

Predição do Cumprimento da Lei dos 60 Dias no Tratamento de Câncer de Mama: Comparação de Algoritmos de Aprendizado de Máquina

Lívia F. Ferrete ¹, Leticia M. Raposo ^{1,2}

¹Departamento de Métodos Quantitativos
Universidade Federal do Estado do Rio de Janeiro (UNIRIO) – Rio de Janeiro, RJ – Brasil

²Programa de Engenharia Biomédica
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brasil

livia.ferrete@edu.unirio.br, leticia.raposo@uniriotec.br

Abstract. *Breast cancer is the most incident malignant neoplasm among women in Brazil, and compliance with the 60-day legal deadline for treatment initiation within the public health system (SUS) remains an ongoing challenge. This study compares machine learning models (Decision Tree, Random Forest, XGBoost, and CatBoost) in predicting violations of this deadline using data from the Hospital Cancer Registry (2013–2024). CatBoost achieved the best predictive performance (AUC-ROC = 0.801; PR-AUC = 0.862). SHAP analysis identified the interval between diagnosis and first oncology appointment and the capital–interior care flow as the main predictors of delay. The findings demonstrate the potential of these models to support oncological care management.*

Resumo. *O câncer de mama é a neoplasia maligna mais incidente entre mulheres no Brasil, e o cumprimento do prazo legal de 60 dias para início do tratamento no SUS permanece um desafio. Este estudo compara modelos de aprendizado de máquina (Decision Tree, Random Forest, XGBoost e CatBoost) na predição da violação desse prazo, utilizando dados do Registro Hospitalar de Câncer (2013–2024). O CatBoost apresentou melhor desempenho preditivo (AUC-ROC = 0,801; PR-AUC = 0,862). A análise SHAP identificou o intervalo entre diagnóstico e primeira consulta oncológica e o fluxo assistencial capital–interior como principais preditores. Os resultados evidenciam o potencial desses modelos para subsidiar a gestão do cuidado oncológico.*

1. Introdução

O câncer de mama é a neoplasia maligna de maior incidência entre mulheres brasileiras, correspondendo a aproximadamente 30% dos novos casos no triênio 2026-2028, excluídos os tumores de pele não melanoma [Martins et al. 2026]. O diagnóstico precoce e o início oportuno do tratamento estão associados a melhores desfechos clínicos e à redução de custos para o sistema de saúde [Ginsburg et al. 2020].

No Brasil, a Lei Federal nº 12.732/2012 garante o início do tratamento oncológico no SUS em até 60 dias após o diagnóstico, posteriormente complementada pela Lei nº 13.896/2019. Apesar desse arcabouço legal, estudos nacionais apontam atrasos persistentes, sobretudo entre mulheres em situação de maior vulnerabilidade

socioeconômica e racial [Alves et al. 2022; Barros et al. 2019; Cabral et al. 2019; Medeiros et al. 2021].

Técnicas de aprendizado de máquina supervisionado têm sido utilizadas para identificar pacientes com maior risco de atraso no acesso ao tratamento. Entretanto, no contexto brasileiro, aplicações ainda se concentram em algoritmos clássicos, como no estudo de [Gioia et al. 2020], carecendo de avaliações com métodos mais robustos.

Diante disso, este estudo compara os algoritmos Decision Tree, Random Forest, XGBoost e CatBoost na predição da violação do prazo legal de 60 dias para o início do tratamento do câncer de mama no Brasil, com o objetivo de avaliar seu desempenho preditivo e identificar os principais fatores associados ao atraso.

2. Metodologia

Trata-se de um estudo observacional, retrospectivo, de base populacional, com abordagem preditiva baseada em técnicas de aprendizado de máquina supervisionado. Foram utilizados dados do Registro Hospitalar de Câncer (RHC) brasileiro, que reúne informações clínicas, demográficas e assistenciais de pacientes atendidos em unidades habilitadas em oncologia no Sistema Único de Saúde (SUS).

A população do estudo incluiu mulheres diagnosticadas com câncer de mama (CID-10: C50) entre 2013 e 2024, período posterior à promulgação da Lei Federal nº 12.732/2012. Foram considerados apenas casos analíticos, com primeiro tumor primário e datas válidas de diagnóstico e início do tratamento, assegurando que todas as pacientes incluídas tivessem a possibilidade de iniciar o tratamento dentro do prazo legal. Casos com tratamento prévio, óbito antes do tratamento, ausência de indicação terapêutica ou inconsistências temporais foram excluídos.

