Predição da Sobrevida em Pacientes com Câncer Colorretal: Modelos Estatísticos e de Aprendizado de Máquina

João Vitor Morimoto Sesma
Instituto Mauá de Tecnologia
São Caetano do Sul, São Paulo, Brasil
23.01516-0@maua.br

Rogério de Oliveira
Instituto Mauá de Tecnologia
Universidade Presbiteriana Mackenzie
São Caetano do Sul | São Paulo, São Paulo, Brasil
rogerio.oliveira@{maua.br|mackenzie.br}

Lucas Buk Cardoso
Instituto Mauá de Tecnologia
São Caetano do Sul, São Paulo, Brasil
lucas.cardoso@maua.br

Tatiana Natasha Toporcov
Faculdade de Saúde Pública da Universidade de São Paulo
São Paulo, São Paulo, Brasil
toporcov@usp.br

Resumo

Epidemiological studies of diseases assess the conditions and determinants of patients' health and are essential to support health policies aimed at the population. In view of technological advances, artificial intelligence has emerged as a tool to support solutions in several areas, including the health sector. This study aims to predict the survival of patients with colorectal cancer, based on epidemiological data, by comparing statistical survival models, classically used, with machine learning models. For this purpose, data from cancer case records in the state of São Paulo from 2000-2023 are used. Preliminary results in this study conclude that machine learning survival models do not outperform traditional statistical models (Cox model) in predicting survival time and suggest the adaptation of purely regressive models for better predictions.

Keywords

Epidemiology, Survival Models, Cox Regression, Machine Learning.

1 INTRODUÇÃO

O câncer é um problema de saúde mundial sendo uma das principais causas de morte e responsável pela redução da expectativa de vida. O aumento do número de casos na última década foi de 20% e até 2030 espera-se haja mais de 25 milhões de casos novos no mundo, sendo estimado cerca de 700 mil novos casos no Brasil para o triênio 2023-2025 [1].

O câncer colorretal é uma doença grave e está entre os tipos de câncer mais incidentes no mundo. No Brasil, o câncer de mama (mulheres) e da próstata (homens) são os mais comuns, mas são seguidos pelo câncer colorretal como o segundo mais incidente para homens e mulheres – levando a mais de 20 mil óbitos no ano de 2019 – sendo a região Sudeste e Sul as mais incidentes [2].

No Brasil, observa-se na última década uma melhora na disponibilidade na qualidade dos registros de câncer de base populacional (RCBP) dados que são fundamentais para se estabelecer o planejamento, monitoramento e avaliação de ações de controle do câncer [1], assim como os registros de base hospitalar (RHC). Sobre esses dados estudos e modelos epidemiológicos permitem fazer estimativas e previsões de incidência, sobrevida, recidiva e outros eventos de interesse, cruciais para fundamentar políticas públicas, a alocação racional dos recursos saúde e tratamentos mais eficazes [3].

Este estudo avalia a sobrevida de pacientes com câncer colorretal, com base no registro hospitalar de câncer (RHC) do estado de São Paulo (Brasil) no período de 2000 a 2023, comparando modelos estatísticos de sobrevida, classicamente empregados na epidemiologia do câncer [4], com recentes modelos de aprendizado de máquina, buscando assim contribuir para estimativas mais precisas e uma melhora nos mecanismos de vigilância do câncer.

2 REFERENCIAL TEÓRICO

Modelos estatísticos são o padrão em análises de sobrevivência na área de saúde, e mais recentemente diversos modelos de aprendizado de máquina vem sendo aplicados a diferentes tipos de câncer como mama [3], tireoide [5] e colorretal [6, 7, 8].

A análise de sobrevivência tal consiste basicamente na análise do tempo e do risco da ocorrência de um evento de interesse. Na medicina, muitas vezes o interesse recai sobre o prognóstico, ou seja, a previsão do tempo até que um evento adverso (por exemplo, morte ou recidiva) ocorra. Um dos desafios dessa análise é que o tempo de eventos passados é apenas parcialmente conhecido. No registro de casos de câncer, por exemplo, apenas um subconjunto de pacientes vai a óbito durante o período analisado e muitos pacientes viverão além do final do estudo. Há ainda casos em que não há notificação. Assim, o tempo exato da ocorrência do evento é conhecido apenas para os pacientes que realmente faleceram durante o período do estudo, enquanto para os demais pacientes, apenas podemos afirmar que houve notificação. Esses são os chamados dados censurados.

