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

Giovani Luiz dos Reis[1,2], Guilherme Cesar Soares Ruppert[1],

[1]Divisao de Métodologias da Computação – DIMEC
CTI Renato Archer – Campinas/SP

[2]Análise e Desenvolvimento de Sistemas
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Hortolândia/SP

giovani.r@aluno.ifsp.edu.br,guilherme.ruppert@cti.gov.br

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.*

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.*

1. Introdução

[**?**] Segundo o Instituto Nacional do Câncer (INCA), o câncer de próstata (CaP) é o tipo de câncer mais comum entre homens no Brasil, com mais de 71.730 casos registrados em

2023. A recidiva — quando a doença retorna após remissão — pode ocorrer em diferentes localizações e está associada a fatores como: estágio avançado no diagnóstico, resposta incompleta ao tratamento, níveis elevados de PSA, algumas mutações genéticas, dentre outros.

O avanço da inteligência artificial tem impulsionado a aplicação de técnicas de aprendizado de máquina na construção de modelos preditivos capazes de capturar padrões não lineares e superar algumas limitações inerentes aos métodos estatísticos convencionais. Entretanto, sua utilização em análises de sobrevivência ainda é limitada, sobretudo devido às dificuldades no tratamento de dados censurados e ausentes.

Este trabalho propõe uma análise comparativa de quatro algoritmos de aprendizado de máquina — *Random Forest*, XGBoost, HistGradientBoosting e *Naive Bayes* — para a predição da recidiva do câncer de próstata, utilizando dados abertos da Fundação Oncocentro. O presente artigo está organizado da seguinte forma: a Seção 2 aborda os conceitos, técnicas, algoritmos e bases de dados utilizados; a Seção 3 detalha a metodologia empregada; a Seção 4 apresenta os resultados obtidos; a Seção 5 discute as conclusões e os encaminhamentos futuros da pesquisa; e a Seção 6 contempla as referências bibliográficas.

2. Conceitos, Técnicas e Algoritmos

2.1. 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.

2.2. 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.

2.3. 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.
- **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.
- **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.

- ***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.

2.4. Base de Dados do Oncocentro

O presente estudo utilizou um conjunto de dados público, anonimizado e de acesso aberto, disponibilizado pela **Fundação Oncocentro de São Paulo (FOSP)**. A base reúne informações de pacientes oncológicos do estado de São Paulo, incluindo variáveis *sociodemográficas* (idade, sexo, escolaridade), *clínico-patológicas* (estadiamento, topografia e histologia tumoral) e *terapêuticas* (cirurgia, radioterapia, quimioterapia). A **recidiva da doença** foi definida como a variável-alvo, a partir do acompanhamento temporal dos pacientes.

Essa base foi escolhida por sua abrangência e representatividade, permitindo a construção de modelos preditivos com potencial de aplicação prática no apoio ao diagnóstico e às políticas de saúde.

3. 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.

3.1. Fonte e Seleção de Dados

A fonte de dados deste estudo foi o Registro Hospitalar de Câncer do Estado de São Paulo (RHC/SP), disponibilizado pela Fundação Oncocentro de São Paulo (FOSP) e compreendendo o período de 2000 a 2024. Os dados originais estavam no formato DBF, exigindo a implementação de um algoritmo de conversão e organização.

A partir da base completa, foi realizada a filtragem dos registros referentes a pacientes diagnosticados com câncer de próstata. Para isso, utilizou-se a Classificação Internacional de Doenças para Oncologia, 3ª edição (CID-O-3), aplicando o código topográfico C61.9 (TOPO = C619).

A variável-alvo utilizada na modelagem foi **RECNENHUM** , um indicador binário da recidiva da doença:

- **0**: ocorrência de recidiva (classe minoritária);
- **1**: ausência de recidiva durante o acompanhamento (classe majoritária).