O desfecho de interesse foi definido como a violação do prazo de 60 dias entre a data de diagnóstico e o início do primeiro tratamento oncológico. A variável resposta foi operacionalizada de forma binária, classificando os casos como “dentro do prazo” (≤ 60 dias) ou “violação” (> 60 dias). As variáveis preditoras incluíram características demográficas, socioeconômicas, clínicas, tumorais, geográficas e de percurso assistencial, considerando apenas informações disponíveis até o momento do diagnóstico, de modo a evitar vazamento de informação. Variáveis relacionadas diretamente ao tratamento ou a eventos posteriores ao desfecho foram excluídas. Após análise de qualidade dos dados, variáveis com variância próxima de zero foram removidas, e apenas registros completos foram mantidos para a análise final.

Os dados foram particionados temporalmente para simular um cenário prospectivo, utilizando-se registros diagnosticados entre 2013 e 2019 para treinamento dos modelos ($n = 46.573$) e casos de 2020 a 2024 para teste ($n = 24.624$). Essa estratégia preserva a ordem cronológica dos dados e evita viés temporal.

Foram treinados e comparados quatro algoritmos de aprendizado de máquina: *Decision Tree*, *Random Forest*, *XGBoost* e *CatBoost*. A otimização de hiperparâmetros foi realizada por validação cruzada estratificada em três partições no conjunto de treinamento, utilizando a área sob a curva ROC (AUC-ROC) como métrica principal de seleção.

O desempenho dos modelos foi avaliado no conjunto de teste por meio de múltiplas métricas, incluindo AUC-ROC, área sob a curva de precisão-revocação (PR-

AUC), sensibilidade (revocação), especificidade, precisão, F1-score e acurácia. As diferenças de desempenho entre os modelos foram avaliadas por meio de *bootstrap* pareado, com geração de 2.000 reamostragens do conjunto de teste e ajuste para múltiplas comparações. O modelo com melhor desempenho global foi submetido à análise de interpretabilidade utilizando valores SHAP (*SHapley Additive exPlanations*), permitindo identificar os principais preditores associados à violação do prazo legal e a direção de seus efeitos sobre as predições.

Todas as análises foram conduzidas no ambiente R, versão 4.4.3, com controle de reprodutibilidade por meio da fixação de sementes aleatórias. O estudo utilizou exclusivamente dados secundários anonimizados e de acesso público, estando dispensado de apreciação por Comitê de Ética em Pesquisa, conforme a legislação vigente.

3. Resultados

Conforme a Tabela 1, os modelos *ensemble* e *boosting* superaram na maioria das métricas, com o *CatBoost* obtendo as maiores médias de AUC-ROC (0,801), AUC-PR (0,862), especificidade (0,867), precisão (0,849) e acurácia (0,715). A Figura 1 evidencia que diversas diferenças pareadas foram estatisticamente significativas, sobretudo nas comparações envolvendo *Decision Tree*.

Tabela 1. Médias das métricas de desempenho preditivo dos modelos de aprendizado de máquina no conjunto de teste, estimadas por *bootstrap* (2.000 reamostragens).

Medida	<i>Decision Tree</i>	<i>Random Forest</i>	<i>XGBoost</i>	<i>CatBoost</i>
ROC-AUC	0.769	0.785	0.797	0.801
PR-AUC	0.833	0.852	0.859	0.862
Sensibilidade	0.611	0.601	0.597	0.593
Especificidade	0.811	0.846	0.858	0.867
Precisão	0.803	0.831	0.841	0.849
F1-score	0.694	0.698	0.699	0.699
Acurácia	0.700	0.710	0.713	0.715

