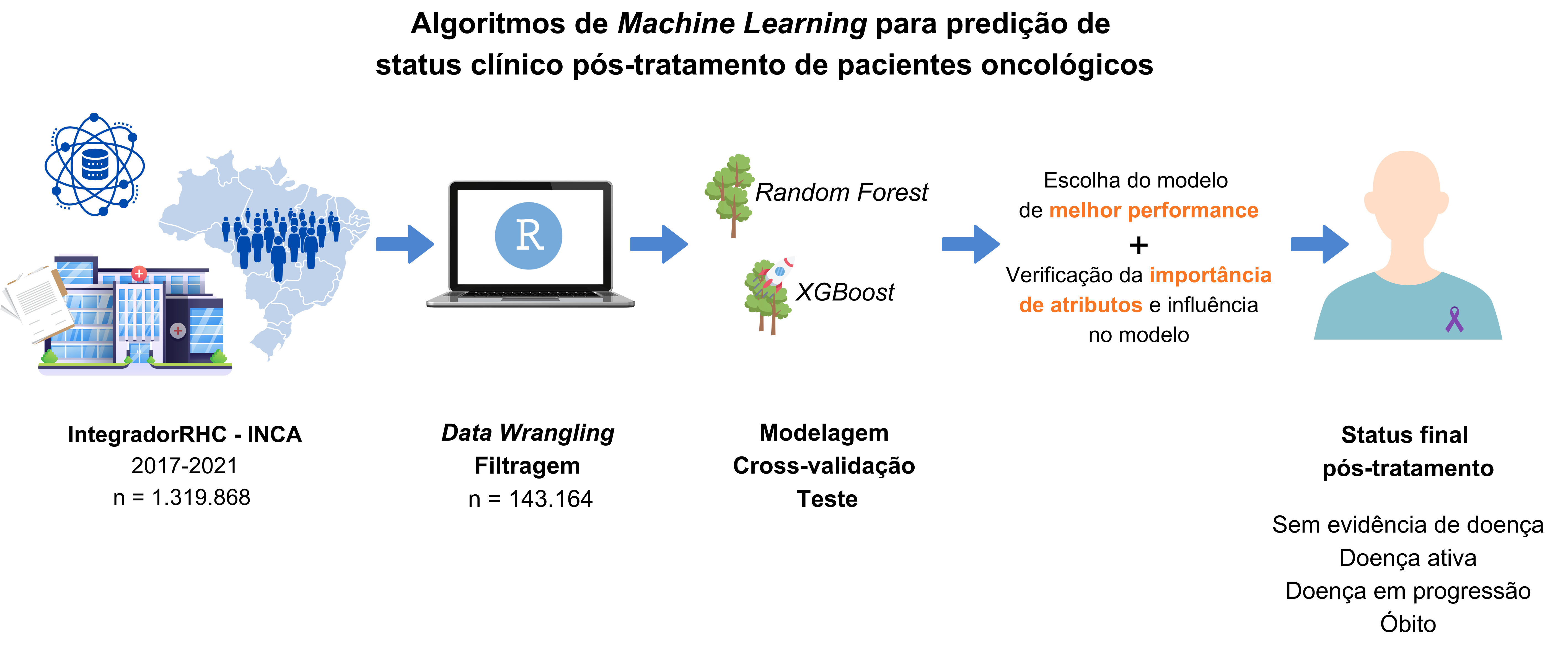
O câncer é uma doença multifatorial e uma das principais questões de saúde pública global, apresentando um aumento significativo nas taxas de incidência e mortalidade nas últimas décadas(INCA - Instituto Nacional de Câncer, [s.d.]; Sung et al., 2021). Esse cenário impõe desafios aos sistemas de saúde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

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 e diagnóstico precoce.

**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.

*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 (Anexo II).

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.

