*Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1*
ISSN: 3085-8461 • doi: 10.5753/reic.2026.6648


**Artigo de Pesquisa/Research Paper**

# Análise Comparativa de Algoritmos de Aprendizado de Máquina para Predição de Recidiva de Câncer de Próstata com Dados da Fundação Oncocentro de São Paulo

## *Comparative Analysis of Machine Learning Algorithms for Predicting Prostate Cancer Recurrence Using Data from the Oncocentro Foundation of São Paulo*

**Giovani Reis** [Centro de Tecnologia da Informação Renato Archer (CTI) / Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) | *giovani.r@aluno.ifsp.edu.br* ]
**Guilherme Ruppert** [Centro de Tecnologia da Informação Renato Archer (CTI) | *guilherme.ruppert@cti.gov.br* ]

*Divisão de Metodologias da Computação – DIMEC, CTI - Centro de Tecnologia da Informação Renato Archer, Rodovia Dom Pedro I (SP-65), Km 143,6 – Chácaras Campos dos Amarais, Campinas-SP, 13069-901, Brasil*

**Resumo.** O câncer de próstata é a neoplasia mais comum entre os homens no Brasil, tornando o estudo dos casos de recidiva um tema de grande interesse para a medicina. O foco principal deste trabalho está na aplicação do aprendizado de máquina supervisionado, por meio de uma análise comparativa dos algoritmos de classificação — Random Forest, XGBoost, HistGradientBoosting e Naive Bayes — para predizer a ocorrência da recidiva. Busca-se principalmente avaliar o desempenho desses modelos na predição do atributo de recidiva com base nos demais atributos presentes na base de dados da Fundação Oncocentro de São Paulo. Espera-se contribuir para o estudo da eficácia de distintas abordagens de aprendizado de máquina em dados clínicos abertos e desbalanceados sobre o câncer de próstata. Os resultados obtidos são significativos e apontam para a superioridade de modelos mais simples neste cenário, além de indicarem a possibilidade de continuidade dos estudos com técnicas mais avançadas.

**Abstract.** Prostate cancer is the most common neoplasm among men in Brazil, making the study of recurrence cases a topic of great interest to medicine. The main focus of this work is the application of supervised machine learning, through a comparative analysis of classification algorithms — Random Forest, XGBoost, HistGradientBoosting, and Naive Bayes — to predict the occurrence of recurrence. The main goal is to evaluate the performance of these models in predicting the recurrence feature based on the other characteristics present in the dataset from the Fundação Oncocentro de São Paulo. It is hoped to contribute to the study of the effectiveness of different machine learning approaches on open and imbalanced clinical data regarding prostate cancer. The results obtained are significant and point to the superiority of simpler models in this scenario, as well as suggesting the possibility of continuing the studies with more advanced techniques.

**Palavras-chave:** Câncer de Próstata, Aprendizado de Máquina, Predição de Recidiva, Dados Desbalanceados, Classificação Supervisionada.

**Keywords:** Prostate Cancer, Machine Learning, Recurrence Prediction, Imbalanced Data, Supervised Classification.

**Recebido/Received:** 07 October 2025 • **Aceito/Accepted:** 27 February 2026 • **Publicado/Published:** 02 April 2026

## 1 Introdução

Segundo o Instituto Nacional de Câncer INCA [2025], o câncer de próstata (CaP) é o tipo de câncer mais comum entre homens no Brasil, com mais de 71.730 casos estimados para o ano de 2023. A recidiva — retorno da doença após a remissão — representa um desafio clínico significativo e está associada a uma complexa interação de fatores, tais como: estágio avançado no diagnóstico, resposta incompleta ao tratamento primário, níveis elevados de Antígeno Prostático Específico (PSA) e perfis de mutações genéticas específicos. A capacidade de prever quais pacientes apresentam maior risco de recidiva é fundamental para a personalização do tratamento e para o aumento da sobrevida Fonseca *et al*. [2007].