3.2. Pré-processamento

A preparação dos dados foi uma etapa fundamental para garantir a qualidade e a relevância das informações utilizadas na modelagem. O processo foi conduzido na seguinte ordem:

- **Conversão de Tipos de Dados:** Atributos de natureza numérica que estavam armazenados como texto (por exemplo, `T`, `N`, `M`, `PSA`, `GLEASON`) foram convertidos para o formato numérico. Valores que não puderam ser convertidos foram tratados como ausentes (`NaN`), permitindo um tratamento padronizado.
- **Remoção de Outliers:** Para mitigar o efeito de valores extremos e melhorar a robustez dos modelos, foi aplicada uma técnica de remoção de *outliers* às variáveis contínuas mais importantes (`IDADE`, `PSA`, `GLEASON`, `DIAGTRAT`, `CONSDIAG`). Utilizou-se o método do intervalo interquartil (IQR), no qual foram excluídos todos os registros cujos valores se encontravam fora dos limites definidos por $Q1 - 1.5 \times IQR$ e $Q3 + 1.5 \times IQR$.
- **Imputação de Valores Ausentes:** Os dados faltantes foram tratados de forma estratégica. Para as variáveis do sistema TNM (`T`, `N`, `M`, `G`, `EC`), foi imputado o valor 9, conforme indicado no dicionário de dados. Para `PSA` e `GLEASON`, optou-se por preencher com -1, tratando a ausência de informação como uma categoria distinta. Para as demais variáveis, foi utilizado o valor da mediana.

Para incorporar conhecimento de domínio, foram criadas três variáveis binárias: `ESTADIO_AVANCADO` (baseado em TNM), `PSA_GLEASON_ALTO` (baseado nos níveis de PSA e escore de Gleason) e `IDADE_65` (indicando pacientes com 65 anos ou mais). Esses novos atributos foram adicionados ao conjunto de dados, resultando em um total de 18 *features* para a modelagem.

3.3. 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 (RECNENHUM). Esta técnica garante que a proporção de casos de recidiva e não recidiva seja a mesma em ambas as amostras, o que é crucial para evitar vieses na avaliação de desempenho, especialmente em cenários com dados desbalanceados.

3.4. Balanceamento de Classes

O conjunto de dados de treino exibia um forte desbalanceamento, com a classe de "recidiva" (0) sendo significativamente minoritária. Para corrigir esse viés e permitir que os modelos aprendessem adequadamente os padrões de ambos os desfechos, foi aplicada a técnica **SMOTE** (*Synthetic Minority Over-sampling Technique*). É importante ressaltar que ela foi aplicada exclusivamente no conjunto de treinamento, para evitar o vazamento de dados (*data leakage*) para o conjunto de teste. A estratégia de amostragem foi customizada de forma a triplicar o número de amostras da classe minoritária, sem exceder a quantidade de amostras da classe majoritária.

3.5. Construção e Treinamento dos Modelos

Para a tarefa de classificação binária, foram empregados quatro algoritmos supervisionados: *GaussianNB(Naive Bayes)* , *RandomForestClassifier (Random Forest)*, *XGBClassifier (XGBoost)* e *HistGradientBoostingClassifier (HistGradientBoosting)*.

Antes do treinamento, foi aplicada a normalização dos dados com *StandardScaler*, que foi ajustada (fit) apenas com os dados de treino (já balanceados) e, em seguida, aplicada (transform) tanto no treino quanto no teste. Os modelos baseados em árvores (Random Forest, XGBoost, HistGradientBoosting) foram treinados com os dados balanceados e não normalizados, pois são inerentemente robustos a variações na escala dos atributos. O modelo GaussianNB, por sua vez, foi treinado com os dados normalizados.

3.6. Avaliação de Desempenho Utilizada

A performance dos modelos foi rigorosamente avaliada no conjunto de teste, utilizando métricas apropriadas para problemas de classificação com dados desbalanceados, onde a correta identificação da classe minoritária (*recidiva*) é o objetivo principal. As métricas selecionadas foram:

- ***Recall* (Sensibilidade):** Métrica de maior importância clínica, medindo a capacidade do modelo de identificar corretamente os casos reais de recidiva.
- ***Precision*:** Indica a proporção de predições de recidiva que foram, de fato, corretas.
- ***F1-Score*:** Média harmônica entre *Precision* e *Recall*. Foi utilizado como critério primário para a seleção do melhor modelo, por oferecer um balanço robusto entre a detecção de falsos negativos e falsos positivos.
- **AUC-ROC** e ***Average Precision* (AP):** Métricas que avaliam a capacidade discriminatória geral do modelo em diferentes limiares de decisão.