Dados censurados impedem a aplicação direta de modelos de aprendizado máquina, como os estimadores de rergressão, que aproximam o mapeamento de entradas e saídas conhecidas, pois não são conhecidas as saídas para os dados censurados que podem constituir uma grande parcela dos dados. Dados censurados não afetam apenas o treinamento dos modelos, mas também a avaliação dos modelos, pois os dados de teste também são à censura [9].

Desse modo, modelos tradicionais de aprendizado de máquina são adaptados para análise de sobrevivência. O tempo do evento é tratado como uma variável contínua ou categorizada, a censura é incorporada como um indicador binário (evento observado ou censurado) e as funções de custo, como erro quadrático médio (MSE) ou sua raiz (RMSE), são substituídas por funções de custo como a verossimilhança parcial de Cox para prever a função de risco (ou de sobrevivência) diretamente. É o caso, por exemplo, da

adaptação do modelo de floresta de árvores de decisão aleatórias de [10]. Diversos outros modelos tradicionais, como modelos de gradiente e máquinas de vetores de suporte vem sendo adaptados do mesmo modo [9].

3 SOLUÇÃO PROPOSTA

Neste estudo são aplicados 3 diferentes modelos para estimativa da sobrevida de pacientes com câncer do coloretal sobre a base de dados de RHC do estado de São Paulo [11].

Dados. Os casos de câncer do colorretal somam 37499 casos. A base populacional encontra-se caracterizada na Tabela 1 destacando-se o grande número de casos censurados (41%) e uma população predominantemente > 25 anos.

Tabela 1: Características da base populacional de pacientes com câncer do colorretal empregada

Atributo	Valor	Proporção
Evento	Observado	0.58
	Censurado	0.41
Gênero	Masculino	0.51
	Feminino	0.48
Idade	< 25	<0.01
	25-60	0.42
	> 60	0.57
Total de casos	37499	-
Período	2000-2023	-

Os dados incluem `escolaridade, idade, sexo`, variáveis binárias sobre tratamentos, como `cirurgia, radio, quimio, hormonio, imuno`, 11 variáveis binárias de topografia do tumor, data do diagnóstico, presença de recidiva ou metástase, num total de 34 potenciais preditores além das variáveis `tempo` e `evento`. O tempo de sobrevivência é estimado a valores contínuos para depois ser discretizado nas classes de tempo de sobrevivência de < 1 ano, entre $3-5$ anos e > 5 anos.

Modelos. Para estimativa da sobrevida, são aplicados o modelo de regressão de Cox, um modelo aprendizado de máquina de sobrevivência e um modelo clássico de aprendizado de máquina regressivo.

(1) **(COX) Regressão Cox.** Emprega-se o pacote `lifelines` [12] para implementar o modelo de Cox. Consideram-se várias combinações de preditores, mas o melhor C-Index (*índice de concordância*) é obtido com o total dos atributos. Três preditores são descartados (`escolaridade, imuno` e `habilit2`) por apresentarem coeficientes não significativos ($pvalue > 0.05$) para a regressão. O tempo de vida é previsto com o método `predict_expectation()`, que integra os valores da curva de sobrevivência estimadas do modelo.

(2) **(RFS) RandomSurvivalForest.** `scikit-survival` [9] é o pacote implementar modelo de floresta de árvores aleatórias adaptado (RFS) [10], e empregado aqui na estimativa do tempo dos eventos com censura. Todos atributos são empregados como preditores. Embora o pacote forneça, além do modelo de RFS, também modelos de *gradiente boost* e *máquinas vetorais de suporte*, eles apresentam resultados similares e apenas o RFS é empregado. O modelo é treinado com 1000 estimadores. Diferentemente do `lifelines`, o `scikit-survival`, não implementa nenhum método para estimativa do tempo de vida e os valores são obtidos integrando-se numericamente os valores das curvas de sobrevivência estimada.

