UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

HENRIQUE MARCHETTI

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA NA PREDIÇÃO DE SOBREVIDA DE PACIENTES ONCOLÓGICOS

Revisado pelo(a) Orientador(a)

Assinatura do(a) Orientador(a)

Data 15/12/2025

PRESIDENTE PRUDENTE

2025

HENRIQUE MARCHETTI

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA NA PREDIÇÃO DE SOBREVIDA DE PACIENTES ONCOLÓGICOS

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista "Júlio de Mesquita Filho" - FCT/UNESP, para aproveitamento na disciplina de Trabalho de Conclusão de Curso II (TCC II).

Orientador: Prof. Dr. Sérgio Minoru Oikawa

PRESIDENTE PRUDENTE

2025

M317u Marchetti, Henrique

Utilização de aprendizado de máquina na predição de sobrevida de pacientes oncológicos / Henrique Marchetti. -- Presidente Prudente,
2025

37 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) Universidade Estadual Paulista (UNESP), Faculdade de Ciências e
Tecnologia, Presidente Prudente

Orientador: Sérgio Minoru Oikawa

1. Análise de sobrevivência. 2. Câncer de estômago. 3. Aprendizado de máquina. 4. Random Survival Forest. 5. Modelo de Cox. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada

TERMO DE APROVAÇÃO

HENRIQUE MARCHETTI

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA NA PREDIÇÃO DE SOBREVIDA DE PACIENTES ONCOLÓGICOS

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Dr. SÉRGIO MINORU OIKAWA

Departamento de Estatística

Documento assinado digitalmente
gov.br
MARIO HISSAMITSU TARUMOTO
Data: 15/12/2025 18:40:28-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. MÁRIO HISSAMITSU TARUMOTO

Departamento de Estatística

Documento assinado digitalmente
gov.br
FERNANDO ANTONIO MOALA
Data: 15/12/2025 22:06:39-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. FERNANDO ANTONIO MOALA

Departamento de Estatística

Presidente Prudente, 15 de dezembro de 2025.

Resumo

A análise de sobrevivência desempenha um papel fundamental na estatística médica, lidando com dados censurados para estimar o tempo até a ocorrência de eventos de interesse, como o óbito em pacientes oncológicos. Neste contexto, o câncer gástrico destaca-se pela alta letalidade e complexidade clínica, demandando ferramentas prognósticas precisas. O presente trabalho teve como objetivo principal realizar uma análise comparativa de desempenho entre o modelo semiparamétrico de Riscos Proporcionais de Cox e o algoritmo de aprendizado de máquina *Random Survival Forest* (RSF). Especificamente, buscou-se identificar os fatores de risco clínicos determinantes, avaliar a capacidade preditiva e determinar a abordagem mais eficaz para prognóstico. A metodologia utilizou uma base de dados real da Fundação Oncocentro de São Paulo (FOSP), compreendendo 942 pacientes diagnosticados com câncer de estômago nos estágios I, II e III, acompanhados entre 2018 e 2024. Foram ajustados modelos de Cox, e modelos RSF, treinados tanto com variáveis originais quanto estratificadas, utilizando métricas como *C-Index*, *Brier Score* (IBS) e VIMP em conjuntos de treino e teste.

Os resultados clínicos via regressão de Cox indicaram que o estadiamento avançado (Estágio III) eleva o risco de óbito em 4,78 vezes comparado ao Estágio I, e que o atendimento pelo SUS está associado a um risco de óbito 92% superior ao da rede privada. Na comparação metodológica, o modelo RSF treinado com variáveis originais superou o modelo com variáveis recategorizadas, alcançando um *C-Index* de 0,74 no conjunto de teste e demonstrando menor erro de predição. Conclui-se que o processo de recategorização exigido pelo modelo de Cox pode acarretar perda de informação relevante. O estudo evidencia que, enquanto o modelo de Cox é indispensável para a interpretação da magnitude dos riscos (Risco Relativo), o *Random Survival Forest* apresenta superioridade na precisão preditiva e generalização, constituindo uma ferramenta robusta para suporte à decisão clínica.

Palavras-chave: Random Survival Forest. Análise de Sobrevivência. Aprendizado de Máquina. Oncologia. Regressão de Cox.

Abstract

Survival analysis plays a fundamental role in medical statistics, dealing with censored data to estimate the time until the occurrence of events of interest, such as death in oncological patients. In this context, gastric cancer stands out due to its high lethality and clinical complexity, demanding precise prognostic tools. The primary objective of this study was to perform a comparative performance analysis between the semiparametric Cox Proportional Hazards model and the Random Survival Forest (RSF) machine learning algorithm. Specifically, the study sought to identify determinant clinical risk factors, evaluate predictive capacity, and determine the most effective approach for prognosis. The methodology utilized a real database from the Fundação Oncocentro de São Paulo (FOSP), comprising 942 patients diagnosed with gastric cancer in stages I, II, and III, followed up between 2018 and 2024. Cox models were fitted, and RSF models were trained with both original and stratified variables, using metrics such as C-Index, Brier Score (IBS), and VIMP on training and testing sets.

The clinical results via Cox regression indicated that advanced staging (Stage III) increases the risk of death by 4.78 times compared to Stage I, and that care provided by the Unified Health System (SUS) is associated with a 92% higher risk of death compared to the private network. In the methodological comparison, the RSF model trained with original variables outperformed the model with recategorized variables, achieving a C-Index of 0.74 in the test set and demonstrating lower prediction error. It is concluded that the recategorization process required by the Cox model may lead to a loss of relevant information. The study highlights that, while the Cox model is indispensable for interpreting the magnitude of risks (Relative Risk), the Random Survival Forest demonstrates superiority in predictive accuracy and generalization, constituting a robust tool for clinical decision support.

Keywords: Random Survival Forest. Survival Analysis. Machine Learning. Oncology. Cox Regression.

SUMÁRIO

1 INTRODUÇÃO 7

2 REFERENCIAL TEÓRICO 9

2.1 ANÁLISE DE SOBREVIVÊNCIA 9
2.1.1 Função de Sobrevivência 10
2.1.2 Função de Risco 11
2.1.3 Função de Risco Acumulada 11
2.1.4 Estimadores não paramétricos 11
2.1.5 Teste Log-rank 12
2.2 MODELO DE REGRESSÃO DE COX 13
2.3 APRENDIZADO DE MÁQUINA 14
2.3.1 Árvore de Sobrevivência 14
2.3.2 Divisão Log-rank 15
2.3.4 Random Survival Forests 16
2.3.5 Estimadores 18
2.3.6 Estatísticas de desempenho 19
2.3.7 Mortalidade 20
2.3.8 C-Index 20
2.3.9 Brier Escore Integrado 21
2.3.10 Importância das variáveis 22

3 METODOLOGIA 23

4 DISCUSSÃO E RESULTADOS 24

4.1 ANÁLISE DESCRITIVA 24
4.2 ANÁLISE DAS CURVAS DE SOBREVIVÊNCIA 25
4.3 MODELO DE REGRESSÃO DE COX 28
4.4 ADEQUAÇÃO DO MODELO 29
4.5 RANDOM SURVIVAL FOREST 32

5 CONCLUSÃO 35

REFERÊNCIAS 36

1 INTRODUÇÃO

A evolução da coleta, processamento, armazenamento e tratamento de dados, impulsionada pelos avanços computacionais, possibilitou a integração das técnicas de aprendizado de máquina (AM) com a estatística. Essas técnicas, originalmente desenvolvidas na década de 60 como parte da inteligência artificial e predominantemente computacionais, passaram a ser amplamente adotadas por estatísticos a partir dos anos 90 (IZBICKI; SANTOS, 2020). As técnicas de AM combinam modelos estatísticos e algoritmos computacionais para extrair informações de conjuntos de dados extensos com múltiplas variáveis (MORETTIN; SINGER, 2021). Geralmente, seu principal objetivo é a predição de novas observações.

Consequentemente, o desenvolvimento de novas formas de aquisição de dados e a tecnologia de big data possibilitaram a coleta de uma ampla variedade de dados e seu acompanhamento ao longo de períodos prolongados (WANG; LI; REDDY, 2019). Em muitos casos, o principal objetivo dessas novas possibilidades é obter estimativas do tempo até a ocorrência de um evento de interesse. No entanto, um dos principais desafios do acompanhamento em longo prazo ocorre quando há censura nos dados, ou seja, o evento de interesse não é observado devido a restrições de tempo da pesquisa ou perda de acompanhamento da observação, por diversos motivos.

Nesse contexto, a análise de sobrevivência é um ramo da estatística que lida com esse tipo de dados, essencialmente em situações médicas envolvendo dados censurados, em que a variável resposta é o tempo até a ocorrência de um evento de interesse (COLOSIMO; GIOLO, 2006). Em geral, os métodos de análise de sobrevivência podem ser classificados em duas categorias principais: métodos estatísticos e métodos baseados em aprendizado de máquina. Ambos os métodos têm como objetivo a previsão do tempo de sobrevivência e a estimativa da probabilidade de sobrevivência nesse tempo. No entanto, os modelos estatísticos concentram-se em caracterizar tanto a distribuição dos tempos dos eventos quanto às propriedades estatísticas do estimador de parâmetros, estimando as curvas de sobrevivência, enquanto os métodos de aprendizado de máquina focam principalmente na previsão da ocorrência do evento de interesse em um momento específico, combinando métodos de análise de sobrevivência tradicionais com técnicas computacionais (WANG; LI; REDDY, 2019).

Este trabalho se concentra no estudo das técnicas de aprendizado de máquina, com ênfase na extensão das conhecidas *random forests* introduzidas por Breiman (2001),

para trabalhar com dados de sobrevivência, a *Random Survival Forests* (ISHWARAN et al., 2008) e, no estudo de técnicas clássicas de estatística para se trabalhar com dados de sobrevivência como o método de estimação proposto por Kaplan e Meier (1958) e o Modelo de regressão de Cox.

Desse modo, o presente trabalho tem como propósito realizar uma análise comparativa entre o algoritmo de aprendizado de máquina Random Survival Forest e o modelo de Riscos Proporcionais de Cox, buscando entender as vantagens de cada um. Para isso, o estudo envolve a compreensão dos fundamentos da análise de sobrevivência e das técnicas de aprendizado de máquina, a investigação dos algoritmos utilizados na predição do tempo até o evento, a comparação dos resultados obtidos pelos modelos e a análise de como as variáveis influenciam na ocorrência da falha dos indivíduos.