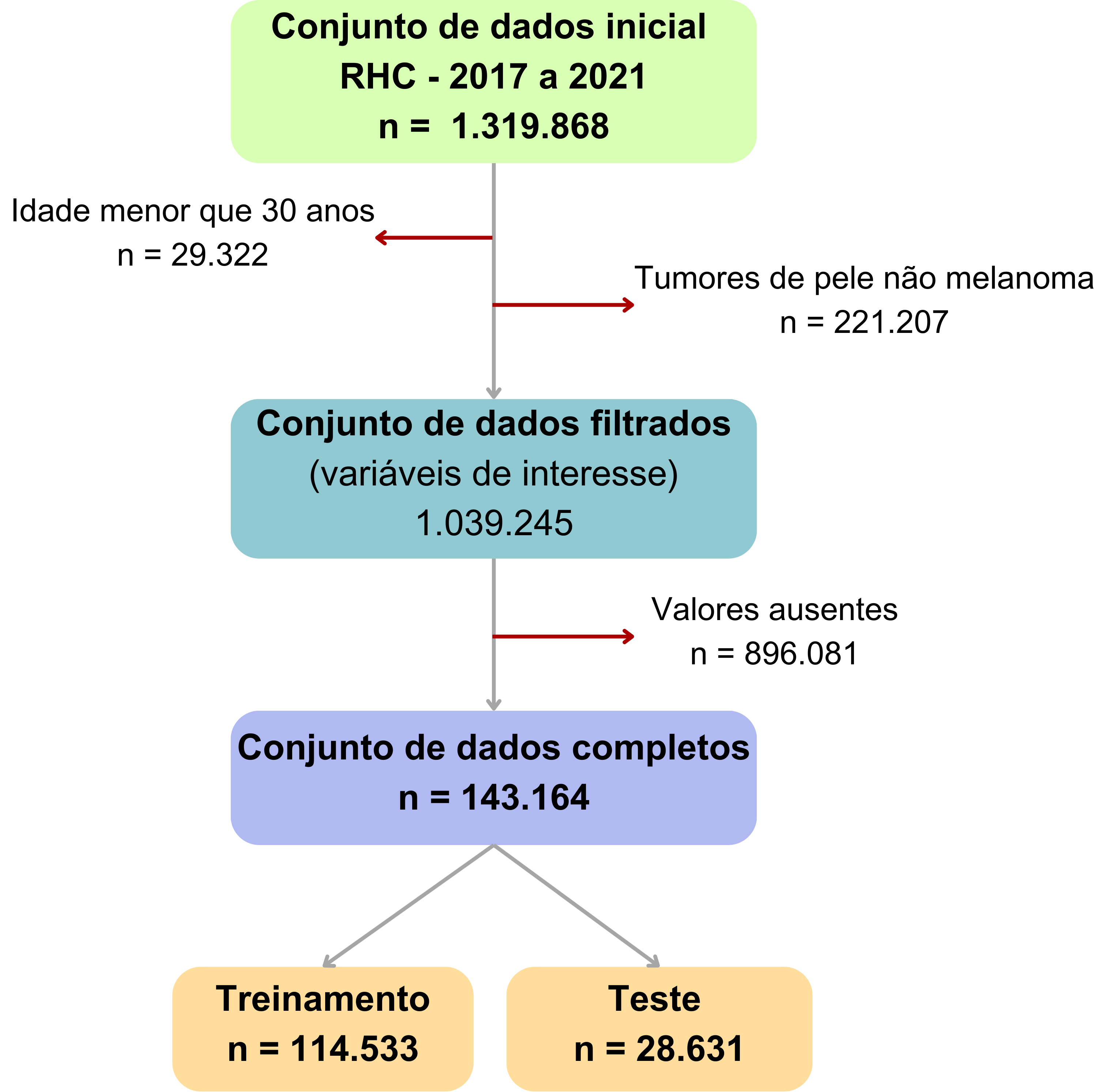
|  |  |  |
| --- | --- | --- |
| Variável | Tipo | Níveis |
| faixa\_etaria | Fator | "30-40 anos", "41-50 anos", "51-60 anos", "61-70 anos", "Mais de 70 anos" |
| sexo | Fator | "Masc", "Fem" |
| tabagismo | Fator | "Ex-tabagista", "Nunca tabagista", "Tabagista ativo" |
| etnia | Fator | "Amarelos e 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 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 |
| *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 “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 (2017-2021) acerca dos atributos selecionados se encontra apresentada na Tabela 4.

Neste 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. Mais detalhes estão descritos na Tabela 4.

**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 de doença/remissão completa | 42.028 (29%) |

N - número de casos.

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 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 ilustra a 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** |
| Faixa etária | 0,063 |
| Sexo | 0,117 |
| Tipo de caso | 0,877 |
| Etnia | 0,442 |
| Tabagismo | 0,866 |
| Historico familiar | 0,311 |
| Estado de residência | 0,215 |
| Respeito a Lei dos 60 dias | 0,117 |
| Tratamento inicial | 0,152 |
| Unidade de tratamento | 0,097 |
| Estadiamento clínico | 0,326 |
| Região | 0,097 |
| Localização do tumor | 0,180 |

*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.