Tradicionalmente, a análise de sobrevivência e o prognóstico oncológico baseiam-se em métodos estatísticos clássicos, como o modelo de Riscos Proporcionais de Cox Zardeto *et al*. [2022]. No entanto, o avanço da inteligência artificial tem impulsionado a aplicação de técnicas de aprendizado de máquina (*Machine Learning*), que demonstraram capacidade superior em capturar padrões não lineares e interações complexas entre variáveis clínicas, superando as limitações de linearidade inerentes aos modelos estatísticos convencionais Kourou *et al*. [2015]; Uddin *et al*. [2019]. Entretanto, a transposição desses algoritmos para a análise de sobrevivência não é trivial. A literatura aponta que a utilização de métodos de classificação padrão enfrenta barreiras significativas, sobretudo devido às dificuldades no tratamento de dados censurados (quando o desfecho não é observado durante o estudo) e à presença de dados ausentes, comuns em registos hospitalares do mundo real Wang *et al*. [2019]. A incapacidade de lidar adequadamente com a censura pode levar a vieses de predição, subestimando o risco em pacientes com seguimento incompleto.

É neste cenário de desafio metodológico e prático que se insere a presente investigação. Embora existam estudos internacionais sobre o tema, há uma lacuna na validação comparativa desses algoritmos em bases de dados públicas brasileiras,

que frequentemente apresentam ruído, desbalanceamento e incompletude. O problema de pesquisa central deste trabalho reside, portanto, em verificar se algoritmos de aprendizado de máquina conseguem superar essas limitações dos dados regionais e oferecer predições robustas. Para tal, este trabalho propõe uma análise comparativa de quatro algoritmos — *Random Forest*, XGBoost, *HistGradientBoosting* e *Naive Bayes* — para a predição da recidiva do câncer de próstata, utilizando dados abertos e reais da Fundação Oncocentro de São Paulo (FOSP).

# 2 Trabalhos Relacionados

A aplicação de aprendizado de máquina no prognóstico oncológico tem crescido exponencialmente. Estudos recentes, como o de Lee *et al*. [2020], demonstraram que modelos de ensemble tendem a superar modelos lineares na predição de falha bioquímica após prostatectomia.

No contexto brasileiro, Antunes *et al*. [2025] exploraram biomarcadores moleculares combinados com *Machine Learning*, obtendo alta acurácia, porém utilizando conjuntos de dados proprietários e pequenos. Maeda *et al*. [2022] realizaram estudos preliminares com dados da FOSP, destacando a dificuldade de lidar com a alta dimensionalidade e dados faltantes.

A lacuna que este trabalho busca preencher é a avaliação de algoritmos modernos de *Gradient Boosting (XGBoost e HistGradientBoosting)* em comparação com métodos clássicos (*Naive Bayes*) e *Bagging (Random Forest)* em um cenário de dados abertos massivos e altamente desbalanceados ($\sim$10% de taxa de recidiva), estabelecendo baselines claros para estudos futuros.

# 3 Conceitos, Técnicas e Algoritmos

## 3.1 Aprendizado de Máquina

O aprendizado de máquina (*Machine Learning* – ML) é uma subárea da inteligência artificial cujo objetivo é desenvolver algoritmos e sistemas capazes de aprender a partir de dados e adquirir conhecimento de forma automática. Essencialmente, seu objetivo é permitir que computadores reconheçam padrões em diferentes tipos de informação — como imagens, sons e dados estruturados — e os utilizem para tomar decisões sem intervenção humana direta. Para tanto, os modelos precisam ser treinados com exemplos rotulados ou não, ajustando seus parâmetros internos de maneira a garantir a capacidade de generalização para novos dados Bigonha [2025].

## 3.2 Tipos de Aprendizado de Máquina e Técnica Utilizada

O aprendizado de máquina pode ser classificado em quatro categorias principais:

- **Supervisionado:** Nesse tipo de classificador, os dados usados para treinar o modelo são rotulados, ou seja, contêm exemplos das entradas (chamadas de atributos) e das saídas corretas (rótulos ou classes). Os algoritmos analisam um grande conjunto de dados com esses pares de treinamento para inferir qual seria o valor de saída desejado quando solicitado a fazer uma previsão com base em dados novos.
- **Não supervisionado:** trabalha com dados não rotulados, buscando identificar estruturas ocultas, como padrões, associações ou agrupamentos.
- **Semi-supervisionado:** combina um pequeno conjunto de dados rotulados com um grande volume de dados não rotulados, aproveitando ambos para melhorar o desempenho preditivo.
- **Por reforço:** baseia-se na interação de um agente com o ambiente, em que o aprendizado ocorre por tentativa e erro, recebendo recompensas ou penalidades de acordo com as ações tomadas.