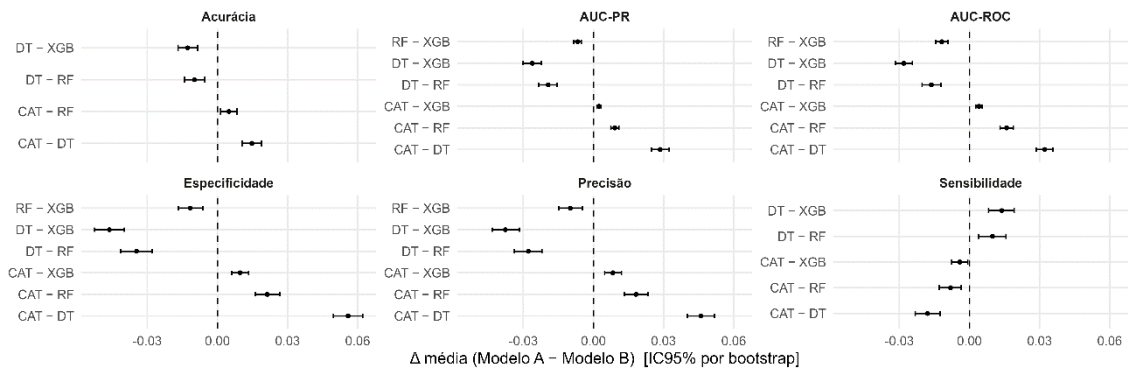


Figura 1. Diferenças entre médias *bootstrap* ($\Delta = \text{Modelo A} - \text{Modelo B}$) nas métricas de desempenho preditivo, com intervalos de confiança de 95% (IC95%), exibindo apenas comparações cujo IC95% não inclui zero. DT: *Decision Tree*; RF: *Random Forest*; XGB: *Extreme Gradient Boosting*; CAT: *CatBoost*.

A Figura 2 apresenta o gráfico *beeswarm* de valores SHAP obtido a partir do modelo CatBoost. O intervalo entre diagnóstico e primeira consulta oncológica constituiu a variável de maior importância global, com valores mais elevados associados a

contribuições SHAP positivas de maior magnitude para a predição de violação do prazo legal. O fluxo assistencial capital–interior foi o segundo preditor de maior impacto, com contribuições predominantemente positivas observadas entre pacientes atendidos sem deslocamento à capital e aqueles encaminhados do interior à capital. A região de residência apresentou contribuições positivas concentradas nas regiões Norte e Sudeste e negativas nas regiões Nordeste e Sul. A idade exibiu efeito positivo moderado, com valores mais elevados associados a maiores contribuições para a violação do prazo. Quanto à origem do encaminhamento, pacientes referenciados via SUS apresentaram contribuições predominantemente positivas, ao passo que encaminhamentos não SUS e por demanda espontânea tenderam a contribuições neutras ou negativas. Escolaridade e raça/cor exibiram menor importância global, com menores níveis de escolaridade e raça/cor negra associados a contribuições positivas para o desfecho.