A análise foi complementada pela matriz de confusão, para uma visualização detalhada dos acertos e erros de cada modelo.

4. 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.

4.1. Desempenho Comparativo dos Modelos

O desempenho dos modelos apresentou resultados notáveis e contraintuitivos, conforme ilustrado na Figura 1 e detalhado nas matrizes de confusão. Os modelos de ensemble, como XGBoost e HistGradientBoosting, apesar de sua complexidade, exibiram um poder discriminatório (AUC-ROC) extremamente baixo, inferior a 0.25, indicando uma performance inferior à aleatória na distinção entre as classes.

A análise aprofundada das matrizes de confusão revela a razão deste baixo desempenho: estes modelos desenvolveram um forte viés para a classe majoritária (”Sem Recidiva”). O XGBoost, por exemplo, classificou corretamente 99.8% dos casos de não recidiva, mas falhou em identificar a classe minoritária, alcançando um recall para ”Recidiva”de apenas 1.8% (identificando somente 49 dos 2.700 casos reais).

Em contrapartida, o classificador *Naive Bayes*, embora mais simples, demonstrou a maior capacidade de identificar os casos de interesse clínico. Conforme a matriz de confusão na Figura 2, o *Naive Bayes* obteve o maior Recall para a classe ”Recidiva”entre

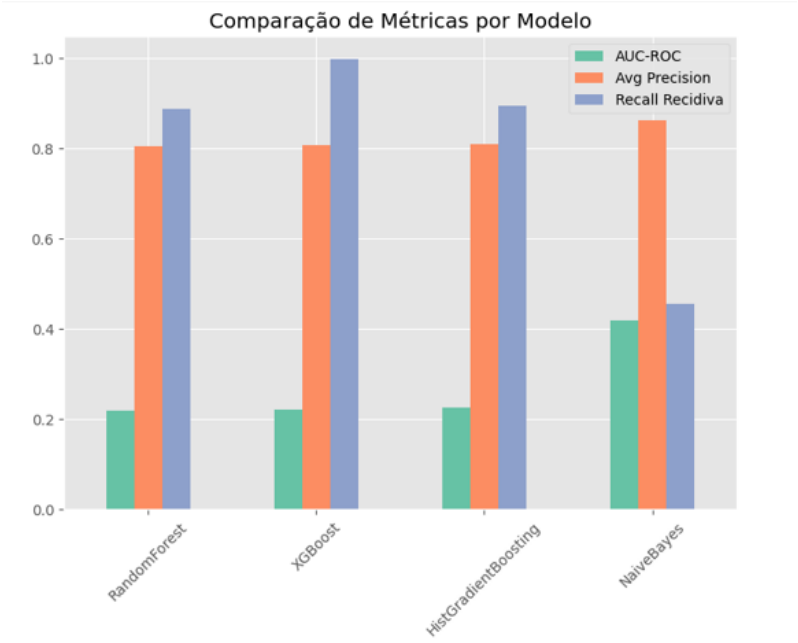

Figura 1. Desempenho dos modelos.

todos os modelos, identificando corretamente 1.768 casos (65.5%). Apesar de sua precisão ser mais baixa (resultando em um número maior de falsos positivos), ele foi o único modelo a prover um balanço mais útil para uma triagem inicial. As curvas Precision-Recall (Figura 3) e o valor de Average Precision (AP) de 0.863 confirmam que o Naive Bayes apresentou a melhor performance geral no contexto desbalanceado.