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 ser possível de se verificar o grau de certeza das predições.

*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 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 ilustrados na Tabela 6.

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

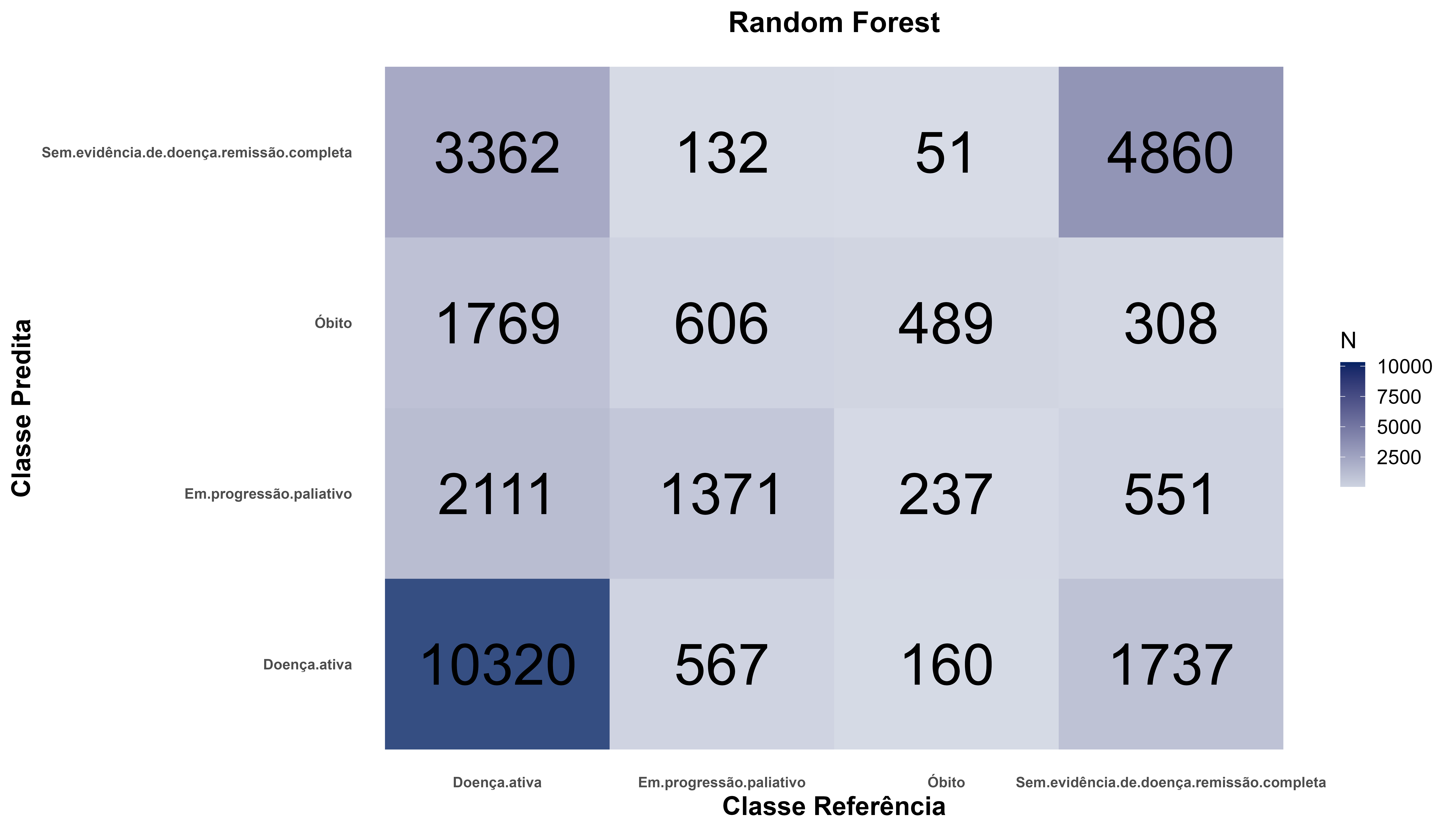
|  |  |
| --- | --- |
| Modelo | Acurácia média ± Desvio padrão |
| *Random Forest* | 59,72 ± 0,36 |
| *XGBoost* | 59,31 ± 0,22 |

*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 nossas classes de maior interesse em predição assertiva do status.