Para realizar esse trabalho, utilizaremos uma base de dados real de pacientes oncológicos. De acordo com o Observatório Global do Câncer (GLOBOCAN), em 2020, estima-se que houve 19,3 milhões de casos de câncer e quase 10 milhões de mortes, tornando-o uma das principais causas de morte e um obstáculo para o aumento da expectativa de vida em todo o mundo. A estimativa da GLOBOCAN para 2040 é de 28,4 milhões de casos de câncer (SUNG et al., 2021).

Dentre os diversos tipos de câncer, o câncer gástrico, é um dos que merece destaque devido à sua alta letalidade e complexidade clínica. Mundialmente, é o quinto tipo de câncer mais comum e a quarta principal causa de morte por câncer (SUNG et al., 2021).

A sobrevivência dos pacientes com câncer de estômago é fortemente influenciada pelo estágio em que a doença é diagnosticada. Enquanto casos detectados precocemente apresentam prognósticos favoráveis, tumores em estágios avançados possuem opções terapêuticas limitadas e baixas taxas de sobrevida. Segundo o Instituto Nacional do Câncer (2022), a cirurgia é o principal tratamento em casos iniciais, em que a doença ainda não se espalhou.

2 REFERENCIAL TEÓRICO

2.1 Análise de Sobrevivência

Análise de Sobrevivência é uma área da estatística que tem como interesse avaliar o tempo decorrido até um ou mais eventos de interesse, como a morte de pacientes submetidos a certo tratamento ou a quebra de um equipamento mecânico, chamados de falha. Sendo assim, os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falhas e, geralmente, pelas censuras. Quando se está interessado em estudos clínicos, um conjunto de covariáveis também é medido para cada paciente (COLOSIMO & GIOLO, 2006).

Para caracterizarmos um tempo de falha, devemos definir o tempo de início do estudo, como por exemplo, o diagnóstico de alguma doença, a escala de medida adotada, e o evento de interesse, como por exemplo a morte do paciente.

A censura, por sua vez, é definida como dados que foram observados parcialmente ou de maneira incompleta, podendo ocorrer por diversas razões, dentre elas, a perda de acompanhamento do paciente e o término do estudo antes do indivíduo falhar. Nos modelos de análise de sobrevivência, os dados censurados são incluídos nas análises. Segundo Colosimo e Giolo (2006) existem dois motivos para tal: “(i) mesmo sendo incompletas, as observações censuradas nos fornecem informações sobre o tempo de vida de pacientes; (ii) a omissão das censuras no cálculo das estatísticas de interesse pode acarretar conclusões viciadas.”

Existem três tipos de censura que podem ser consideradas em estudos de sobrevivência, são elas: a censura a direita, para a qual se conhece o instante em que a característica de interesse ocorreu, porém a falha não é observada até a conclusão do estudo, a censura a esquerda, onde não se conhece o instante inicial que corre a característica de interesse, porém a falha ocorre durante o estudo e, por fim a censura intervalar, para a qual não se conhece o instante exato que ocorre a falha, porém sabe-se o intervalo de tempo que ela ocorreu. Nesse trabalho utilizaremos dados com censura à direita, que são os casos mais comuns em estudos de medicina e engenharia. A censura à direita pode ainda ser classificada como censura do tipo I, que é quando o estudo é terminado após um período pré-estabelecido de tempo e a censura do tipo II, em que o estudo é terminado após a observação do evento de interesse em um número pré-estabelecido de indivíduos. Existe um terceiro tipo, chamado de censura aleatória, sendo

o mais comum na prática médica, ocorre quando um paciente é retirado durante o andamento do estudo sem ter ocorrido a falha. Ocorre, por exemplo, se o paciente morre de uma causa diferente da estudada.

A predição da taxa de sobrevivência é um dos maiores objetivos da análise de sobrevivência. No caso de observação de dados até o evento de interesse, o modelo de regressão de Cox é o método mais utilizado. O modelo de Cox pode ser utilizado para identificar as variáveis que significativamente afetam a variável resposta. (YOSEFIAN et al., 2018). No entanto, esse método possui suposições restritivas, como é o caso dos riscos proporcionais.

Sendo assim, *Random Survival Forests* (ISHWARAN et al., 2008), uma extensão das florestas aleatórias para se trabalhar com dados de sobrevivência contendo censura à direita, se torna uma alternativa, quando se está interessado em um modelo preditivo.

Algumas funções extensamente utilizadas em análise de sobrevivência serão descritas a seguir, estas, serão utilizadas nos próximos passos.

2.1.1 Função de Sobrevivência

A função de sobrevivência é descrita como a probabilidade de uma observação não falhar (ou sobreviver) até um certo tempo t. Em termos probabilísticos, temos
Sendo:

$$S(t) = P(T \geq t)$$

• T é uma variável aleatória que representa o tempo até a ocorrência do evento de interesse.
• t é o tempo específico para o qual a probabilidade de sobrevivência está sendo calculada.
• P (T > t) é a probabilidade de que a variável T seja maior do que t, ou seja, a probabilidade de sobrevivência além do tempo t.

2.1.2 Função de Risco

A função de Risco $h(t)$ é extensamente utilizada. Ela descreve a taxa de risco de um evento ocorrer em um determinado intervalo de tempo, condicionado ao fato de o indivíduo ter sobrevivido até aquele momento.

Ela é definida como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(\, t \leq T < t + \Delta t \mid T \geq t\,)}{\Delta t}$$

Observa-se que quando Δt é muito pequeno, $\mathrm{h}(t)$ representa a taxa de falha instantânea para t, condicional a sobrevivência até o tempo t.

2.1.3 Função de Risco Acumulada

A função de risco acumulada fornece a taxa de falha acumulada do indivíduo. É definida por:

$$H(t) = \int_0^t \mathrm{h(u)du}.$$

Segundo Colosimo e Giolo (2006), $\mathrm{H}(t)$ não tem interpretação direta, mas pode ser útil para estimar $h(t)$ pois é mais fácil de estimar por métodos não-paramétricos que a função de riscos.

2.1.4 Estimadores não paramétricos

O estimador de Kaplan-Meier é um método não paramétrico utilizado para estimar a função de sobrevivência $S(t)$, proposto por Kaplan e Meier (1958), continua sendo o método mais utilizado para este fim (COLOSIMO & GIOLO, 2006). Esse modelo é aplicado por ser não viciado para grandes amostras e também por incluir no estimador, casos censurados. Considerando:

$t_1 < t_2 < \cdots < t_k$ os k tempos distintos e ordenados de falha;

d_j o número de falhas ocorridos no tempo t_j , j = 1, ..., k

n_j o número de indivíduos em risco no tempo t_j,
O estimador de Kaplan-Meier é definido como:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(\frac{d_j - n_j}{n_j} \right) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j} \right)$$

O estimador de Kaplan-Meier permite a visualização dos tempos de falha por meio da curva de sobrevivência. Tal estimativa resulta em uma função do tipo escada, cujos degraus demarcam os tempos distintos das falhas observadas.

Segundo Colosimo e Giolo (2006) $\widehat{S}(t)$ é o estimador de máxima verossimilhança de $S(t)$, sendo não viciado para grandes amostras e converge assintoticamente para um processo gaussiano.

2.1.5 Teste Log-rank

O teste Log-rank (MANTEL, 1966) é uma das ferramentas mais utilizadas na análise de sobrevivência para a comparação de funções de sobrevivência entre diferentes grupos. Segundo Colosimo e Giolo (2006), este teste é apropriado quando as populações comparadas satisfazem a propriedade de riscos proporcionais, isto é, quando a razão entre as taxas de falha dos grupos permanece aproximadamente constante ao longo do tempo.

A estatística do teste baseia-se na comparação entre o número de falhas observadas em cada grupo e o número de falhas que seriam esperadas caso a hipótese nula seja verdadeira. As hipóteses a serem testadas pelo teste Log-rank são dados por:

$$H_0: as\ curvas\ de\ sobrevivência\ são\ iguais$$
$$H_A: as\ curvas\ de\ sobrevivência\ não\ são\ iguais$$

A estatística Log-rank é dada da forma:

$$\chi^2_{LR} = \sum \frac{\left(\sum O_{jt} - \sum E_{jt} \right)^2}{\sum E_{jt}} \sim \chi^2_{k-1}$$

Onde $\sum O_{jt}$ representa a soma do número observado de eventos no j-ésimo grupo, $\sum E_{jt}$ representa a soma do número esperado de eventos no j-ésimo grupo e k representa o número de grupos em comparação. A hipótese H_0 é rejeitada se $\chi^2_{LR} > \chi^2_{k-1}$.

2.2 Modelo de regressão de COX

Conforme apontado por Colosimo e Giolo (2006), o modelo de regressão de Cox destaca-se como uma das técnicas mais empregada em estudos clínicos devido à sua grande flexibilidade. A introdução deste método por Cox (1972) marcou uma revolução na modelagem de dados de sobrevivência.

A principal utilidade desse modelo é analisar dados de tempo de vida, relacionando o tempo até a ocorrência de um evento, com diversas covariáveis. Em sua configuração mais elementar, podemos imaginar o modelo com apenas uma covariável que divide os indivíduos em dois grupos distintos: um que recebe o tratamento padrão e outro submetido a um tratamento novo.