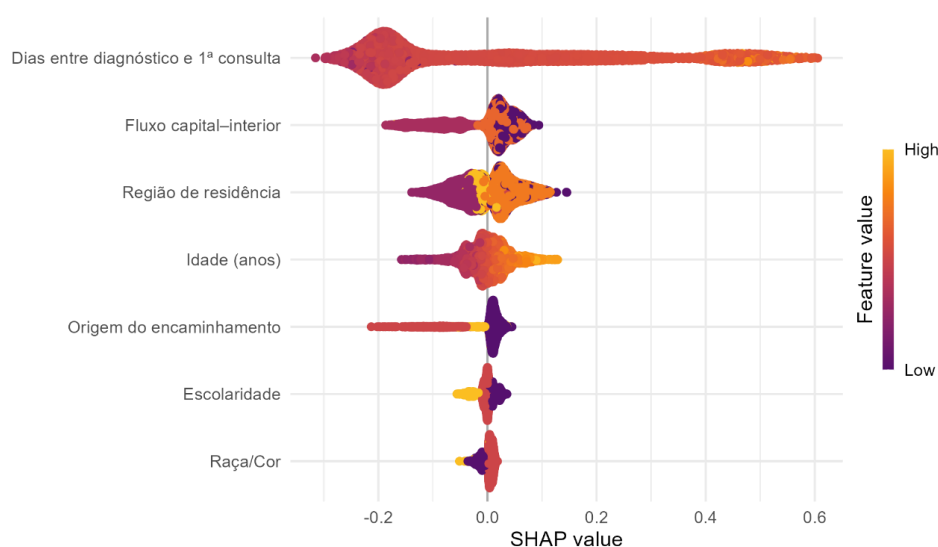


Figura 2. Gráfico *beeswarm* dos valores de SHAP do modelo *CatBoost* para as principais variáveis associadas à violação do prazo legal para início do tratamento. A posição horizontal representa o impacto do preditor na predição individual (valor de SHAP). As cores indicam o valor da variável: contínuo para variáveis numéricas e, para variáveis categóricas, os níveis na seguinte ordem (do menor índice ao maior, do roxo ao amarelo): Fluxo capital–interior (Sem deslocamento–capital; Sem deslocamento–interior; Interior→capital; Capital→interior), Região de residência (Norte; Nordeste; Centro-Oeste; Sudeste; Sul), Origem do encaminhamento (SUS; Não SUS; Conta própria), Escolaridade (Baixa; Média; Alta) e Raça/Cor (Branca; Negra; Outras).

4. Discussão

Os resultados evidenciaram que variáveis sociais, econômicas e regionais foram relevantes para a predição do cumprimento da Lei dos 60 dias em casos de câncer de mama na população feminina brasileira. A análise SHAP identificou raça/cor, escolaridade e origem do encaminhamento como preditores de destaque, corroborando achados prévios que associam mulheres negras, com baixa escolaridade e entrada pelo SUS a maiores intervalos entre diagnóstico e início do tratamento [Alves et al. 2022; Ferreira et al. 2020]. A idade, quarto preditor em ordem de importância, apresentou associação positiva com o atraso, embora a literatura registre tendências distintas, com maior demora observada tanto em pacientes mais jovens quanto em mais idosas

[Medeiros et al. 2021]. O intervalo entre diagnóstico e primeira consulta oncológica demonstrou comportamento consistente, com valores mais elevados associados a maiores contribuições para a violação do prazo. Pacientes residentes nas capitais e aqueles que necessitaram de deslocamento para atendimento hospitalar apresentaram maior tempo até o início do tratamento, em consonância com achados anteriores [Medeiros et al. 2021; Nogueira et al. 2023]. Quanto à região de residência, Nordeste e Sul exibiram contribuições negativas para a violação do prazo, sugerindo efeito protetor sobre o cumprimento da Lei, enquanto as demais regiões contribuíram para sua violação, conforme observado por [Nogueira et al. 2023]. Ressalta-se, contudo, que os valores SHAP representam contribuições para a predição do modelo e não efeitos causais, limitação inerente a estudos preditivos que deve ser considerada na interpretação dos resultados.

Os quatro algoritmos avaliados apresentaram desempenho satisfatório, com especificidade variando entre 0,811 (*Decision Tree*) e 0,867 (*CatBoost*). Considerando a AUC-ROC como métrica de referência para comparação entre classificadores binários [Li 2024], o *CatBoost* obteve o melhor desempenho (0,801), sendo sua superioridade atribuída à maior capacidade de lidar com variáveis categóricas e ao emprego do *boosting* ordenado, características metodologicamente adequadas aos dados provenientes do RHC. Em consonância com [McElfresh et al. 2024], algoritmos baseados em árvores com *gradient boosting*, como *CatBoost* e *XGBoost*, demonstraram desempenho superior em dados tabulares frente a redes neurais artificiais em diversos cenários.

Do ponto de vista metodológico, a comparação estatística dos desempenhos via *bootstrap* pareado permitiu avaliar os algoritmos sob os mesmos subconjuntos de dados, conferindo maior rigor à análise comparativa em relação à simples comparação isolada de métricas. A aplicação desses algoritmos representa uma ferramenta auxiliar com potencial de integração ao Programa de Navegação da Pessoa com Diagnóstico de Câncer, instituído no SUS pela Portaria GM/MS nº 6.592/2025, podendo contribuir para a redução do intervalo entre diagnóstico e início do tratamento oncológico nas redes pública e privada de saúde.