Neste trabalho foi adotado o aprendizado supervisionado, por se tratar de um problema de classificação binária, em que a variável-alvo representa a ocorrência ou não da recidiva do câncer de próstata.

## 3.3 Algoritmos Selecionados

Quatro algoritmos supervisionados foram escolhidos para compor a análise comparativa, contemplando diferentes paradigmas de aprendizado:

- ***Random Forest*:** O algoritmo *Random Forest* (RF), ou Floresta Aleatória, é um meta-estimador do tipo ensemble que opera através da construção de múltiplas árvores de decisão durante o treinamento. Sua arquitetura se fundamenta em duas fontes principais de aleatoriedade: a técnica de *bagging* (*Bootstrap Aggregating*), onde cada árvore é treinada sobre uma subamostra aleatória dos dados (com reposição), e a seleção aleatória de um subconjunto de características (*features*) em cada nó para determinar a melhor divisão. Essa dupla aleatoriedade descorrelaciona as árvores individuais, resultando em um modelo com menor variância e alta robustez contra o *overfitting*. A predição final é agregada a partir de todas as árvores, utilizando o voto majoritário para tarefas de classificação e a média para regressão. Sua capacidade de lidar nativamente com variáveis numéricas e categóricas o torna uma ferramenta poderosa para aplicações complexas, como em análises de dados médicos Shalev-Shwartz and Ben-David [2014].
- **XGBoost:** O *eXtreme Gradient Boosting (XGBoost)* é uma implementação avançada e altamente otimizada do framework de *Gradient Boosting*. Utilizando o método de ensemble do tipo boosting, ele constrói árvores de decisão de forma sequencial e aditiva, onde cada nova árvore é treinada para corrigir os erros do modelo anterior. Mais especificamente, o algoritmo utiliza a descida de gradiente (gradient descent) para minimizar uma função de perda, ajustando cada nova árvore aos gradientes negativos (pseudo-resíduos) da iteração anterior. O diferencial do XGBoost reside em suas otimizações de sistema e avanços algorítmicos, que incluem a regularização da complexidade da árvore (termos L1 e L2 para evitar *overfitting*), o processamento paralelo durante a construção das árvores e o tratamento eficiente de dados faltantes. Essas características resultam em uma performance superior em velocidade e precisão, tornando-o uma ferramenta de referência para problemas de classificação e regressão, especialmente em cenários com dados estruturados e de grande volume Shalev-Shwartz

and Ben-David [2014].

- **HistGradientBoosting:** O *Histogram-based Gradient Boosting* (HistGBM) é uma evolução do algoritmo clássico de *Gradient Boosting* projetada para otimizar drasticamente a eficiência computacional e a escalabilidade, especialmente em grandes volumes de dados.Sua inovação central reside na discretização das variáveis contínuas em um número fixo de "bins"(ou caixas) antes do treinamento, criando histogramas das características. Durante a construção das árvores, em vez de avaliar cada valor único para encontrar o ponto de divisão ótimo — um processo computacionalmente caro —, o algoritmo itera apenas sobre os bins. Essa abordagem reduz significativamente o consumo de memória e acelera o treinamento, ao mesmo tempo que a discretização atua como uma forma de regularização, melhorando a capacidade de generalização do modelo. Adicionalmente, o HistGBM possui suporte nativo para valores ausentes, aprendendo durante o treino a direção ótima (esquerda ou direita) para alocar amostras com dados faltantes com base no ganho de informação, o que simplifica o pré-processamento dos dados Shalev-Shwartz and Ben-David [2014].
- ***Naive Bayes*:** O classificador *Naive Bayes* é um algoritmo de aprendizado supervisionado de natureza probabilística, fundamentado na aplicação do Teorema de Bayes. Seu princípio operacional consiste em calcular a probabilidade posterior de uma amostra pertencer a uma determinada classe, com base nas probabilidades observadas dos seus atributos. O adjetivo "ingênuo"(*naive*) advém da sua premissa central e simplificadora: a assunção de independência condicional entre todos os atributos (*features*), dado o valor da classe. Embora essa premissa seja raramente satisfeita em problemas reais, ela reduz drasticamente a complexidade computacional do modelo, permitindo um treinamento extremamente rápido. Surpreendentemente, essa simplicidade não impede que o *Naive Bayes* atinja alta performance, especialmente em tarefas de classificação de texto (como filtragem de spam) e outros problemas com alta dimensionalidade, onde sua eficiência e robustez o consolidam como um *baseline* poderoso Bonaccorso [2017].