Sejam $\lambda_0(t)$ e $\lambda_1(t)$ as taxas de falha para o grupo padrão e o grupo de teste, respectivamente. O modelo assume que a razão entre essas taxas é uma constante K durante todo o tempo t do estudo, expressa por:

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K$$

Considerando x como uma variável indicadora (onde $x = 0$ para o grupo padrão e $x = 1$ para o grupo teste) e definindo a constante de proporcionalidade como $K = \exp\{\beta x\}$ a equação para uma única covariável assume a forma:

$$\lambda(t) = \lambda_0(t)\exp\{\beta x\}$$

,

Ou seja:

$$\lambda(t) = \begin{cases} \lambda_0(t)\exp\{\beta x\}, se\ x = 1 \\ \lambda_0(t), \ \ se\ x = 0 \end{cases}$$

Expandindo esse conceito para uma situação mais realista com múltiplas variáveis, onde $x = (x_1, \ldots, x_p)'$ representa um vetor de covariáveis, a formulação geral do modelo de Cox é:

$$\lambda(t) = \lambda_0(t)g(x'\beta)$$

Nesta equação, o modelo combina dois componentes distintos:

Componente não paramétrico: Representado por $\lambda_0(t)$, é uma função não negativa do tempo que não precisa ser especificada. Componente paramétrico: Representado pela

função $g(\boldsymbol{x}'\beta)$. A forma mais comum de utilização é a multiplicativa, dada por $exp(\boldsymbol{x}'\beta)$, pois a função exponencial garante que o resultado seja sempre positivo.
Assim, temos:

$$\lambda(t|x) = \lambda_0(t)\exp\{x'\beta\} = \lambda_0(t)\exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}$$

Onde β é o vetor dos coeficientes que medem o efeito de cada covariável.
Este método é popularmente conhecido como modelo de riscos proporcionais. Essa denominação deve-se ao fato de que, ao compararmos dois indivíduos (i e j) com características diferentes, a razão entre suas taxas de falha anula o componente do tempo $\lambda_0(t)$, resultando em um valor constante no tempo:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{x_i'\beta\}}{\lambda_0(t)\exp\{x_j'\beta\}} = \exp\{x_i'\beta - x_j'\beta\}$$

Isso significa que o risco relativo entre dois pacientes não muda com o passar do tempo. Por exemplo, se no início do acompanhamento um paciente apresenta um risco de óbito duas vezes maior que outro, essa proporção de risco (2 para 1) se manterá inalterada até o final do estudo.

2.3 APRENDIZADO DE MÁQUINA

2.3.1 Árvore de Sobrevivência

Segundo Izbicki (2020), as árvores de regressão são métodos não paramétricos que proporcionam resultados com alta interpretabilidade, onde a árvore é construída particionando recursivamente o espaço das covariáveis, com intuito de gerar dois subconjuntos mais homogêneos em relação a variável resposta. A árvore é particionada até um critério de parada, onde o número de observações é mínimo ou a homogeneidade alcançada é máxima, cada partição da árvore é chamada de nó e cada resultado final, de folha.

Por sua vez, segundo Ishwaran et al. (2008), às árvores de sobrevivência são adaptadas para trabalhar com dados com presença de censura, característicos de um estudo de sobrevivência. O critério adotado para a partição em grupos homogêneos utiliza uma regra de divisão que busca maximizar a diferença de sobrevivência entre os

subconjuntos gerados. O processo de divisão continua até um nó terminal ou folha, em que um critério de parada foi alcançado.

Ao alcançar o critério de parada, a árvore se torna saturada, desse modo, todos os indivíduos presentes nas folhas têm uma sobrevivência homogênea, e assim, podemos atribuir a estes, as mesmas estimativas não paramétricas de sobrevivência.

2.3.2 Divisão Log-rank

De acordo com Ishwaran e Lu (2018), a divisão log-rank é uma das regras mais populares quando trabalhamos com as árvores de sobrevivência. Baseado no teste estatístico log-rank, habitualmente utilizado para compararmos dois ou mais grupos em dados de sobrevivência, neste caso, adaptado para maximizar a diferença entre os nós.

A divisão log-rank é dada como segue Ishwaran e Lu (2018), denotamos o nó a ser dividido como h e os dados como $(T_1, X_1, \delta_1), \dots, (T_n, X_n, \delta_n)$, onde X_i são as covariáveis do indivíduo i. T_i e δ_i representam o tempo observado e os indicadores de censura para o indivíduo i, respectivamente.

Supondo que X seja uma característica específica do vetor X_i, a divisão ocorre ao dividir o nó h em nós filhos à esquerda e à direita, denotados por E e D, respectivamente, utilizando a condição $X \leq c$ e $X > c$. Temos $t_1, t_2, \dots, t_m$ como os tempos de morte distintos. Além disso, denotamos $d_{j,E}, d_{j,D}$ e $Y_{j,E}, Y_{j,D}$ como o número de mortes e o número de indivíduos em risco no tempo t_j nos nós filhos E e D, respectivamente. O termo “em risco” refere-se aos indivíduos que continuam vivos no tempo t_j ou que experimentam o evento de interesse nesse tempo. Temos então:

$$Y_{j,E} = \{T_i \geq t_j, X_i \leq c\}$$

$$e$$

$$Y_{j,D} = \{T_i \geq t_j, X_i > c\}$$

Definindo

$$Y_j = Y_{j,E} + Y_{j,D} \;\; e \;\; d_j = d_{j,E} + d_{j,D}$$

A estatística de divisão log-rank para a divisão $E = \{X_i \leq c\}$ e $D = \{X_i > c\}$ é:

$$L(X,c) = \frac{\sum_{j=1}^{m}(d_{j,E} - Y_{j,E}\frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,E}}{Y_j}\left(1-\frac{Y_{j,E}}{Y_j}\right)\left(\frac{Y_j - d_j}{Y_j - 1}\right)d_j}}$$

Onde $|L(X,c)|$ é uma medida de separação de nós. Quanto maior o valor, maior será a diferença de sobrevivência entre os novos subgrupos E e R, desse modo, quanto maior o valor, melhor será a divisão. A melhor divisão é determinada achando a característica X^* e o valor de divisão c^* que maximizam esse valor, ou seja:

$$|L(X^*,c^*)| \geq |L(X,c)|, \text{ para todo X e c}$$

2.3.4 Random Survival Forests

O *Random Survival Forest* (RSF) é um método de aprendizado de máquina não paramétrico projetado para análise de dados de sobrevivência com censura. Desenvolvido como uma extensão das florestas aleatórias tradicionais, o RSF se destaca por sua capacidade de lidar com dados complexos sem exigir suposições restritivas sobre a distribuição dos tempos de sobrevivência ou a proporcionalidade dos riscos, características que limitam modelos tradicionais como a regressão de Cox (ISHWARAN et al., 2008).

Essa flexibilidade torna o RSF particularmente útil em estudos clínicos, onde os dados frequentemente apresentam relações não lineares entre variáveis e padrões complexos de censura.

O algoritmo do RSF opera através de um processo iterativo que combina técnicas de ensemble learning com métodos específicos para análise de sobrevivência. Inicialmente, são geradas múltiplas árvores de sobrevivência, cada uma construída a partir de uma amostra *bootstrap* do conjunto de dados original. Cerca de um terço das observações não são incluídas em cada amostra *bootstrap*, formando o conjunto *out-of-bag* (OOB), que é utilizado posteriormente para validação interna do modelo (ISHWARAN et al., 2008).

Durante a construção de cada árvore, em cada nó é selecionado aleatoriamente um subconjunto de variáveis preditoras, o que reduz a correlação entre as árvores e

melhora a generalização do modelo. A divisão dos nós é guiada pelo critério baseado no teste log-rank, adaptado para maximizar a diferença entre as curvas de sobrevivência dos subgrupos gerados (Ishwaran & Lu, 2018).

Nos nós terminais de cada árvore, são armazenadas estimativas não paramétricas da função de sobrevivência, usando o estimador de *Kaplan-Meier*, e da função de risco acumulada, estimada via *Nelson-Aalen*. Quando uma nova observação precisa ser avaliada, ela é percorrida em cada árvore até um nó terminal, e as estimativas são combinadas por média para gerar previsões robustas (ISHWARAN et al., 2008).

A avaliação do desempenho do RSF é comumente feita por meio de duas métricas: o índice de concordância, *C-index*, que mede a capacidade do modelo de ordenar corretamente os tempos de falha (Harrell et al., 1982), e o *Brier Score*, que avalia o erro quadrático médio entre as probabilidades previstas e os eventos observados ao longo do tempo (Gerds & Schumacher, 2006).

O algoritmo é dado como se segue: (ISHWARAN et al., 2008).

Passo 1. Obtenha B amostras *bootstrap* e cultive uma árvore para cada amostra criada. Em cada nó será selecionado aleatoriamente as covariáveis de divisão, com intuito de maximizar a diferença de sobrevivência entre nós-filhas.

Passo 2. Cresça a árvore até o tamanho máximo com a restrição de que um nó terminal não deve ter um tamanho menor do que nó de mortes únicas.

Passo 3. Calcule a função de risco acumulado para cada árvore que é estimada pelo estimador Nelson-Aalen e a função de sobrevivência pelo estimador de Kaplan-Meier. Calculando a média dessas estimativas é possível estimar a função de risco acumulado de uma nova observação x. Encontrando o conjunto de funções de riscos acumulados.

Passo 4. Usando dados de teste, calcular a taxa de erro de previsão para os estimadores encontrados no passo anterior.

O algoritmo *Random Survival Forests* estima tanto a função de risco acumulado quanto a função de sobrevivência e o critério de divisão deve ser escolhido.

2.3.5 Estimadores

Ao alcançarem os nós terminais, as árvores cultivadas individualmente de amostras diferentes, serão utilizadas para a estimação da função de sobrevivência e da função de risco acumulado. Podemos estimar a função de risco acumulada, pelo estimador de Nelson-Aalen e a função de sobrevivência, pelo estimador de Kaplan-Meier onde (ISHWARAN & LU ,2018).

Os preditores da árvore de sobrevivência são definidos por atribuir a todos os casos no nó terminal os mesmos estimadores. Isso acontece porque o propósito da árvore de sobrevivência é particionar os dados em grupos homogêneos de indivíduos, com comportamento de sobrevivência semelhante (ISHWARAN et al., 2008).

Sendo x_i um vetor com todas as covariáveis do indivíduo i, para estimar $H(x_i)\ e\ S(x_i)$ para uma nova observação i, inserimos x_i na árvore treinada. Devido à natureza de uma árvore de sobrevivência, x_i chegará a um nó terminal **h** único. Como todos os casos de um nó terminal tem os mesmos estimadores da função de sobrevivência e da função de risco acumulado, o estimador de $H(x_i)\ e\ de\ S(x_i)$ será o estimador do nó terminal **h**, portanto:

$$\widehat{H}(x_i) = \widehat{H_h}(t), \text{ se } x_i \in h$$
$$\hat{S}(x_i) = \widehat{S_h}(t), \text{ se } x_i \in h$$

Nota-se que os estimadores foram derivados de apenas uma árvore. Para o conjunto de árvores, devemos calcular a média das estimativas de **B** árvores contidas na floresta.

Cada árvore da floresta foi crescida utilizando uma amostra **b** própria, desse modo, ao introduzirmos um indivíduo x_i contido nos dados, com objetivo de realizar o teste do modelo, pode ocorrer que a observação tenha sido utilizada para treinar algumas das árvores da floresta e, em outros casos, não tenha sido utilizada.