5. Conclusão

Este estudo demonstrou a relevância de variáveis sociodemográficas e regionais na predição do cumprimento da Lei dos 60 dias em mulheres com câncer de mama no Brasil. Os modelos baseados em *gradient boosted decision trees* apresentaram desempenho satisfatório, com o *CatBoost* obtendo a maior AUC-ROC entre os classificadores avaliados. A aplicação da análise SHAP conferiu interpretabilidade ao modelo, identificando o intervalo entre diagnóstico e primeira consulta oncológica e o fluxo assistencial capital–interior como os principais preditores do atraso no início do tratamento. Ambas as variáveis foram derivadas dos dados brutos do RHC por meio de *feature engineering*, evidenciando que a construção de novos atributos representa uma estratégia promissora para estudos futuros nesta temática. Os resultados reforçam a vulnerabilidade sistêmica de grupos historicamente desfavorecidos, como mulheres negras, de baixa escolaridade e usuárias exclusivas do SUS, sem que se estabeleça, contudo, relação de causalidade direta. A comparação de desempenho via *bootstrap* pareado constitui contribuição metodológica adicional relevante. Espera-se que os modelos aqui desenvolvidos possam subsidiar gestores e iniciativas do SUS na

identificação precoce de perfis com maior risco de violação do prazo legal, contribuindo para a aceleração do início do tratamento oncológico no país.

6. Referências

- Alves, M., Monteiro, M., Alves, F. and Figueiredo, F. (2022). Determinants of Lack of Access to Treatment for Women Diagnosed with Breast Cancer in Brazil. *International Journal of Environmental Research and Public Health*, 19(13):7635.
- Barros, A., Araújo, J., Murta-Nascimento, C. and Dias, A. (2019). Clinical pathways of breast cancer patients treated in the Federal District, Brazil. *Revista de Saúde Pública*, 53:14.
- Cabral, A., Giatti, L., Casale, C. and Cherchiglia, M. (2019). Social vulnerability and breast cancer: differentials in the interval between diagnosis and treatment of women with different sociodemographic profiles. *Ciência & Saúde Coletiva*, 24(2):613-622.
- Ferreira, N., Schoueri, J., Sorpreso, I., Adami, F. and Figueiredo, F. (2020). Waiting Time between Breast Cancer Diagnosis and Treatment in Brazilian Women: An Analysis of Cases from 1998 to 2012. *International Journal of Environmental Research and Public Health*, 17(11):4030.
- Ginsburg, O., Yip, C., Brooks, A., Cabanes, A., Caleffi, M., Yataco, J., Gyawali, B., McCormack, V., Anderson, M., Mehrota, R. et al. (2020). Breast Cancer Early Detection: A Phased Approach to Implementation. *Cancer*, 126(10):2379-2393.
- Gioia, S., Galdino, R., Brigagão, L., Valadares, A., Secol, F., Miguel, S., Bukowski, A., Krush, L. and Goss, P. (2020). Prediction of Attendance to the "Law of 60 Days" in Breast Cancer Patients using Machine Learning Classifiers. *Acta Scientific Cancer Biology*, 4(3):16-28.
- Li, J. (2024). Area under the ROC Curve has the most consistent evaluation for binary classification. *PLoS ONE*, 19(12):e0316019.
- Martins, L., Chaves, G., Oliveira, J., Souza, L., Neto, P., Carvalho, F., Vasconcelos, G., Dias, M. and Mello, M. (2026). Epidemiological Profile of Cancer Incidence in Brazil and Regions: Estimates for the 2026-2028 Triennium. *Revista Brasileira de Cancerologia*, 72(2):e-025587.
- McElfresh, D., Khandagale, S., Valverde, J., Prasad, V., Ramakrishnan, G., Goldblum, M. and White, C. When Do Neural Nets Outperform Boosted Trees on Tabular Data? (2024). In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*. NeurIPS Foundation.
- Medeiros, G., Thuler, L. and Bergmann, A. (2021). Determinants of delay from cancer diagnosis to treatment initiation in a cohort of brazilian women with breast cancer. *Health and Social Care in the Community*, 00:1-10.
- Nogueira, M., Atty, A., Tomazelli, J., Jardim, B., Bustamante-Teixeira, M. and Azevedo e Silva, G. Frequency and factors associated with delay in breast cancer treatment in Brazil, according to data from the Oncology Panel, 2019-2020. *Epidemiologia e Serviços de Saúde*, 32(1):e2022563.