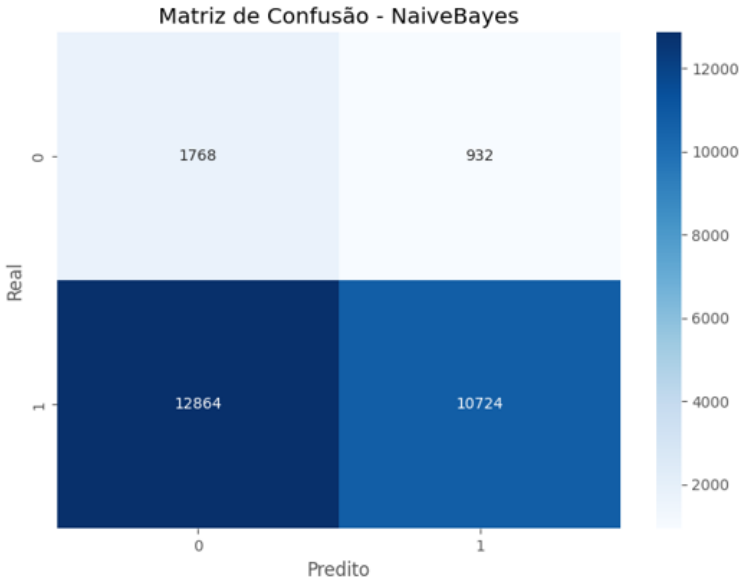

Figura 2. Matriz de Confusão - Naive Bayes

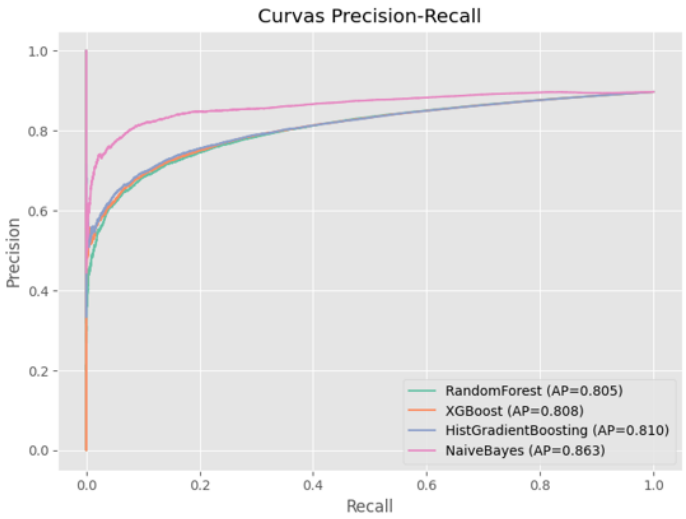

Figura 3. Curvas Precision-Recall

4.2. Análise de Importância de Atributos

Para investigar os fatores mais influentes na predição, foi analisada a importância dos atributos do modelo RandomForest (Figura 4). A variável TRATAMENTO emergiu como o preditor de maior impacto, sugerindo que a modalidade terapêutica é um fator crucial para o prognóstico da recidiva. Em seguida, destacam-se G (grau histológico), INSTITU (instituição de tratamento), EC (estadiamento clínico), PSA e GLEASON, todos consistentes com os fatores de risco conhecidos na literatura médica. Variáveis temporais como DIAGTRAT (tempo do diagnóstico ao tratamento) e demográficas como IDADE também demonstraram relevância, confirmando a natureza multifatorial do problema.