Sendo assim, podemos gerar estimativas conjuntas excluindo as árvores em que x_i participou do treinamento. Definimos $I_{i,b} = 1$ se i não participou do treinamento e, $I_{i,b} = 0$ se i está presente na amostra utilizada para o crescimento da árvore. Sendo $H_b(x_i)$ e $S_b(x_i)$ o estimador Nelson-Aalen da função de risco acumulada e o estimador de Kaplan-Meier da função de sobrevivência, da árvore **b**, crescidas a partir da b-ésima amostra bootstrap. Temos que o estimador da função de risco acumulado conjunto, que será utilizado para validação do modelo, para a observação i, é dada por:

$$H_c^*(x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b(x_i)}{\sum_{b=1}^{B} I_{i,b}}, i = 1, \ldots, n$$

Por sua vez, o estimador da função de sobrevivência do conjunto de árvores, para testes, é dado por:

$$S_c^*(x_i) = \frac{\sum_{b=1}^{B} I_{i,b} S_b(x_i)}{\sum_{b=1}^{B} I_{i,b}}$$

Os estimadores são uteis para inferência nos dados de treino e para estimar erro de predição, e, só são aplicáveis quando a observação x_i está contida nos dados originais.

Por outro lado, os estimadores conjuntos da função de sobrevivência e da função de risco acumulado:

$$H_c^{**}(x) = \frac{1}{B} \sum_{b=1}^{B} H_b(x) \text{ , e}$$

$$S_c^{**}(x) = \frac{1}{B} \sum_{b=1}^{B} S_b(x)$$

São utilizados para predição e podem ser aplicados para qualquer novo paciente, por exemplo, se quisermos estimar as funções de sobrevivência e de risco acumulado de um novo paciente, inserimos ele em uma floresta, em cada árvore ele termina em um nó terminal diferente, devido a natureza das árvores de sobrevivência, todos os pacientes que acabam em um nó terminal recebem as mesmas estimativas $\widehat{H}(x_i)\ e\ \hat{S}(x_i)$, ou seja, em cada nó terminal é estimado a função de sobrevivência e a função de risco acumulado com base nos tempos de sobrevivência observados dos pacientes que chegam nesse nó. A partir da média entre as **B** estimativas calculadas, temos os estimadores conjuntos.

2.3.6 Estatísticas de desempenho

O C-Index e o Brier Score são métricas comumente utilizadas na avaliação de modelos de Random Survival Forest, o C-Index avalia a capacidade de classificação do modelo RSF, enquanto o Brier Score mede a precisão e calibração das probabilidades de falha previstas pelo modelo.

2.3.7 Mortalidade

Para calcular o C-Index, é necessário definir o que constitui um resultado predito pior. Para os modelos de sobrevivência, definimos como mortalidade, que é o valor predito usado pelo Random Survival Forest. Denotando $t_1 < \cdots < t_m$ como todo o conjunto de tempos de evento únicos para os dados de aprendizado. Para compararmos dois casos, $i\ e\ j$ com vetores de covariáveis $X_i\ e\ X_j$, dizemos que i tem um resultado previsto pior que j, se:

$$\sum_{l=1}^{m} H_c^*(x_i) > \sum_{l=1}^{m} H_c^*(x_j)$$

Os lados direito e esquerdo da equação denotam a mortalidade para i e j respectivamente, que é o número de mortes esperadas se todos os indivíduos tivessem as mesmas características de $X_i\ e\ X_j$. (ISHWARAN & LU ,2018)

2.3.8 C-Index

O C-Index, também conhecido como concordance index ou índice de concordância de Harrell (HARRELL et al., 1982), é uma medida de desempenho que avalia a capacidade de classificação do modelo Random Survival Forest. Ele mede a proporção de pares de amostras em que o tempo de falha real é previsto corretamente, ou seja, a proporção de vezes em que o paciente que falhou primeiro é classificado corretamente em relação a um paciente que ainda está sob risco de falha.

Para calcularmos o C-Index, devemos seguir os seguintes passos, como descrito por Ishwaran et al. (2008):

Passo 1. Formar todos os pares de indivíduos possíveis a partir dos dados.

Passo 2. Omitir os pares cujo tempo de sobrevivência menor é censurado. Omitir os pares $i\ e\ j$ se $T_i = T_j$ a menos que um deles seja uma morte. Denotamos como admissível o total de pares que não foram omitidos.

Passo 3. Para cada um dos pares admitidos, onde $T_i \neq T_j$, conte 1 se o tempo de sobrevivência mais curto tiver um pior resultado predito; conte 0,5 se os resultados preditos forem iguais. Para cada par admitido, onde $T_i = T_j$ e ambos são óbitos, conte 1 se os resultados preditos forem iguais. Para cada par admitido, onde $T_i = T_j$, mas apenas

um dos dois é uma morte, conte 1 se a morte tem um resultado predito pior, de outra forma, conte 0,5. Seja denotado Concordância como a soma de todos os pares admitidos. **Passo 4**. O C-Index, C, é definido como:

$$C = \frac{Concordância}{Admissível}$$

A taxa de erro de estimação é $PE = 1 - C$. Note que $0 \leq PE \leq 1$ e que $PE = 0{,}5$ significa que o modelo não acerta mais que uma adivinhação aleatória e, $PE = 0$ indica previsões perfeitas.

2.3.9 Brier Escore Integrado

Seja $BS(t)$ para o tempo t definido por, (Gerds & Schumacher, 2006)

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\left(0 - \hat{S}(x_i)\right)^2}{\hat{G}(t_i)} I(T_i \leq t, \sigma_i = 1) + \frac{\left(1 - \hat{S}(x_i)\right)^2}{\hat{G}(t)} I(T_i > t)\right\}$$

.

onde, N representa o número total de observações no conjunto de dados, a soma é realizada para cada observação i, onde i varia de 1 a N e, $\hat{G}(t) = P(C_i > t)$, onde, o estimador da função de sobrevivência censorizada $\hat{G}(t_i)$ no tempo t_i pode ser obtido considerando apenas as observações censuradas até o tempo t_i. Isso pode ser calculado usando a função de sobrevivência de Kaplan-Meier considerando apenas as observações censuradas.

O estimador da função de sobrevivência censurada $\hat{G}(t)$ no tempo t, por sua vez, é obtido considerando todas as observações censuradas até o tempo t. Esse estimador também pode ser calculado usando a função de sobrevivência de Kaplan-Meier, levando em consideração todas as observações censuradas até o tempo t.

Dentro do somatório temos duas partes, na primeira $\frac{\left(0-\hat{S}(x_i)\right)^2}{\hat{G}(t_i)} I(T_i \leq t, \sigma_i = 1)$, $\hat{S}(x_i)$ é o estimador da função de sobrevivência para o tempo t, para a observação i, $\hat{G}(t_i)$ é o estimador da função de sobrevivência censorizada nos tempos t_i, e, por fim, $I(T_i \leq t, \sigma_i = 1)$ é uma função indicadora que é igual a 1 se a observação i não está censurada no tempo t, ou seja, $\sigma_i = 1$, e se $T_i \leq t$, 0, caso contrário. Na segunda parte $\frac{\left(1-\hat{S}(x_i)\right)^2}{\hat{G}(t)} I(T_i > t)$

temos, novamente $\hat{S}(x_i)$ o estimador da função de sobrevivência no tempo t, para a observação i, e $\hat{G}(t)$ é o estimador da função de sobrevivência censurada para o tempo t, $I(T_i > t)$ uma função indicadora que é igual a 1 se $(T_i > t)$ e 0 caso contrário.

A curva de erro de previsão é obtida a partir do Brier Escore, através dos tempos. Desse modo, o Brier Escore Integrado é a curva do erro de previsão acumulado sobre o tempo, e é dada por,

$$IBS = \frac{1}{(t_i)} \int_0^{(t_i)} BS(t)dt.$$

Valores de IBS próximos de 1 indicam baixa capacidade preditiva do modelo, enquanto valores próximos de 0 refletem bom desempenho.

2.3.10 Importância das variáveis

O RSF permite calcular a importância das variáveis (*VIMP*) com base no aumento do erro de predição ao permutar os valores de cada variável de forma aleatória. Esse recurso oferece uma medida útil para identificar fatores com maior impacto na sobrevida dos indivíduos (ISHWARAN et al., 2008).

3 METODOLOGIA

O presente estudo utilizou dados de 942 pacientes diagnosticados com câncer de estômago nos estágios I, II e III no estado de São Paulo, em 2018, e acompanhados até dezembro de 2024. As informações foram obtidas gratuitamente através da base de dados da Fundação Oncocentro de São Paulo (2025).

A base original continha 104 variáveis, mas para a análise, foram mantidas apenas oito: Sexo, Morfologia do Câncer, Grupo de Estadiamento Clínico, Tratamentos Recebidos, Faixa Etária, Categoria de Atendimento, Tempo de Seguimento (em dias) e Status. O Tempo de Seguimento foi calculado desde a data do diagnóstico até a última informação registrada, que foi usada para classificar o desfecho: Evento de Interesse (óbito por câncer) ou Censura (pacientes vivos ou falecidos por outras causas).

Para garantir a homogeneidade da amostra, foram excluídos pacientes no Estágio IV, devido à alta mortalidade, aqueles que não receberam nenhum tratamento e pacientes com Grupo de Estadiamento Clínico X ou Y, pois esta classificação indica a impossibilidade de determinar a extensão da doença.

Algumas covariáveis foram recategorizadas para otimizar a análise e atender aos pressupostos de riscos proporcionais: a Morfologia foi resumida nos três tipos mais comuns e "Outro"; os nove tipos originais de Tratamentos foram dicotomizados em “Tratamentos Cirúrgicos” e “Tratamentos Não Cirúrgicos”; e a Faixa Etária, originalmente em intervalos de 10 anos, foi condensada em três grupos: “10 a 49 anos”, “50 a 69 anos” e “70 ou mais”. As variáveis Categorias sobre “Atendimento” e “Sexo” foram mantidas em suas classificações originais.

Às análises e gráficos foram feitos no software estatístico R. Para a aplicação dos algoritmos de Random Survival Forests (RSF), utilizou-se o pacote randomForestSRC (Ishwaran; Kogalur, 2025), enquanto o ajuste do Modelo de Riscos Proporcionais de Cox foi realizado através do pacote survival (Therneau, 2024).

