UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

JHEYMISSON THIAGO SOUSA SILVA

APLICAÇÃO DE RANDOM SURVIVAL FOREST NA PREDIÇÃO DE SOBREVIDA DE PACIENTES COM CÂNCER DE CÉREBRO

CAMPINA GRANDE - PB
2025

JHEYMISSON THIAGO SOUSA SILVA

APLICAÇÃO DE RANDOM SURVIVAL FOREST NA PREDIÇÃO DE SOBREVIDA DE PACIENTES COM CÂNCER DE CÉREBRO

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira
Coorientador(a): Profa. Ma. Débora de Sousa Cordeiro

CAMPINA GRANDE - PB
2025

S586a Silva, Jheymisson Thiago.
Aplicação de random survival forest na predição de sobrevida de pacientes com câncer de cérebro [manuscrito] / Jheymisson Thiago Silva. - 2025.
43 f. : il.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2025.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Departamento de Estatística - CCT".

"Coorientação: Prof. Ma. Débora de Sousa Cordeiro, Departamento de Estatística - CCT".

1. Tumor cerebral. 2. Random Survival Forest. 3. Modelo de riscos proporcionais de Cox. 4. Algoritmos de Machine Learning. I. Título

21. ed. CDD 519.5

Elaborada por Bruno Rafael Freitas de Lima - CRB - 15/1021 BC

JHEYMISSON THIAGO SOUSA SILVA

APLICAÇÃO DE RANDOM SURVIVAL FOREST NA PREDIÇÃO DE SOBREVIDA DE PACIENTES COM CÂNCER DE CÉREBRO

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Estatística da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Estatística

Aprovada em: 06/06/2025.

BANCA EXAMINADORA

Documento assinado eletronicamente por:

- **Tiago Almeida de Oliveira** (***.448.214-**), em **26/06/2025 11:52:04** com chave **200ff1ca529d11f092981a7cc27eb1f9**.
- **Pedro Monteiro de Almeida Júnior** (***.163.384-**), em **26/06/2025 12:09:59** com chave **a05e6e22529f11f0978c1a7cc27eb1f9**.
- **Érika Fialho Morais Xavier** (***.821.464-**), em **26/06/2025 12:17:34** com chave **afc61ddc52a011f0ab7d06adb0a3afce**.

Documento emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.uepb.edu.br/comum/autenticar_documento/ e informe os dados a seguir.
Tipo de Documento: Folha de Aprovação do Projeto Final
Data da Emissão: 26/06/2025
Código de Autenticação: b52d5f

A minha família, pela dedicação, companheirismo e amizade, DEDICO.

AGRADECIMENTOS

A Deus, por ter me dado tudo que sempre precisei.

Ao professor Thiago Almeida de Oliveira, pela paciência e apoio em me orientar no desenvolvimento deste trabalho.

Minha coorientadora Débora de Sousa Cordeiro, cuja paciência e dedicação foram fundamentais ao me orientar durante a elaboração deste trabalho.

A todos os funcionários do Departamento de Exatas pela simpatia e bom atendimento.

A todos os professores da Graduação em Estatística da UEPB, especialmente aos professores Silvio Fernando Alves, Ricardo Alves, Vladimir Costa, Ana Patricia.

A todos os colegas da Graduação em Estatística.

Aos meus familiares, principalmente, minha mãe Ana lúcia Sousa Silva, meu pai Gonzaga José de Souza e a minha esposa Franciely Rodriguês Pereira por toda confiança e apoio.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	Critério de Informação de Akaike
AICc	Critério de Informação de Akaike Corrigido
BIC	Critério de Informação Bayesiano
CART	Classification and Regression Tree
EQM	Erro Quadrático Médio
GAM	Modelos Aditivos Generalizados
IA	Inteligência Artificial
IBS	Brier Score Index
LRT	Likelihood Radio Test
OOB	Out-Of-Bag
RF	Random Forest
RSF	Random Survival Forest
SVM	Máquinas de Vetores de Suporte

SUMÁRIO

APLICAÇÃO DE RANDOM SURVIVAL FOREST NA PREDIÇÃO DE SOBREVIDA DE PACIENTES COM CÂNCER DE CÉREBRO

Jheymisson Thiago Sousa Silva*
Tiago Almeida de Oliveira†

RESUMO

O câncer de cérebro refere-se ao crescimento anormal de células no tecido cerebral, podendo ser primário, quando se origina no próprio cérebro, ou secundário, quando resulta de metástase de outros órgãos. A previsão precisa do prognóstico é essencial para decisões terapêuticas em pacientes com câncer de cérebro. Assim, nesse estudo foi realizado uma análise preditiva sobre a sobrevida de 456 pacientes pelos estados do Brasil que contém o câncer de cérebro com o objetivo de analisar amplamente os fatores que mais influenciam na morte dos pacientes, contribuindo para que decisões clínicas possam ser tomadas com mais precisão, confiabilidade e menor evasivo na sobrevida desses pacientes. Foi realizada uma comparação do modelo de *Random Survival Forest* (RSF) e um modelo típico de riscos proporcionais de Cox. Os cálculos foram feitos com a ajuda do *software* estatístico R. Nesse sentido, o modelo de previsão baseado em algoritmos de aprendizado de máquina é uma alternativa para prever com mais precisão a taxa de sobrevivência e o período de sobrevivência do câncer de cérebro em comparação com outros modelos. Os resultados indicaram que o RSF é uma opção promissora na prática médica, a aplicação permitiu não apenas prever com maior precisão o tempo de sobrevida de pacientes com câncer cerebral, mas também ofereceu subsídios concretos para personalizar condutas terapêuticas.