# 4 Metodologia

A metodologia deste estudo foi dividida em uma sequência de etapas, começando com o preparo dos dados brutos e terminando com a avaliação dos modelos de aprendizado de máquina. O processo é detalhado a seguir.

## 4.1 Fonte e Seleção de Dados

O presente estudo utilizou um conjunto de dados público, anonimizado e de acesso aberto, disponibilizado pela Fundação Oncocentro de São Paulo (FOSP). A base completa da FOSP representa um dos maiores repositórios de dados oncológicos da América Latina, contendo um histórico superior a **1 milhão de registros tumorais** de diversas tipologias.

Os dados originais encontravam-se no formato **DBF** (dBase File), um padrão de arquivo de banco de dados estruturado amplamente utilizado em sistemas legados de gestão de saúde pública. Este formato organiza os dados em registros de tamanho fixo com cabeçalhos descritivos, mas requer a implementação de um pipeline de extração e conversão para formatos analíticos modernos (como CSV) para a manipulação eficiente em ambientes de Ciência de Dados, processo realizado nesta etapa.

A partir desse universo massivo, foi aplicada a filtragem rigorosa para selecionar apenas os registros referentes a pacientes diagnosticados com câncer de próstata. Para isso, adotou-se a Classificação Internacional de Doenças para Oncologia, 3ª edição (CID-O-3), selecionando-se exclusivamente os casos associados ao código topográfico C61.9 (TOPO = C619). Este recorte resultou em um *dataset* final de **131.437 registros**, volume que garante alta significância estatística para o treinamento dos modelos.

Destes, a classe minoritária (Recidiva) representava 10,27% (13.498 casos), enquanto a classe majoritária (Sem Recidiva) compunha 89,73% (117.939 casos). A variável-alvo utilizada na modelagem foi `RECNENHUM`, um indicador binário onde 0 representa a ocorrência de recidiva e 1 a ausência.

## 4.2 Pré-processamento e Dicionário de Dados

A preparação dos dados foi fundamental para garantir a qualidade da modelagem. Foram selecionadas **20 variáveis preditoras** (*features*) baseadas na relevância clínica descrita na literatura e disponibilidade na base. O Quadro 1 apresenta o dicionário de dados, descrevendo cada variável e o tratamento aplicado.

*Nota: Valores ausentes em variáveis numéricas foram preenchidos com a mediana e variáveis categóricas foram transformadas utilizando Label Encoding.*

## 4.3 Divisão de Dados

O conjunto de dados processado foi particionado em dois subconjuntos mutuamente exclusivos: treino (80%) e teste (20%). A divisão foi realizada por meio de amostragem aleatória estratificada em relação à variável-alvo. Esta técnica garantiu que a proporção original de casos de recidiva fosse mantida em ambas as amostras, evitando vieses na avaliação.

- **Conjunto de Treino:** 105.149 amostras.
- **Conjunto de Teste:** 26.288 amostras.

## 4.4 Balanceamento de Classes (SMOTE)

Para corrigir o desbalanceamento severo no conjunto de treino, optou-se pela aplicação da técnica **SMOTE (Synthetic Minority Over-sampling Technique)**.

A principal vantagem do uso do SMOTE em comparação a técnicas simples de *undersampling* (remoção de dados da classe majoritária) é a **preservação integral da informação**. Em datasets médicos, onde cada registro pode conter padrões clínicos raros e valiosos, descartar dados da classe majoritária poderia levar à perda de capacidade de generalização. Por outro lado, comparado ao *random oversampling* (simples duplicação), o SMOTE reduz o risco de *overfitting*, pois cria exemplos sintéticos novos através da interpolação de vizinhos próximos no espaço de características.