4 DISCUSSÃO E RESULTADOS

4.1 Análise descritiva

Esta seção tem como objetivo apresentar uma análise descritiva das características dos pacientes incluídos no estudo. A Tabela 1 detalha as frequências absolutas e relativas dos pacientes em cada categoria das covariáveis, permitindo uma melhor compreensão do perfil da amostra. Além disso, a Tabela 1 também exibe a distribuição de mortes e dos casos de censura em cada um dos grupos, fornecendo o panorama inicial para a análise de sobrevivência.

Tabela 1 – Frequência das variáveis e relação de falha / censura

COVARIÁVEL	CATEGORIA	FREQUÊNCIA (%)	FALHA / CENSURA (%)
SEXO	Masculino	576 (61,1%)	201 (34,9%) / 375 (65,1%)
	Feminino	366 (38,9%)	103 (28,1%) / 263 (71,9%)
CATEATEND	Convênio	99 (10,5%)	15 (15.2%) / 84 (84.8%)
	SUS	843 (89,5%)	289 (34.3%) / 554 (65.7%)
ECGRUP	I	274 (29,1%)	32 (11,7%) / 242 (88,3%)
	II	271 (28,8%)	84 (31,0%) / 187 (69,0%)
	III	397 (42,1%)	188 (47,4%) / 209 (52,6%)
TRATAMENTO	Não Cirúrgicos	227 (24,1%)	98 (43,2%) / 129 (56,8%)
	Cirúrgicos	715 (75,9%)	206 (28,8%) / 509 (71,2%)
FAIXAETAR	10-49	144 (15,3%)	33 (22,9%) / 111 (77,1%)
	50-69	508 (53,9%)	160 (31,5%) / 348 (68,5%)
	70+	290 (30,8%)	111 (38,3%) / 179 (61,7%)
MORFO	ADENOCARCINOMA SOE	315 (33,4%)	117 (37,1%) / 198 (62,9%)
	CARCINOMA DE CELULAS EM ANEL DE SINETE	171 (18,2%)	73 (42,7%) / 98 (57,3%)
	ADENOCARCINOMA TUBULAR	217 (23%)	70 (32,3%) / 147 (67,7%)
	Outros	239 (25,4%)	44 (18,4%) / 195 (81,6%)
TOTAL		942 (100%)	304 (32,3%) / 638 (67,7%)

Fonte: Elaborado pelo autor, 2025.

A análise da Tabela 1, nos ajuda a entender o perfil clínico e demográfico dos pacientes. O grupo masculino representa a maioria dos casos, sendo 576 pacientes, ou 61,2% do total, enquanto o grupo feminino é composto por 366 pacientes. Em relação à mortalidade, a taxa é ligeiramente maior entre os homens, cerca de 34,9%, em comparação com os 28,1% das mulheres.

Grande parte dos pacientes recebeu atendimento pelo SUS, 843 pacientes, ou 89,5%. A mortalidade é significativamente mais alta no grupo do SUS, sendo de 34,3% em comparação com o grupo que recebeu atendimento em convênio de 15,2%. Quanto ao estadiamento clínico (ECGRUP), o Estágio III é o mais frequente, sendo constatado em 397 pacientes, seguido pelos Estágios I e II, ambos com cerca de 270 pacientes. A

taxa de mortalidade aumenta progressivamente com o avanço do estágio: apenas 11,7% no Estágio I, saltando para 31,0% no Estágio II e atingindo o pico de 47,4% no Estágio III.

A maioria dos pacientes foi submetida a Tratamentos Cirúrgicos, que apresentaram uma mortalidade de 28,8%. Os pacientes em Tratamentos Não Cirúrgicos apresentaram uma mortalidade consideravelmente mais alta, de 43,2%. Em relação à Faixa Etária, a maioria dos pacientes concentra-se no grupo de 50-69 anos. A mortalidade aumenta com a idade, sendo mais baixa no grupo de 10-49 anos 22,9% e atingindo o nível mais alto no grupo 70+ 38,3%.

Por fim, a morfologia mais comum é o ADENOCARCINOMA SOE (315 casos), com mortalidade de 37,1%. O CARCINOMA DE CÉLULAS EM ANEL DE SINETE (171 casos) apresentou a maior mortalidade (42,7%), e os casos classificados como "Outros" foram os que tiveram a menor taxa de mortalidade (18,4%). 32,3% dos pacientes com câncer gástrico em estágio I, II e III chegaram a óbito nesse estudo.

Tabela 2 – Distribuição da variável Tempo em dias

Variável	Mínimo	1ºQ	Mediana	Média	3ºQ	Máximo
Tempo	1,00	406,00	960,50	1096,20	1875,80	2500,00

Fonte: Elaborado pelo autor, 2025.

Com base na Tabela 2, que detalha o Tempo de Seguimento dos pacientes em dias, a principal medida observada é a mediana de 960,50 dias. Esse valor indica que metade dos pacientes incluídos no estudo chegou à morte ou censura em um período de aproximadamente 2 anos e meio.

O acompanhamento dos pacientes variou significativamente, começando com um tempo mínimo de apenas 1 dia e estendendo-se até um máximo de 2.500 dias. A diferença entre a média, de 1.096,20 dias e a mediana nos mostra que a distribuição do tempo é ligeiramente assimétrica, com alguns pacientes sendo acompanhados por períodos consideravelmente mais longos.

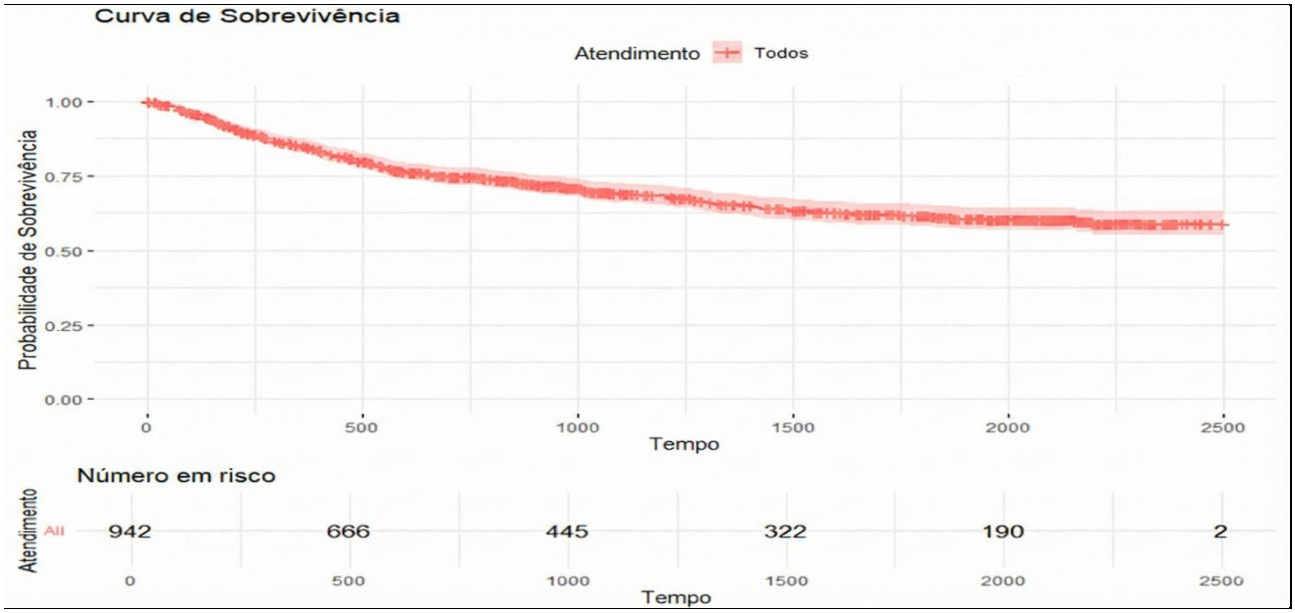
Na próxima seção, entenderemos melhor a relação da mortalidade com o tempo, a partir da comparação das curvas de Kaplan- Meier.

4.2 Análise das curvas de sobrevivência

Nesta seção foram apresentadas as curvas de sobrevivência ajustadas pelo método não paramétrico de *Kaplan-Meier* para cada uma das variáveis do estudo, bem

como a comparação destas pelo teste *log-rank*, a fim de tentar entender melhor a sobrevivência destes pacientes e analisar se as covariáveis entrarão no modelo.