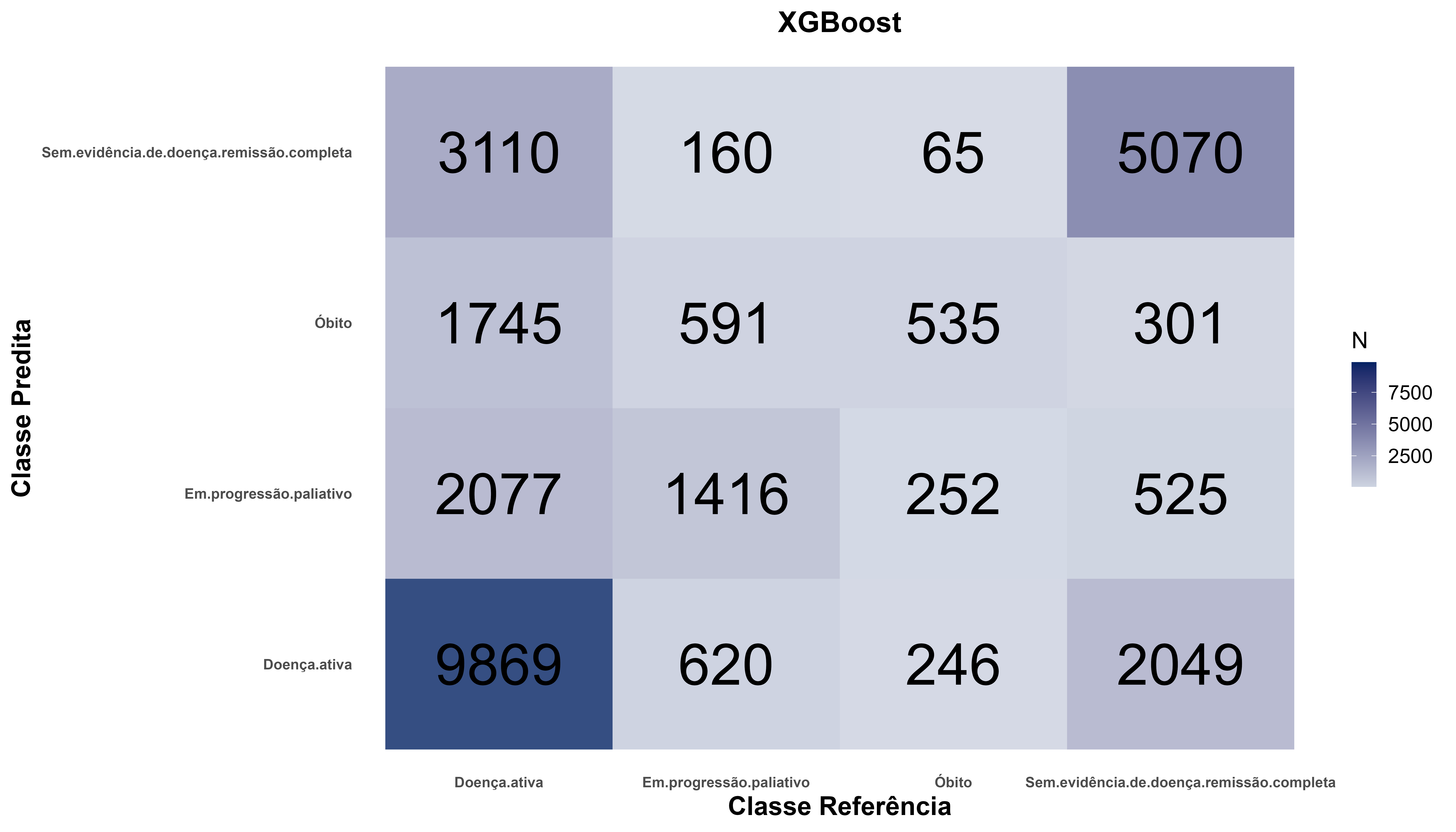
As métricas de desempenho incluem acurácia geral, precisão, *recall* (sensibilidade), especificidade e F1-score para as quatro classes da variável de interesse: “Doença ativa”, “Em progressão/paliativo”, “Óbito” e “Sem evidência de doença/remissão completa”.