Após a aplicação do SMOTE, as classes no treino foram equalizadas, resultando em 94.351 registros para cada classe.

**Tabela 1.** Dicionário de Variáveis do Conjunto de Dados (RHC/FOSP)

| Variável | Tipo | Descrição |
|---|---|---|
| IDADE | Numérico | Idade do paciente no momento do diagnóstico |
| PSA | Numérico | Nível de Antígeno Prostático Específico (ng/mL) |
| GLEASON | Numérico | Escore de Gleason (grau de agressividade do tumor) |
| TOPO | Categórico | Topografia do tumor (Código CID-O) |
| MORFO | Categórico | Morfologia do tumor (tipo celular) |
| EC | Categórico | Estadiamento clínico agrupado (I, II, III, IV) |
| T, N, M | Categórico | Classificação TNM (Tumor, Nódulo, Metástase) |
| G | Categórico | Grau de diferenciação histológica |
| LATERALI | Categórico | Lateralidade do tumor primário |
| TRATAMENTO | Categórico | Tipo de tratamento principal realizado |
| CIRURGIA | Binário | Indicador de realização de cirurgia |
| QUIMIO | Binário | Indicador de realização de quimioterapia |
| RADIO | Binário | Indicador de realização de radioterapia |
| HORMONIO | Binário | Indicador de realização de hormonioterapia |
| CONSDIAG | Numérico | Tempo (dias) entre a primeira consulta e o diagnóstico |
| DIAGTRAT | Numérico | Tempo (dias) entre o diagnóstico e o início do tratamento |
| HABILIT | Categórico | Código de habilitação da unidade hospitalar |
| INSTITU | Categórico | Identificador anônimo da instituição de saúde |

O conjunto de teste permaneceu com sua distribuição original para garantir uma avaliação realista.

## 4.5 Construção e Treinamento dos Modelos

Para a tarefa de classificação binária, foram empregados quatro algoritmos supervisionados: **Gaussian Naive Bayes**, **Random Forest Classifier**, **XGBoost** (eXtreme Gradient Boosting) e **HistGradientBoosting**.

Antes do treinamento, foi aplicada a normalização dos dados com *StandardScaler*, ajustada apenas aos dados de treino para evitar vazamento de dados (*data leakage*).

## 4.6 Avaliação de Desempenho

A performance dos modelos foi avaliada no conjunto de teste independente, utilizando métricas robustas para dados desbalanceados:

- **Recall (Sensibilidade):** Capacidade de detectar os casos reais de recidiva.
- **Precision:** Proporção de acertos entre as predições de recidiva.
- **F1-Score:** Média harmônica entre Precision e Recall.
- **AUC-ROC:** Área sob a Curva da Característica de Operação do Receptor, que mede a capacidade de discriminação global do modelo.

Para validar a eficácia do aprendizado, os resultados foram comparados com modelos de referência aleatória (*Dummy Classifiers*) utilizando estratégias estratificadas e de frequência majoritária.

# 5 Resultados

A análise comparativa dos quatro algoritmos de aprendizado de máquina foi conduzida sobre um conjunto de teste com 26.288 amostras, das quais 2.700 (10.27%) correspondiam a casos de recidiva (classe 0). Os resultados quantitativos e qualitativos são detalhados a seguir.

## 5.1 Desempenho Comparativo dos Modelos

Ao contrário de abordagens que ignoram o desbalanceamento, a aplicação do SMOTE combinada com modelos de *ensemble* resultou em uma capacidade discriminatória robusta. A Tabela 2 apresenta o comparativo das métricas obtidas. Observa-se que todos os modelos de aprendizado de máquina superaram significativamente os *baselines* aleatórios (Dummy Classifiers) na métrica AUC-ROC, comprovando a eficácia do treinamento.