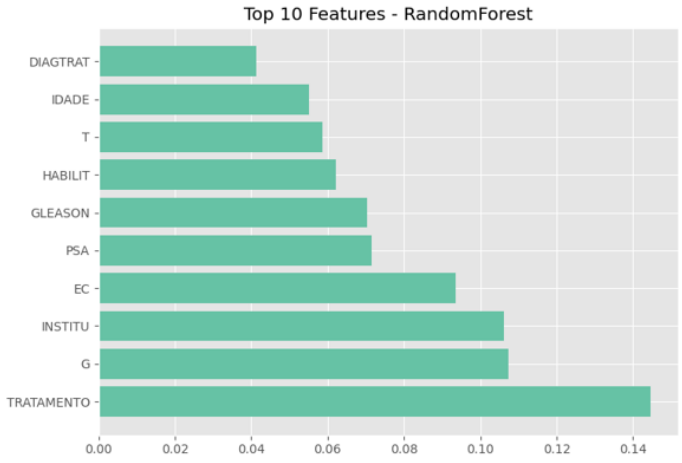

Figura 4. Top 10 Features - RandomForest

5. Conclusões e Trabalhos Futuros

5.1. Conclusões

Este estudo comparativo revelou que, no contexto de um conjunto de dados de recidiva de câncer de próstata com forte desbalanceamento de classes, a complexidade do modelo não se traduziu diretamente em melhor performance. O resultado mais significativo foi a superioridade do classificador probabilístico Naive Bayes em detrimento de algoritmos de ensemble avançados como XGBoost e RandomForest na tarefa de identificar corretamente os casos de recidiva (Recall da classe minoritária).

O resultado indica que os modelos baseados em árvores, mesmo após o rebalanceamento dos dados de treino com SMOTE, desenvolveram um viés para a classe majoritária, uma estratégia que maximiza a acurácia geral mas falha no objetivo clínico principal. Os baixos valores de AUC-ROC em todos os modelos indicam a dificuldade intrínseca do problema e a limitação dos atributos disponíveis para uma separação clara entre as classes. A performance superior do Naive Bayes, neste cenário, pode ser atribuída à sua menor sensibilidade ao sobreajuste dos padrões da classe majoritária, destacando a importância de selecionar modelos e métricas de avaliação que estejam estritamente alinhados ao objetivo clínico.

5.2. Trabalhos Futuros

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

- **Técnicas de Reamostragem Avançadas:** Explorar métodos de *under-sampling* para a classe majoritária ou variantes do SMOTE, como o ADASYN, que gera amostras sintéticas de forma mais adaptativa.
- **Engenharia de Atributos:** Aprofundar a criação de variáveis, explorando interações não lineares entre os preditores mais importantes e incorporando outras informações clínicas que não estavam disponíveis nesta análise.
- **Aprendizado Sensível ao Custo:** Implementar funções de custo que penalizem mais severamente os falsos negativos (pacientes com recidiva classificados como sadios), forçando os algoritmos a priorizarem o *recall* da classe minoritária.
- **Otimização de Hiperparâmetros:** Conduzir uma busca sistemática por hiperparâmetros (por exemplo, *Grid Search*, Otimização Bayesiana) para extrair o máximo potencial de cada algoritmo, especialmente dos modelos de ensemble.
- **Modelos Alternativos:** Avaliar outros paradigmas de modelagem, como Redes Neurais ou *Support Vector Machines* (SVM), que podem ser mais eficazes em capturar a complexidade do problema.

6. Referências

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

[2] Antunes, M. E.; Araújo, T. G.; Till, T. M.; Pantaleão, E.; Mancera, P. F. A.; Oliveira, M. H. de. Machine learning models for predicting prostate cancer recurrence and identifying potential molecular biomarkers. PubMed, 2025. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40034593/.

[3] Bigonha, R. S. Fundamentos do Aprendizado de Máquina. Disponível em: https://homepages.dcc.ufmg.br/~bigonha/Livros/ia-aprendizado.pdf.

[4] Maeda, A. E.; Crocco, P. F.; Ruppert, G. C. S.; Dametto, M.; Bonacin, R. Um Estudo sobre a Predição da Recidiva de Câncer Usando Técnicas de Aprendizado de Máquina. XXIV Jornada de Iniciação Científica do Centro de Tecnologia da Informação Renato Archer - JICC'2022 PIBIC/CNPq/CTI, Campinas, SP, 2022. Disponível em: https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/jicc/xxiv-jicc-2022. Último acesso em 28 de agosto de 2025.

[5] Antunes, M. E.; Mancera, P. F. A.; Araújo, T. G.; Pantaleão, E.; Oliveira, M. H. Seleção de Potenciais Biomarcadores para Previsão da Recidiva do Câncer de Próstata Utilizando Modelos de Classificação. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, XLIII CNMAC, Porto de Galinhas, PE, 2024, v. 11, n. 1.