Figura 1 – Curva de sobrevivência geral



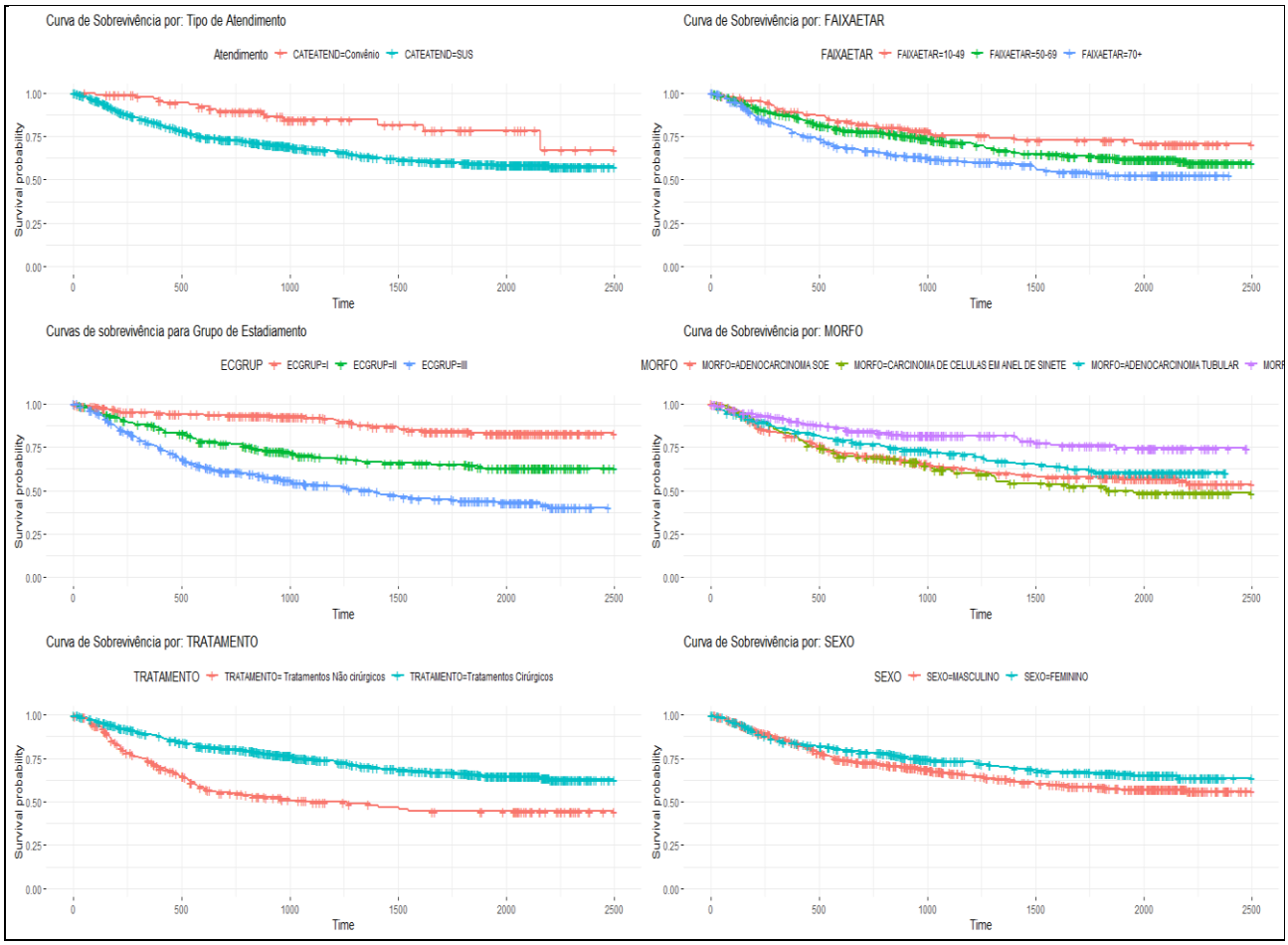
Fonte: Elaborado pelo autor, 2025.

A Figura 1 apresenta a curva de sobrevivência estimada pelo método de Kaplan-Meier para a amostra total de pacientes. A curva demonstra que a probabilidade de sobrevida para os pacientes do estudo é relativamente alta a longo prazo, alcançando aproximadamente 60% ao final do seguimento de 2.500 dias. No entanto, a curva evidencia uma mortalidade concentrada e precoce: a probabilidade de sobrevida decresce rapidamente, de 100% para cerca de 75% nos primeiros 500 dias, e então permanece mais estável até o fim do estudo. Esse padrão pode ser observado na figura 2, nas curvas de sobrevivência estratificadas pelo estágio do câncer, podemos observar que pacientes em estágio III morrem mais rápido que pacientes nos outros estágios, sendo o estágio 1 o menos letal.

A Figura 2 exibe as curvas de sobrevivência de Kaplan-Meier estratificadas por todas as covariáveis do estudo. Visualmente, é possível identificar uma clara separação entre as curvas em todos os gráficos, mas, para confirmar, foi utilizado o teste Log-rank, apresentado na Tabela 3. Adicionalmente, observa-se que as curvas, em sua maioria, não se cruzam ao longo do tempo indicando a manutenção do pressuposto de riscos proporcionais entre os grupos, o que é um requisito essencial para a aplicação do Modelo de Regressão de Cox na análise multivariada subsequente.

Além disso, as maiores diferença de sobrevivência entre os grupos, parecem ser entre os grupos de estadiamento e entre o tipo de tratamento, a variável que apresenta graficamente a menor diferença é o sexo do participante, apresentando uma sobrevivência semelhante.

Figura 2 –Curvas de Sobrevivência de Kaplan-Meier para as covariáveis do estudo.



Fonte: Elaborado pelo autor, 2025.

Realizando o teste Log-rank para comparação de curvas com todas as variáveis do estudo, sendo utilizado um nível de significância α = 5%, verifica-se na Tabela 3 que, todas as variáveis possuem pelo menos uma curva que difere estatisticamente entre a(s) outra(s).

Tabela 3 –Teste Log-rank com todas as variáveis do estudo

VARIÁVEL	ESTATÍSTICA LOG-RANK	P-VALOR
CATEATEND	10,9	$p = 0,001$
ECGRUP	97,6	$p < 0,0001$
TRATAMENTO	41,9	$p < 0,0001$
FAIXAETAR	13,9	$p = 0,001$
MORFO	22,4	$p < 0,0001$
SEXO	3,8	$p = 0,05$

Fonte: Elaborado pelo autor, 2025.

4.3 Modelo de Regressão de Cox

Com base na análise das curvas de *Kaplan-Meier* (Figura 2) e nos resultados do teste *Log-Rank* (Tabela 3), que demonstraram diferença estatística significativa na sobrevida entre os grupos de todas as covariáveis, a próxima etapa é a construção de um Modelo de Regressão de Riscos Proporcionais de Cox.

A análise visual prévia das curvas de sobrevivência sugeriu que o pressuposto de riscos proporcionais é satisfeito para todas as covariáveis. Para confirmar essa premissa e garantir a validade do modelo de Cox, um teste formal foi realizado após o ajuste do modelo.

Foi empregado um método de seleção de variáveis (COLLETT,1994). Este procedimento visa identificar o melhor subconjunto de variáveis que deve ser incluído no modelo de Cox.

Na Tabela 4, descrevemos o passo a passo da seleção de variáveis até o melhor modelo possível, para facilitar, trataremos as variáveis como ECGRUP = V1, TRATAMENTO = V2, FAIXAETAR = V3, MORFO = V4, SEXO = V5 e CATEATEND = V6. No passo 1, comparamos os modelos de Cox ajustados para cada uma das covariáveis com o modelo nulo, todas as covariáveis se mostraram significativas em $\alpha = 5\%$. No passo 2, ajustamos um modelo com todas as covariáveis, retiramos uma a uma e comparamos, no passo 2 a covariável SEXO se mostrou não significativa no modelo conjunto. No passo 3, ajustamos um modelo sem a covariável SEXO e retiramos uma a uma, todas as covariáveis foram significativas para $\alpha = 5\%$. Sendo assim o modelo que seguiremos será ajustado com as covariáveis:

ECGRUP+TRATAMENTO+FAIXAETAR+MORFO+CATEATEND.

Tabela 4 – Seleção de variáveis usando o modelo de regressão de Cox

PASSOS	MODELO	-2LOG L	ESTATÍSTICA	VALOR- P
PASSO 1	NULO	3918,20		
	V1	3812,60	105,58	P < 0,001
	V2	3882,40	35,883	P < 0,001
	V3	3904,60	13,716	P = 0,001
	V4	3894,00	24,221	P < 0,001
	V5	3914,40	3,9014	P = 0,048
	V6	3905,00	13,355	P < 0,001
PASSO 2	V1+V2+V3+V4+V5+V6	3741,40		
	V1+V2+V3+V4+V5	3748,40	7,051	P = 0,008
	V1+V2+V3+V4+V6	3741,40	0,1293	P = 0,719
	V1+V2+V3+V5+V6	3756,40	15,062	P = 0,002
	V1+V2+V4+V5+V6	3755,00	13,647	P = 0,001
	V1+V3+V4+V5+V6	3774,20	32,829	P < 0,001
	V2+V3+V4+V5+V6	3831,00	89,633	P < 0,001
PASSO 3	V1+V2+V3+V4+V6	3741,40		
	V2+V3+V4+V6	3832,60	91,268	P < 0,001
	V1+V3+V4+V6	3774,60	33,131	P < 0,001
	V1+V2+V4+V6	3755,00	14,064	P < 0,001
	V1+V2+V3+V6	3756,60	15,145	P = 0,001
	V1+V2+V3+V4	3748,40	7,0293	P = 0,008
MODELO FINAL	V1+V2+V3+V4+V6	3741,40		

Fonte: Elaborado pelo autor, 2025.

4.4 Adequação do modelo

Nesta seção verificamos a adequação do modelo de Cox final, a partir de técnicas gráficas, bem como a utilização do teste de resíduos padronizados de Schoenfeld, a fim de avaliar a suposição dos riscos proporcionais ao longo do tempo.

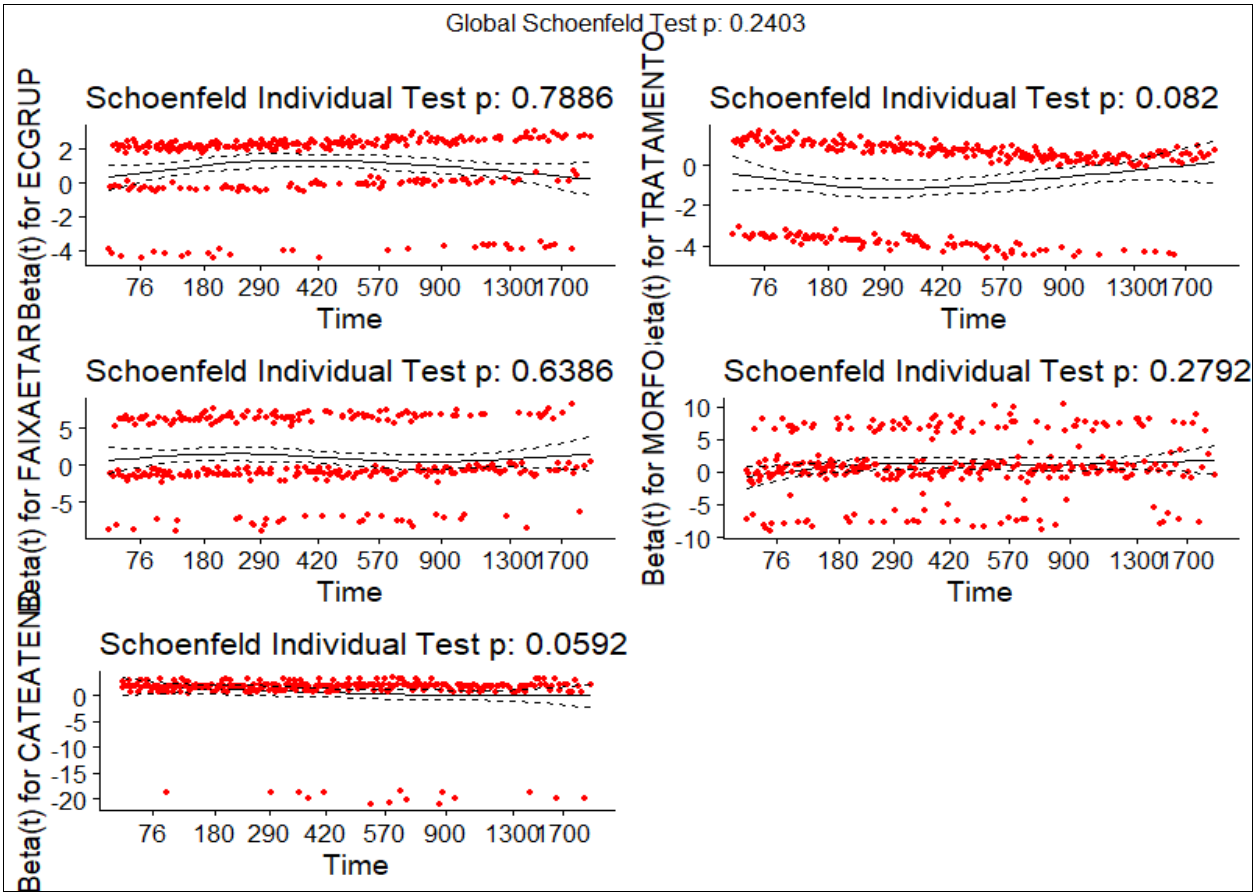
A validade do modelo de Cox foi verificada através da análise dos resíduos de Schoenfeld na Tabela 5 e na Figura 3. O teste global não indicou violação da suposição de riscos proporcionais, com p-valor = 0,2405. Da mesma forma, todas as covariáveis apresentaram comportamento satisfatório individualmente, com p-valor > 0,05, confirmando que os efeitos das variáveis sobre o risco de óbito se mantêm constantes ao longo do tempo de acompanhamento.