**Tabela 2.** Comparação de Métricas de Desempenho (Foco na Classe Recidiva)

| Modelo | AUC-ROC | F1-Score | Recall | Precision | AP |
|---|---|---|---|---|---|
| **Random Forest** | **0.783** | **0.367** | 0.445 | 0.312 | **0.312** |
| HistGradientBoosting | 0.781 | 0.347 | 0.391 | 0.311 | 0.289 |
| XGBoost | 0.776 | 0.083 | 0.045 | **0.498** | 0.286 |
| Naive Bayes | 0.582 | 0.204 | **0.655** | 0.121 | 0.158 |
| *Dummy (Stratified)* | 0.499 | 0.170 | 0.500 | 0.102 | 0.103 |
| *Dummy (Most Freq)* | 0.500 | 0.000 | 0.000 | 0.000 | 0.103 |

O modelo **Random Forest** emergiu como a abordagem mais equilibrada, apresentando a maior AUC-ROC (0.783) e o melhor F1-Score para a classe de recidiva. Isso indica que o modelo consegue discriminar bem entre pacientes que terão ou não recidiva, mantendo um compromisso aceitável entre precisão e sensibilidade.

O algoritmo *Naive Bayes*, por outro lado, priorizou a sensibilidade, alcançando o maior Recall (0.655) do estudo e identificando cerca de dois terços dos casos reais. No entanto, sua baixa precisão (0.121) implica uma alta taxa de falsos positivos. Já o *XGBoost* mostrou-se excessivamente conservador, com alta precisão (0.498) mas falhando em detectar a grande maioria dos casos positivos (Recall de 0.045).

A **Figura 1** ilustra as Curvas Precision-Recall e a Matriz de Confusão do melhor modelo geral (Random Forest). Nota-se graficamente que os modelos baseados em árvores (linhas superiores no gráfico da esquerda) mantêm uma precisão superior em diferentes faixas de sensibilidade.

## 5.2 Análise de Correlação de Variáveis

Para investigar as relações entre as variáveis clínicas e o desfecho de recidiva, foi realizada uma análise de correlação de Spearman, adequada para dados que podem apresentar relações monotônicas não-lineares. O mapa de calor na **Figura 2** destaca essas relações.