O modelo *Random Forest* (Figura 5) 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 5.** Matriz de confusão da predição de desfechos clínicos no conjunto de teste para o modelo *Random Fores*t de melhor performance. As cores refletem o número de observações classificadas em cada classe, com o gradiente de cores.

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



**Figura 6.** 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 precisão 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, 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 foi relativamente baixo (15,42%), o que significa que muitos casos de óbito não foram identificados corretamente. Já o modelo *XGBoost* apresentou uma especificidade ligeiramente inferior para a classe “Óbito” (97,79%), mas seu recall foi ligeiramente superior (16,87%), sugerindo uma melhor capacidade de identificar óbitos verdadeiros.

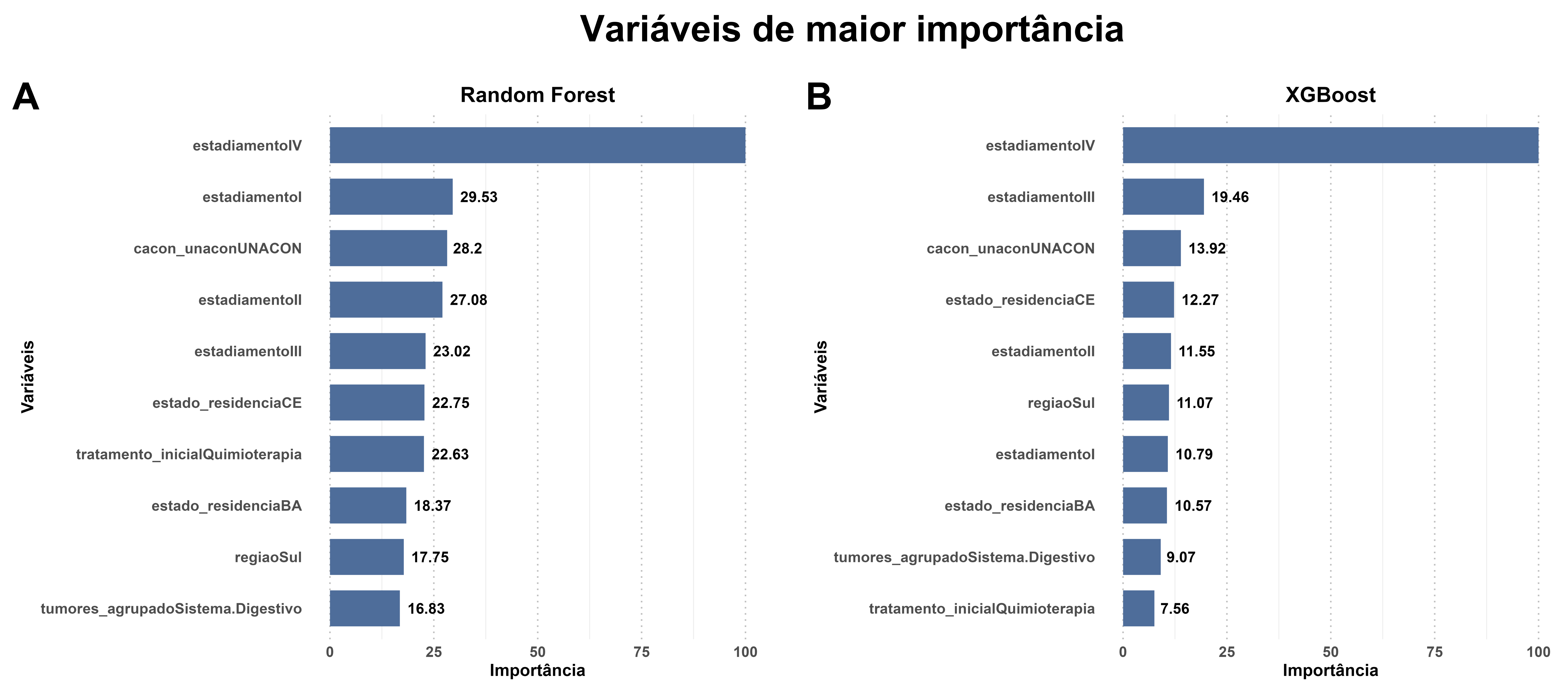
Para predição da classe “Em progressão/paliativo”, o modelo *Random Forest* apresentou uma precisão de 51,23% e um recall de 32,11%, resultando em um F1-score de 39,48%. Já o modelo *XGBoost*, apresentou uma precisão de 50,81% e um *recall* de 33,16%, refletindo um F1-score de 40,13%.