Tabela 5 –Testes da proporcionalidade dos riscos do modelo ajustado

VARIÁVEL	CHISQ	P-VALOR
ECGRUP	0,475	0,789
TRATAMENTO	3,025	0,082
FAIXAETAR	0,897	0,639
MORFO	3,841	0,279
CATEATEND	3,559	0,059
GLOBAL	11,543	0,240

Fonte: Elaborado pelo autor, 2025.

Figura 3 – Resíduos de Schoenfeld



Fonte: Elaborado pelo autor, 2025.

A partir da verificação da violação dos riscos proporcionais, é possível seguir para a interpretação dos coeficientes, das razões de risco e das análises de sobrevivências. Na Tabela 6, descrevemos os resultados do modelo de riscos proporcionais de Cox, as estimativas das Razões de Risco (RR) e seus respectivos intervalos de confiança. Observou-se que o estadiamento avançado foi o maior fator de risco. Pacientes classificados no Estágio III apresentaram um risco de óbito 4,78 vezes maior em relação ao Estágio I (IC 95%: 3,27 – 6,97; p < 0,001), enquanto o Estágio II apresentou um risco 2,65 vezes superior (IC 95%: 1,76 – 3,98; p < 0,001). A realização de tratamentos

cirúrgicos foi associada a uma redução de 54% no risco de mortalidade (IC 95%: 0,36 – 0,60; p < 0,001) quando comparada à tratamentos não cirúrgicos.

Pacientes idosos (70 anos ou mais) tiveram um risco de óbito 95% maior comparados aos pacientes com menos de 50 anos. Além disso, o atendimento via SUS esteve associado a um aumento de 92% no risco de óbito em comparação ao convênio (RR = 1,92; IC 95%: 1,13 – 3,26; p = 0,016). Quanto à morfologia, o Carcinoma de Células em Anel de Sinete eleva o risco de óbito em 39% (RR = 1,39; p = 0,032) em relação ao Adenocarcinoma SOE.

Tabela 6 – Resultados da análise multivariada de Cox para a sobrevida global dos pacientes.

Variável / Categoria	RR	IC 95%	P-valor
Estadiamento (ECGRUP)			
Estágio I (Referência)	1,00	—	—
Estágio II	2,65	1,76– 3,98	< 0,001
Estágio III	4,78	3,27 – 6,97	< 0,001
Tratamento			
Não Cirúrgico	1,00	—	—
Tratamentos Cirúrgicos	0,46	0,36 – 0,60	< 0,001
Faixa Etária			
< 50 anos (Referência)	1,00	—	—
50 - 69 anos	1,36	0,93 – 2,00	0,112
70 anos ou mais	1,95	1,31 – 2,89	< 0,001
Morfologia			
Adenocarcinoma SOE	1,00	—	—
Carcinoma de Células em Anel de Sinete	1,39	1,03 – 1,88	0,032
Adenocarcinoma Tubular	0,95	0,70 – 1,28	0,715
Outros	0,66	0,46 – 0,94	0,023
Categoria de Atendimento			
Particular / Convênio (Ref.)	1,00	—	—
SUS	1,92	1,13 – 3,26	0,016

Fonte: Elaborado pelo autor, 2025.

O grupo de pacientes entre 50 e 69 anos não mostrou uma diferença significativa em sobrevivência, quando comparado aos pacientes com menos de 50 anos, além disso, pacientes com a morfologia Adenocarcinoma Circular, não representaram diferença significativa de sobrevivência com relação ao Adenocarcinoma SOE, isso mostra a semelhança nesses dois tipos morfológicos de câncer.

4.5 Random Survival Forest

Nesta seção, foram apresentados dois modelos de *Random Survival Forests* (RSF). O primeiro utilizou as variáveis em sua forma original, conforme coletadas. Já o segundo utilizou as mesmas covariáveis do modelo de Cox final, mantendo as estratificações realizadas. O objetivo foi verificar se houve perda significativa na predição devido à forma como estratificamos as covariáveis para atender aos pressupostos de riscos proporcionais. Como comumente feito em modelos de *Machine Learning*, os dados foram separados em 70% para o conjunto de treino e 30% para o conjunto de testes.

Tabela 7 – RSF com a base de dados original (modelo 1) e RSF com as variáveis recategorizadas (modelo 2)

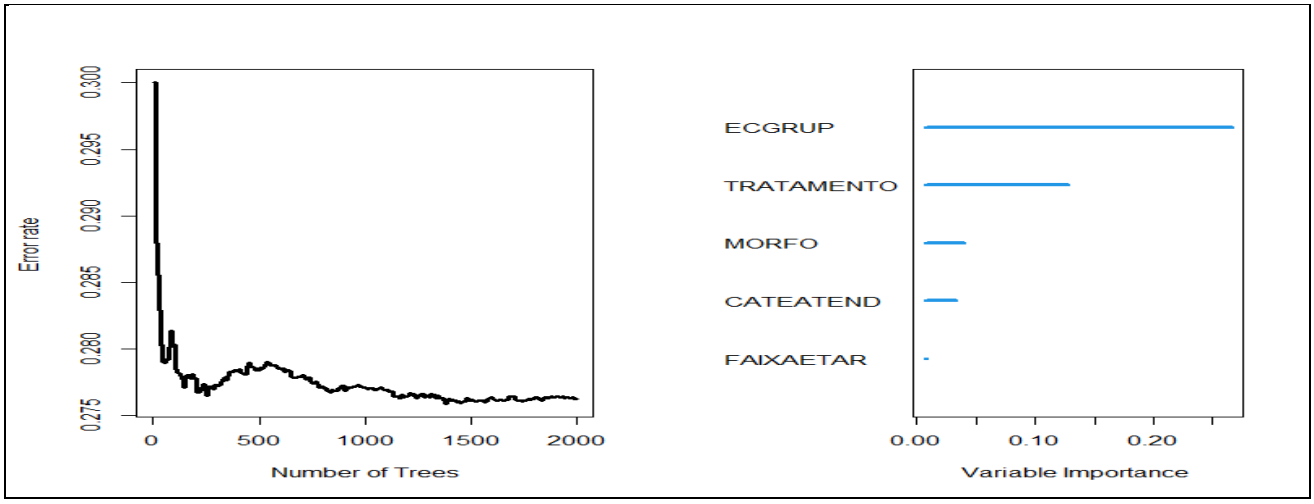
Métrica de Performance	**Modelo 1**	**Modelo 2**
Erro de Predição (OOB)	0,276 (27,6%)	0,326 (32,6%)
C-Index Estimado (1 - Erro)	0,724	0,674
IBS (OOB)	15,85%	17,24%
Nós Terminais (Média)	28,47	25,28
N (Amostra / Óbitos)	659 / 218	659 / 218

Fonte: Elaborado pelo autor, 2025.

A Tabela 7 apresentou a diferença entre Modelo 1 e o Modelo 2, mantendo-se constante o tamanho amostral (n=659). Podemos observar que o modelo com os dados originais performou melhor em todas as métricas, indicando que houve perda de informação na estratificação dos dados. O erro de predição *Out-of-Bag* foi reduzido de 32,6%, para 27,6%, o que resultou em um aumento do C-Index estimado de 0,67 para 0,72. Da mesma forma, o IBS, que avaliou a calibração e precisão das probabilidades de sobrevivência, foi menor no modelo original, indicando um melhor ajuste aos dados. A Figura 4 representou a importância de cada variável no model (VIMP), consideradas pelo

modelo RSF. Podemos observar que todas as covariáveis foram consideradas importantes para o modelo, ECGRUP = 0,2603 representou o maior valor de VIMP. Observa-se que após cerca de 1300 árvores crescidas, o erro de previsão estabilizou.