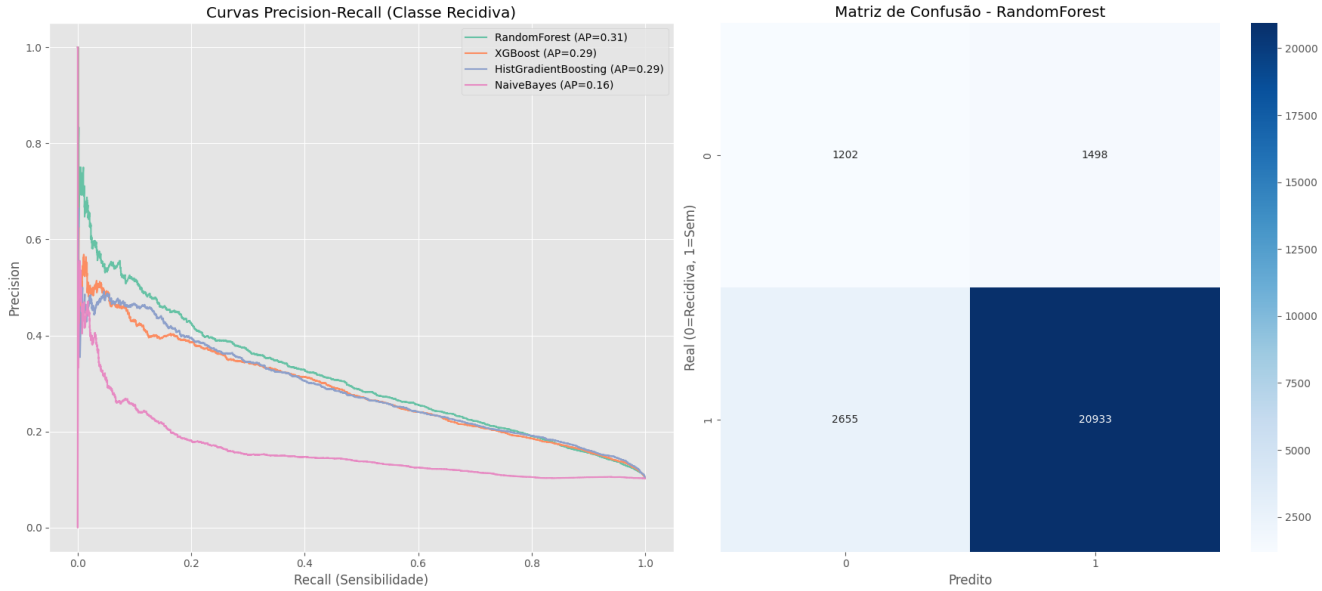


**Figura 1.** Esquerda: Curvas Precision-Recall comparativas. Direita: Matriz de Confusão do modelo Random Forest.

Variáveis como o tratamento hormonal (HORMONIO), estágio do tumor (T) e escore de Gleason apresentaram as correlações mais fortes com a variável alvo, corroborando a literatura médica que associa estadiamentos avançados e tratamentos mais agressivos a prognósticos mais reservados.

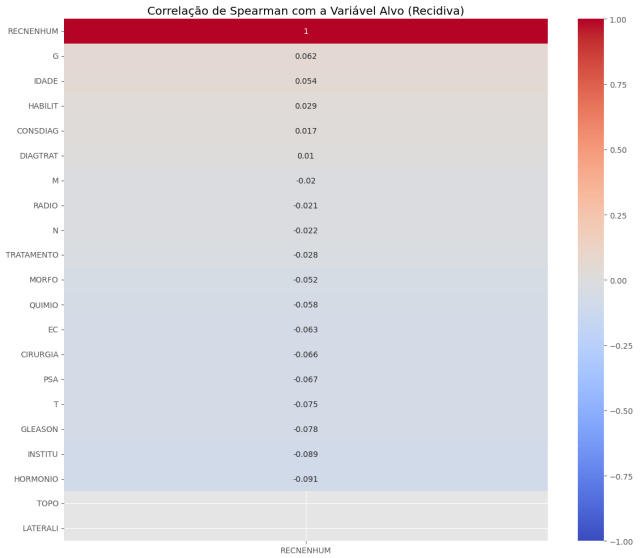


**Figura 2.** Correlação de Spearman das variáveis preditoras com a variável alvo (Recidiva).

# 6 Conclusão e Trabalhos Futuros

Este estudo demonstrou que algoritmos de aprendizado de máquina, quando combinados com técnicas de balanceamento de dados como o SMOTE, são ferramentas eficazes para a predição de recidiva de câncer de próstata utilizando dados públicos do RHC/SP.

Os resultados refutam a hipótese de desempenho aleatório: o algoritmo **Random Forest** alcançou uma AUC-ROC de 0.78, superando significativamente os *baselines* e demonstrando capacidade de generalização. Embora o *Naive Bayes* tenha oferecido maior sensibilidade, sua aplicação prática é limitada pela alta taxa de alarmes falsos, sendo mais indicado apenas para triagens iniciais de baixo custo.

Conclui-se que modelos de *ensemble* baseados em árvores oferecem o melhor caminho para sistemas de auxílio à decisão clínica neste domínio, equilibrando a detecção de casos com a confiabilidade das predições.

## 6.1 Trabalhos Futuros

Os resultados deste trabalho abrem caminho para diversas investigações futuras com o objetivo de aprimorar a capacidade preditiva dos modelos:

- **Otimização de Hiperparâmetros:** Conduzir uma busca sistemática (ex: *Grid Search* ou Otimização Bayesiana) para refinar o Random Forest, visando aumentar o Recall sem sacrificar drasticamente a precisão.
- **Aprendizado Sensível ao Custo:** Implementar funções de custo personalizadas que penalizem mais severamente os falsos negativos (pacientes com recidiva classificados como sadios) durante o treinamento do XGBoost e HistGradientBoosting.
- **Engenharia de Atributos e XAI:** Aprofundar a análise de variáveis com técnicas de *Explainable AI* (como SHAP values) para entender melhor quais interações clínicas estão guiando as decisões do modelo.
- **Modelos Alternativos:** Avaliar a aplicação de Redes Neurais Profundas (*Deep Learning*) em conjunto com dados tabulares para verificar se arquiteturas mais complexas podem extrair padrões latentes não capturados pelos modelos de árvore.