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, pode-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 6.** Gráfico de importância das variáveis para os modelos *Random Forest* (A) e *XGBoost* (B) com valores relativos.

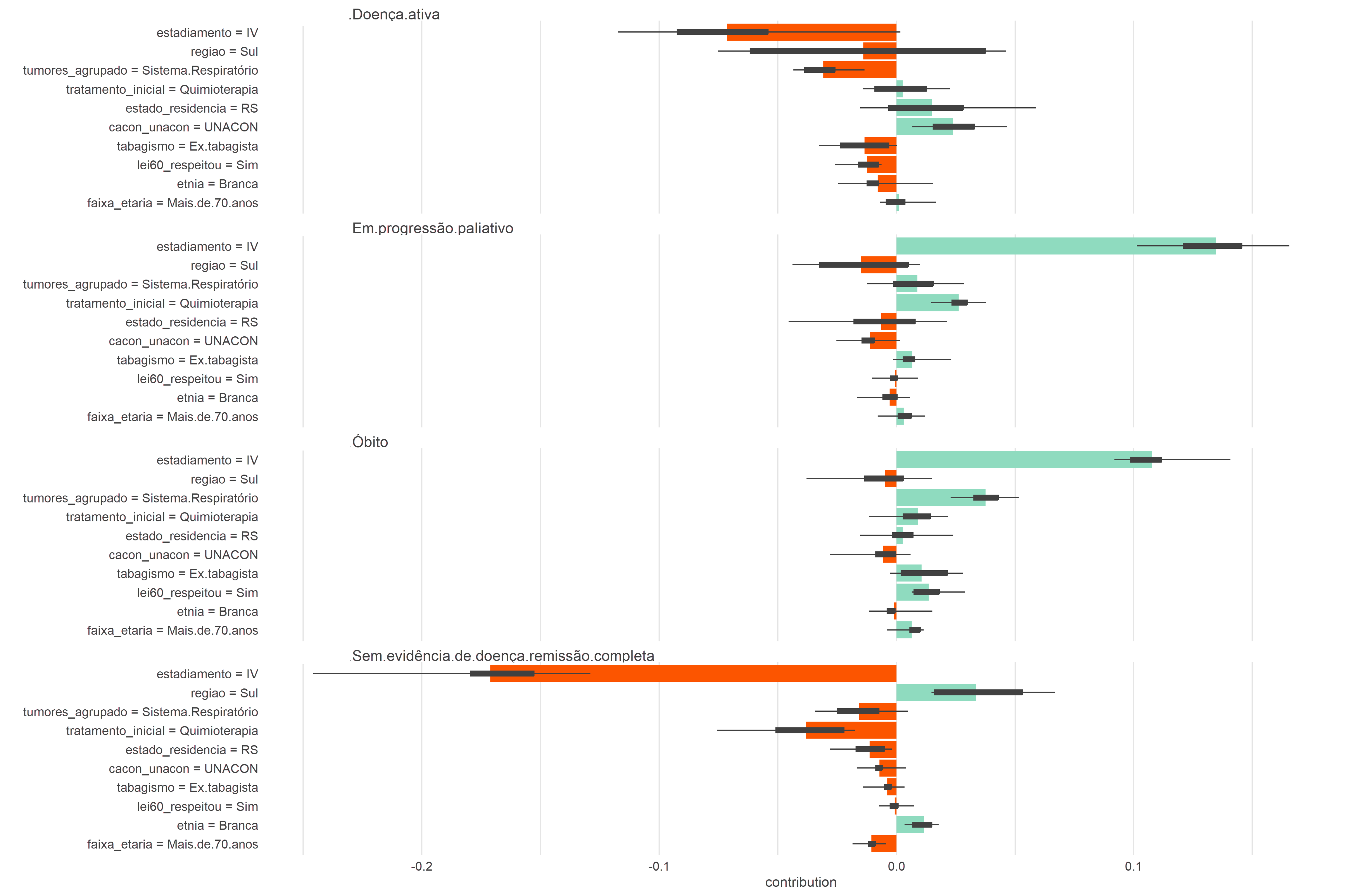
A análise de importância das variáveis revela que o modelo *Random Forest* apresentou um melhor aproveitamento das variáveis disponíveis, distribuindo de forma mais equilibrada a relevância entre um maior número de preditores e se propõe a explorar de maneira mais eficaz as variáveis presentes atribuindo-lhes importância significativa.