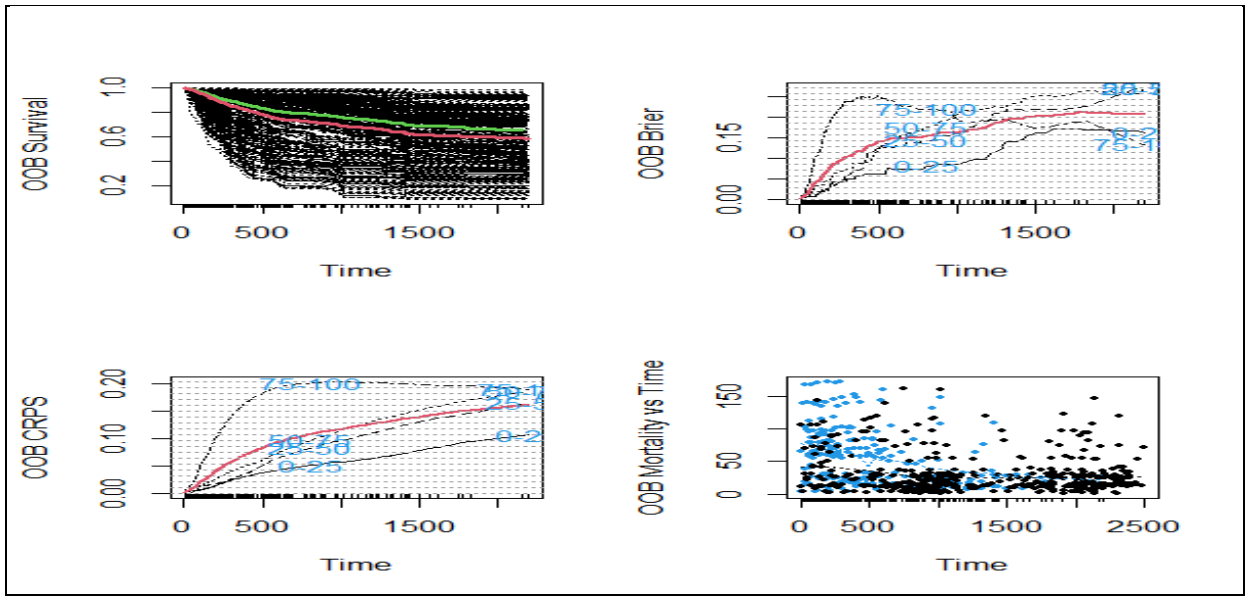
Figura 4 – Relação do erro de previsão e número de árvores e gráfico do VIMP



Fonte: Elaborado pelo autor, 2025.

A Figura 5 apresenta os diagnósticos de desempenho do modelo *Random Survival Forest* baseados nas estimativas *Out-of-Bag*. Inicialmente, o gráfico de sobrevivência (canto superior esquerdo) demonstra a estimativa da floresta da função de sobrevivência para cada indivíduo em linhas pretas, destaca a média geral do modelo pela linha vermelha, que acompanha de perto o estimador clássico, indicado pela linha verde.

Figura 5 – Gráfico referente ao modelo RSF selecionado



Fonte: Elaborado pelo autor, 2025.

Isso nos mostra que o algoritmo capturou corretamente a tendência de queda na sobrevida ao longo do tempo. Em relação à precisão, os gráficos de *Brier Score* e *CRPS* mostram que a taxa de erro aumenta no início do acompanhamento, mas tende a se estabilizar após cerca de 1.000 dias. Por fim, o gráfico de dispersão (canto inferior direito) ilustrou a relação entre a mortalidade prevista e o tempo observado, em que podemos observar uma grande relação de eventos de interesse (pontos azuis) no começo do estudo, entre 0 e 1500 dias, os pontos pretos representam a censura.

Tabela 8 – Predição com os dados de teste, utilizando o modelo RSF selecionado

Métrica de Performance	**Resultados do modelo**
Erro de Predição	0,2554
C-Index Estimado (1 - Erro)	0,7446
IBS	25,78%
Nós Terminais (Média)	26,51

Fonte: Elaborado pelo autor, 2025.

Essas métricas representam o desempenho do modelo RSF ao prever a sobrevivência para novos pacientes, ou seja, para os 30% dos dados reservados exclusivamente para teste.

O erro de predição foi de 0,2554, o que corresponde a um C-Index estimado de 0,7446. Na prática, isso significa que o modelo é capaz de ordenar corretamente cerca de 74% das comparações entre dois pacientes, distinguindo bem quem tem maior risco de óbito ao longo do tempo. Como esse valor vem do conjunto de teste, ele reflete um desempenho consistente e não inflado pelo treinamento.

O *IBS (Integrated Brier Score)* foi de 25,78%, indicando que as curvas de sobrevivência previstas pelo modelo ficaram, em média, relativamente próximas das curvas observadas. Um IBS abaixo de 30% é considerado um resultado bom em estudos de sobrevivência, sugerindo boa calibração das previsões. No conjunto, esses resultados mostram que o RSF apresenta boa discriminação, boa capacidade de generalização e calibração adequada, tornando-se uma abordagem robusta para prever a evolução de novos pacientes.

5 CONCLUSÃO

O presente trabalho alcançou seu objetivo principal ao aplicar o algoritmo Random Survival Forest (RSF) e o modelo de regressão de Cox na análise de sobrevivência de pacientes com câncer gástrico. A utilização de dados reais da Fundação Oncocentro de São Paulo evidenciou os desafios de ajustar modelos clássicos na prática. Contudo, a regressão de Cox permitiu compreender como as variáveis influenciam o tempo de vida dos pacientes, oferecendo uma interpretação clara por meio dos riscos relativos.

Do ponto de vista clínico, a análise descritiva e o Modelo de Cox evidenciaram que o estadiamento avançado é o fator de maior impacto na mortalidade, com o Estágio III apresentando um risco de óbito 4,78 vezes maior que o Estágio I. Além disso, o estudo confirmou a importância da intervenção cirúrgica, que reduziu em 54% o risco de mortalidade. Observou-se também que pacientes atendidos pelo SUS apresentaram um risco de óbito 92% superior aos pacientes de convênio. A análise de importância de variáveis (VIMP) do RSF corroborou estes achados, destacando o Grupo de Estadiamento Clínico (ECGRUP) como a variável de maior poder preditivo.

A construção de um modelo clássico e de um modelo de aprendizado de máquina permitiu entender a vantagem de cada abordagem. O estudo demonstrou que o modelo RSF treinado com as variáveis originais (Modelo 1) obteve desempenho superior ao modelo treinado com variáveis estratificadas (Modelo 2). O Modelo 1 alcançou um C-Index de 0,724 contra 0,674 do Modelo 2, além de apresentar um menor erro de predição (27,6% contra 32,6%). Este resultado indicou que o processo de recategorização e estratificação de variáveis, utilizado para satisfazer o pressuposto de riscos proporcionais, pode acarretar perda de informação relevante, prejudicando a capacidade preditiva. Além disso, o Modelo 1 apresentou um C-Index de 0,74 quando aplicado ao conjunto de testes, confirmando que o RSF ajustado possui um elevado poder de generalização e predição.

Em suma, a escolha entre um modelo clássico, como a regressão de Cox, e um modelo de aprendizado de máquina, como o Random Survival Forest, depende fundamentalmente do objetivo da pesquisa. Neste estudo, observou-se que o modelo de Cox é essencial para entender a relação e a magnitude do efeito das covariáveis sobre o risco de falha ou seja, nos fornece interpretabilidade, enquanto o modelo RSF se destaca quando o objetivo primordial é a maximização da precisão preditiva.

Referências

BREIMAN, L. **Random forests**. Machine Learning, 45: 5-32, 2001.

COLLETT, D. *Modelling Survival Data in Medical Research.* London: Chapman & Hall, 1994.

COLOSIMO, E. A; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. 1.ed. São Paulo, SP: Editora Edgard Blucher, 2006, 367p.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, London, v. 34, n. 2, p. 187-220, 1972.

GERDS, T. A.; SCHUMACHER, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. **Biometrical Journal**, [S. l.], v. 48, n. 6, p. 1029-1040, 2006.

HARRELL, F. E. et al. Evaluating the yield of medical tests. **JAMA**, [S. l.], v. 247, n. 18, p. 2543-2546, 1982.

ISHWARAN, H.; KOGALUR, U.B., BLACKSTONE, E.H., LAUER, M.S. **Random survival forests.** The Annals of Applied Statistics, v.2, n.3, p.841-860, 2008.

ISHWARAN, H.; KOGALUR, U. B**.** ***randomForestSRC: Fast Unified Random Forests for Survival, Regression and Classification (RF-SRC).*** R package version 3.4.3, 2025. Disponível em: https://cran.r-project.org/package=randomForestSRC. Acesso em: 20 nov. 2025.

ISHWARAN, H; LU, Min. **Random Survival Forests.** Wiley StatsRef, 99 1-12, 2018

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina: uma abordagem estatística**. São Carlos, SP: Rafael Izbicki, 2020. 253 p.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, [S. l.], v. 53, n. 282, p. 457-481, 1958.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**, [S. l.], v. 50, n. 3, p. 163-170, 1966.

MOGENSEN, U.B.; ISHWARAN, H.; GERDS, T.A. **Evaluating random forests for survival analysis using prediction error curves.** Journal of statistical software, v.50, n.11, p.1-23, 2012.

MORETTIN, Pedro A.; SINGER, Julio M. **Estatística e Ciência de Dados**. São Paulo, SP: GEN I Grupo Editorial Nacional, 2021. 542 p.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024. Disponível em: https://www.R-project.org/. Acesso em: 20 nov. 2025.

THENEAU, T. M**.** ***survival: A Package for Survival Analysis in R*****.** R package version 3.8-3, 2024. Disponível em: https://cran.r-project.org/package=survival. Acesso em: 20 nov. 2025.

SUNG, H; FEARLEY, J; SIEGEL, RL; LAVERSANNE, M; SOERJOMATARAM, I; JEMAL, A; BRAY, F. **Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin*. 2021: 71: 209-249. https://doi.org/10.3322/caac.21660

WANG, H.; LI, G. **A Selective Review on Random Survival Forests for High Dimensional Data**. Quant Biosci. 2017;36(2):85-96. doi: 10.22283/qbs.2017.36.2.85. PMID: 30740388; PMCID: PMC6364686.

WANG, P.; LI, Y.; REDDY, C. K. **Machine learning for survival analysis**. ACM Computing Surveys, Association for Computing Machinery (ACM), v. 51, n. 6, p. 1–36, feb 2019.

YOSEFIAN, L; FARKHANI, E.M.; BANESHI, M.R. **Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction.** Computational and mathematical methods in medicine, v.2015, p.1-6, 2015.

INSTITUTO NACIONAL DE CÂNCER (Brasil). **Estimativa 2023**: incidência de câncer no Brasil. Rio de Janeiro: INCA, 2022. Disponível em: https://www.inca.gov.br/estimativa. Acesso em: 20 nov. 2025.

FUNDAÇÃO ONCOCENTRO DE SÃO PAULO (FOSP). **Base de dados de câncer hospitalar**. São Paulo: FOSP. Disponível em: https://fosp.saude.sp.gov.br/fosp. Acesso em: 20 nov. 2025.