(3) **(RFR) RandomForestRegressor, com *encode* do evento.** Um modelo clássico de floresta de árvores aleatórias para estimativa de valores contínuos RFR (regressivo) é encontrado no popular pacote `scikit-learn` [13]. A indicação de dados censurados, entretanto, não pode ser um preditor do modelo e uma alternativa é buscar *encode* dessa informação em conjunto com a variável alvo (o tempo de vida). Para isso adota-se o seguinte esquema para a variável alvo:

$$y = \begin{cases} -\lambda y, \text{ se evento observado} \\ y, \text{ se evento censurado} \end{cases}$$

O valor $\lambda = -0.75$ é selecionado de modo experimental, minimizando-se o erro (MSE) do modelo para valores de $\lambda \in [-2, -0.5]$. Outros métodos de seleção ou função devem ser explorados no futuro. Como no modelo anterior também são empregados todos atributos e 1000 estimadores.

4 RESULTADOS

As Tabelas 2, 3, 4 resumem os resultados respectivamente dos modelos **COX**, **RFS** e **RFR**.

Tabela 2: Classificação do tempo sobrevida - COX, Regressão de Cox

Anos	Precisão	Revocação	F1
< 1.0	0.82	0.18	0.30
1.0-3.0	0.47	0.29	0.36
5.0	0.40	0.92	0.55
Acuracidade			0.44
F1 macro			0.40

Tabela 3: Classificação do tempo sobrevida - RFS, RandomSurvivalForest

Anos	Precisão	Revocação	F1
< 1.0	0.95	0.11	0.20
1.0-3.0	0.46	0.27	0.34
5.0	0.39	0.95	0.55
Acuracidade			0.43
F1 macro			0.37

Os melhores resultados de acuracidade e F1 são obtidos pelo modelo **RFR**. As métricas de precisão, revocação e F1, permitem verificar se os erros e acertos não se concentram em um certo número de classes.

Os modelos de sobrevivência, por outro lado, apresentam um C-Index bastante superior como apresentado na Tabela 5.

Tabela 4: Classificação do tempo sobrevida - RFR, RandomForestRegressor, com *encode* do evento.

Anos	Precisão	Revocação	F1
< 1.0	0.83	0.67	0.75
1.0-3.0	0.27	0.63	0.38
5.0	0.51	0.26	0.34
Acuracidade			0.58
F1 macro			0.49

Tabela 5: Comparativo dos resultados dos diferentes modelos

Modelo	Acuracidade	F1	C-index
Regressão de Cox	0.44	0.40	0.75
RandomSurvivalForest	0.43	0.37	0.77
RandomForestRegressor	0.58	0.49	0.34

5 CONSIDERAÇÕES FINAIS

O modelo de regressão de Cox e os modelos de aprendizado de máquina para sobrevivência têm princípios semelhantes que priorizam a determinação do risco buscando maximizar o índice de concordância [10]. Tendo o mesmo princípio chegam a resultados bastante próximos na predição de valores dos tempos de sobrevivência, com de acuracidade de ≈ 0.44, o mesmo acontecendo com outros modelos adaptados de sobrevivência (*survival gradient boosted, survival support vector machine*). Essa semelhança de princípios pode ser observada proximidade das curvas de sobrevivência estimadas de ambos modelos (Figura 1).

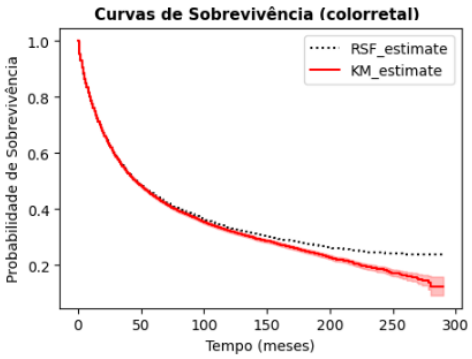

Figura 1: Curvas de sobrevivência estimadas dos modelos de Regressão de Cox e RandomForestSurvival. Ambos, seguindo princípios semelhantes de otimização, geram estimativas muito próximas.

Já o modelo aprendizado de máquina para regressão, adaptado com o encode do evento à variável alvo, tendo como princípio a minimização do erro (como MSE, RMSE), apresenta resultados melhores de predição com de acuracidade para as mesmas classes de tempo de vida.