Em suma, a exploração conjunta dessas frentes — do pré-processamento avançado à otimização de modelos — representa um caminho metodológico promissor para o desenvolvimento de ferramentas prognósticas mais acuradas e clinicamente úteis.

# Declarações complementares

## Agradecimentos

Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio de uma bolsa do Programa Institucional de Bolsas de Iniciação Científica (PIBIC). Gostaría de expressar minha gratidão ao CTI Renato Archer e à Divisão de Metodologias de Computação (DIMEC) por todo o suporte oferecido.Ao meu orientador, Dr. Guilherme Ruppert, agradeço imensamente pela confiança depositada neste projeto, pelo seu tempo e por todo o apoio fundamental. Aos meus pais, minha mais profunda gratidão por todo o incentivo e suporte ao longo desta jornada.

## Financiamento

Esta pesquisa foi financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

## Contribuições dos autores

Reis contribuiu para a concepção, investigação, análise formal dos dados e foi o principal autor na redação do rascunho original do manuscrito. Ruppert contribuiu com a supervisão, revisão e edição crítica do trabalho. Todos os autores leram e aprovaram o manuscrito final.

## Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

## Disponibilidade de dados e materiais

Os conjuntos de dados analisados durante o estudo atual estão disponíveis em: Fundação Oncocentro de São Paulo (FOSP).

# Referências

Antunes, M. E., Araújo, T. G., Till, T. T., Pantaleão, E., Mancera, P. F. A., and Oliveira, M. H. d. (2025). Machine learning models for predicting prostate cancer recurrence and identifying potential molecular biomarkers. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40034593/. Último acesso em 28 de agosto de 2025.

Bigonha, R. S. (2025). *Fundamentos do Aprendizado de Máquina*. Disponível em: https://homepages.dcc.ufmg.br/~bigonha/Livros/ia-aprendizado.pdf. Último acesso em 28 de agosto de 2025.

Bonaccorso, G. (2017). *Machine Learning Algorithms: Reference Guide for Popular Algorithms for Data Science and Machine Learning*. Packt Publishing, Birmingham, UK. Disponível em: https://balasahebtarle.wordpress.com/wp-content/uploads/2020/01/machine-learning-algorithms_text-book.pdf. Acesso em: 8 de dezembro de 2025.

Fonseca, R. P., Fernandes Junior, A. S., Lima, V. S., Lima, S. S. S., Castro, A. F. d., Horta, H. d. L. e., and Favato Neto, B. (2007). Recidiva bioquímica em câncer de próstata: artigo de revisão. *Revista Brasileira de Cancerologia*, 53(2):167–172. DOI: 10.32635/2176-9745.RBC.2007v53n2.1812.

INCA (2025). Números do câncer. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros. Último acesso em 28 de agosto de 2025.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17. DOI: 10.1016/j.csbj.2014.11.005.

Lee, S. J., Yu, S. H., Kim, Y., Kim, J. K., Hong, J. H., Kim, C.-S., Seo, S. I., Byun, S.-S., Jeong, C. W., Lee, J. Y., and Choi, I. Y. (2020). Prediction system for prostate cancer recurrence using machine learning. *Journal of Clinical Medicine*. DOI: 10.3390/jcm9041200.

Maeda, A. E., Crocco, P. F., Ruppert, G. C. S., Dametto, M., and Bonacin, R. (2022). Um estudo sobre a predição da recidiva de câncer usando técnicas de aprendizado de máquina. In *XXIV Jornada de Iniciação Científica do Centro de Tecnologia da Informação Renato Archer (JICC 2022)*, Campinas, SP. Disponível em: https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/jicc/xxiv-jicc-2022. Último acesso em 28 de agosto de 2025.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. Disponível em: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning. Acesso em: 8 de dezembro de 2025.

Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*. Disponível em: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8. Último acesso em 28 de agosto de 2025.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*. Disponível em: https://arxiv.org/abs/1708.04649. Último acesso em 28 de agosto de 2025.

Zardeto, H. N., Schmidt, T. P., and Schneider, I. J. C. (2022). Prostate cancer: Analysis of survival and prognostic factors by age at diagnosis. *Research, Society and Development*, 11(8):e49411831344. DOI: 10.33448/rsd-v11i8.31344.