*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 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.

Focando na predição das classes com piores desfechos, observou-se que, para o modelo Random Forest, na classe "Em progressão/paliativo", o estadiamento IV exerce a maior influência, seguido pela região Sul e pelo tratamento inicial com quimioterapia. Esses resultados sugerem que essas variáveis estão associadas à progressão da doença ou à necessidade de cuidados paliativos. No que diz respeito à classe "Óbito", o estadiamento IV permanece como a variável mais significativa, contribuindo para um aumento expressivo na probabilidade de predição de óbito. Outras variáveis importantes incluem o estado de residência na região Sul e o tratamento em UNACON, que também estão associadas a uma maior probabilidade de mortalidade (Figura 8).

**Figura 8.** 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.

Estes achados reforçam a relevância do estadiamento clínico e de fatores regionais e de tratamento na determinação do prognóstico de pacientes oncológicos, evidenciando a capacidade do modelo *Random Forest* em identificar variáveis-chave que influenciam negativamente os desfechos clínicos

*Discussão*

Os resultados deste estudo reforçam o potencial dos algoritmos de machine learning na predição de desfechos clínicos de pacientes oncológicos, especialmente em contextos de saúde pública, como o SUS. A aplicação dos modelos *Random Forest* e *XGBoost* para prever o status clínico dos pacientes após tratamento ofereceu importantes *insights* sobre padrões entre variáveis clínicas e sociodemográficas que impactam diretamente os desfechos dos pacientes.

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 59,72%, comparada a 59,31% do *XGBoost*. Apesar de a diferença ser pequena em termos percentuais, 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" com uma especificidade de 98,24%, enquanto o *XGBoost* apresentou um *recall* ligeiramente superior. Isso sugere que, em cenários onde a detecção de casos críticos é crucial, o *Random Forest* pode ser uma escolha mais adequada, devido à sua maior precisão em evitar diagnósticos incorretos de óbito.

A análise da importância das variáveis, utilizando o índice de Gini, destacou o estadiamento clínico como o principal preditor de desfechos em ambos os modelos. Pacientes diagnosticados em estágios mais avançados (III e IV) apresentaram um prognóstico pior em comparação com aqueles diagnosticados em estágios iniciais, enfatizando a necessidade de diagnósticos precoces e a importância de estratégias de rastreamento e prevenção. Essas estratégias são cruciais para identificar pacientes em fases iniciais da doença, quando as chances de tratamento bem-sucedido são maiores.

No *Random Forest*, além do estadiamento clínico, variáveis como unidade de tratamento (UNACON ou CACON), tratamento inicial e estado/região de residência tiveram uma forte influência nos desfechos clínicos. Ainda, o modelo se mostrou mais eficiente em capturar e utilizar uma maior diversidade de variáveis, o que lhe confere uma maior flexibilidade e robustez para identificar padrões complexos e sutis nos dados. A distribuição mais homogênea da importância entre as variáveis observada no *Random Forest* pode garantir uma maior estabilidade do modelo, ao evitar que seu desempenho dependa de um menor número de variáveis. Em cenários clínicos, onde múltiplos fatores precisam ser considerados simultaneamente, essa abordagem é particularmente vantajosa.

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. Isso indica que, além da agilidade no início do tratamento, fatores como o estadiamento clínico e a qualidade do tratamento recebido 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, como no SUS(Bisong et al., 2024).(Shahid et al., 2023)

Acurácias moderadas, como as observadas neste estudo, são comuns em pesquisas que utilizam dados públicos ou prontuários médicos. Esses conjuntos de dados são frequentemente heterogêneos, desbalanceados e incompletos. Diante dessa realidade, a escolha por modelos baseados em árvores, como o Random Forest, se justifica por sua robustez e capacidade de generalizar bem em ambientes complexos, mesmo quando há limitações nos dados. Essa característica permite que o Random Forest lide eficazmente com inconsistências e forneça predições úteis, 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, ele 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 simplicidade, robustez 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.

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.

Na prática clínica, uma acurácia de 59,52% 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.

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 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 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.

**Apêndice ou Anexo**

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