Independente do modelo, os resultados parecem a princípio bastante baixos. A acuracidade do melhor modelo, 0.58, está abaixo dos valores >> 0.7 esperado de bons modelos. Mas o valor encontrado parece encontra-se dentro dos valores encontrados na literatura e na revisão de Li *et al.* [3] são encontrados valores de acuracidade de 0.51 a 0.97, mas com alto risco de viés na maior parte dos estudos. Também não são muitas vezes totalmente claros os dados e métodos empregados e, certamente, poucos ou a ausência de dados censurados podem levar a obtenção de resultados de grande acuracidade facilmente.

Importante destacar que este resumo apresenta resultados prelimares de um estudo ainda em andamento. Apresenta, assim, algumas limitações, em particular quanto a discussão dos resultados. Algumas dessas limitações devem ser tratadas no futuro com a ampliação dos dados (outros tipos de câncer) e das análises. De qualquer modo, em seu estágio atual, a comparação dos métodos estatísticos tradicionais com os modelos de aprendizado de máquina, de sobrevivência ou clássicos, não apresentou necessariamente um melhoria dos resultados. Os modelos de aprendizado para sobrevivência, não parecem permitir um ganho muito grande com relação aos modelos estatísticos, mas parece ser promissor o uso modelos de aprendizado de máquina que incorporem, de algum modo à variável alvo, a característica do evento (observado ou censurado) e outras técnicas, além da que foi aplicada neste estudo, ainda podem ser exploradas. A aplicação a outros tipos de câncer (mama, pulmão, próstata etc., mais comuns e com maior número de casos) são ainda requeridos para uma melhor avaliação desses resultados preliminares.

Referências

[1] Marceli de Oliveira Santos, Fernanda Cristina da Silva de Lima, Luís Felipe Leite Martins, Julio Fernando Pinto Oliveira, Liz Maria de Almeida e Marianna de Camargo Cancela. 2023. Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, 69, 1.

[2] Jesnaira Leite da Silva e Agnes Sousa Silva. 2022. Epidemiologia e os tipos de câncer de maior incidência no brasil: revisão integrativa de literatura epidemiology and the most common types of cancer in brazil: an integrative literature review. *Brazilian Journal of Development*, 8, 7, 51703–51711.

[3] Jiaxin Li, Zijun Zhou, Jianyu Dong, Ying Fu, Yuan Li, Ze Luan e Xin Peng. 2021. Predicting breast cancer 5-year survival using machine learning: a systematic review. *PloS one*, 16, 4, e0250370.

[4] Mei-Jie Zhang. 2002. Cox proportional hazards regression models for survival data in cancer research. *Biostatistical applications in cancer research*, 59–70.

[5] Lizhen Xu, Liangchun Cai, Zheng Zhu e Gang Chen. 2023. Comparison of the cox regression to machine learning in predicting the survival of anaplastic thyroid carcinoma. *BMC endocrine disorders*, 23, 1, 129.

[6] Shayeste Alinia, Mohammad Asghari-Jafarabadi, Leila Mahmoudi, Solmaz Norouzi, Maliheh Safari e Ghodratollah Roshanaei. 2023. Survival prediction and prognostic factors in colorectal cancer after curative surgery: insights from cox regression and neural networks. *Scientific Reports*, 13, 1, 15675.

[7] Lucas Buk Cardoso, Vanderlei Cunha Parro, Stela Verzinhasse Peres, Maria Paula Curado, Gisele Aparecida Fernandes, Victor Wünsch Filho e Tatiana Natasha Toporcov. 2023. Machine learning for predicting survival of colorectal cancer patients. *Scientific reports*, 13, 1, 8874.

[8] Oliver Kennion, Stuart Maitland e Richard Brady. 2022. Machine learning as a new horizon for colorectal cancer risk prediction? a systematic review. *Health Sciences Review*, 4, 100041.

[9] Sebastian Pölsterl. 2020. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21, 212, 1–6. http://jmlr.org/papers/v21/20-729.html.

[10] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone e Michael S Lauer. 2008. Random survival forests.

[11] FOSP. 2024. Diretoria adjunta de informação e epidemiologia: banco de dados do rhc. (2024). Retrieved 2 de dezembro de 2024 from https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/.

[12] Cameron Davidson-Pilon. 2019. Lifelines: survival analysis in python. *Journal of Open Source Software*, 4, 40, 1317. DOI: 10.21105/joss.01317.

[13] F. Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.