Palavras-chaves: tumor cerebral; prognóstico; modelo de riscos proporcionais de cox; algoritmos de *machine learning*; *random survival forest.*

* Jheymisson Thiago, Depto Estatística, UEPB, Campina Grande, PB, jheymisson.silva@aluno.uepb.edu.br
† Prof. Tiago Almeida de Oliveira, Depto Estatística, UEPB, Campina Grande, PB, tadolive@servidor.uepb.edu.br

ABSTRACT

Brain cancer refers to the abnormal growth of cells in brain tissue. It can be primary, when it originates in the brain itself, or secondary, when it results from metastasis from other organs. Accurate prognosis prediction is essential for therapeutic decisions in patients with brain cancer. Thus, in this study, a predictive analysis was performed on the survival of 456 patients in Brazilian states with brain cancer, with the aim of broadly analyzing the factors that most influence patient death, contributing to more accurate, reliable, and less evasive clinical decisions regarding the survival of these patients. A comparison was made between the Random Survival Forest (RSF) model (RSF) model and a typical Cox proportional hazards model. The calculations were performed using the statistical software R. In this sense, the prediction model based on machine learning algorithms is an alternative for more accurately predicting the survival rate and survival period of brain cancer compared to other models. The results indicated that RSF is a promising option in medical practice. Its application not only allowed for more accurate prediction of the survival time of brain cancer patients, but also provided concrete support for personalizing therapeutic approaches.

Key-words: brain tumor; prognosis; cox proportional hazards model; machine learning algorithms; random survival forest.

1 INTRODUÇÃO

O câncer de cérebro é caracterizado pelo crescimento descontrolado de células anormais no tecido cerebral, podendo ser classificado como primário (originado no próprio cérebro) ou secundário (metástase de tumores em outras partes do corpo) (Instituto Vencer o Câncer, 2025). Entre os tumores primários, os gliomas — como os glioblastomas — são os mais comuns e agressivos. Historicamente, a compreensão médica sobre tumores cerebrais evoluiu significativamente ao longo dos séculos. Registros antigos, como os do papiro de Edwin Smith (datado de cerca de 1600 a.C.), já descreviam lesões cranianas e sintomas neurológicos que podem ser interpretados atualmente como sinais de patologias intracranianas. No entanto, foi somente no final do século XIX, com o avanço da neurocirurgia e da histologia, que os tumores cerebrais começaram a ser classificados de forma sistemática e tratados com mais eficácia (Weller *et al.*, 2015).

Apesar dos avanços diagnósticos, os tumores cerebrais continuam representando um grande desafio para a medicina moderna. De acordo com, o INCA, o câncer cerebral corresponde a aproximadamente 1,4% a 1,8% de todos os tumores malignos globalmente, existem vários tipos, cada um com características específicas e prognósticos variados. Dentre os tumores malignos do encéfalo mais frequentes, destaca-se o glioma, que se origina das células da glia, que dão suporte e nutrição aos neurônios (Celestino *et al.*, 2024). Já no Brasil, o número de óbitos por câncer cerebral foi de 9.684, com 4.996 mortes em homens e 4.686 em mulheres no ano de 2021. As taxas de mortalidade elevadas, em grande parte são devido à dificuldade de diagnóstico precoce, heterogeneidade biológica dos tumores e limitações nas terapias disponíveis (Machado; Haertel, 2013). Diante disso, modelos estatísticos e de *Machine learning* têm se mostrado promissores ao fornecer apoio à tomada de decisão clínica. Um exemplo disso, Jiang *et al.* (2023) demonstraram o uso de um modelo de rede neural artificial para prever a taxa de sobrevivência de pacientes diagnosticados com neoplasias neuroendócrinas pancreáticas, alavancando informações

clínicas e mostrando a importância dos modelos de *Machine Learning* em resultados clínicos.

Nesse contexto, a estatística tem desempenhado um papel crucial na medicina moderna, sobretudo por meio de modelos de regressão aplicados à análise de sobrevivência. Modelos como a regressão logística e o modelo de Cox são amplamente utilizados para estimar o risco de eventos ao longo do tempo (Kleinbaum; Klein, 2012). O modelo de Cox Proporcional de Riscos, em particular, é um dos mais adotados em estudos clínicos, por sua flexibilidade e capacidade de lidar com dados censurados. No entanto, ele apresenta limitações importantes, como a suposição de proporcionalidade dos riscos e a incapacidade de capturar relações não lineares ou interações complexas entre variáveis. Como alternativa, modelos baseados em árvores de decisão, como as *Random Survival Forests* (RSF), tem-se ganhado destaque por sua capacidade de lidar com dados de alta dimensão, modelar efeitos não lineares e detectar interações automaticamente. As RSF's estendem a ideia das *Random Forest* de Breiman (Breiman, 1986) para dados de sobrevivência, utilizando amostragem *bootstrap*, árvores de decisão e agregação de predições para gerar estimativas robustas do risco de evento (Ishwaran; Kogalur, 2007). Estudos como o de Lunetta e Smith (2004) mostraram que modelos de *Machine Learning* superam métodos tradicionais em tarefas de predição clínica.

Dessa forma, o presente estudo visa predizer o tempo até morte de pacientes com câncer no cérebro, por meio da técnica de *Machine Learning* e *Random Survival Forest* aplicadas a análise de sobrevivência. Adicionalmente, o modelo de Cox será ajustado e seus resultados serão comparados com diferentes configurações de variáveis preditoras no RSF, a fim de encontrar qual modelo tem melhor capacidade preditiva para se ter uma análise de como os modelos de RSF se desempenha comparados aos modelos de sobrevivência tradicionais. Mensurações de performance preditiva de *Brier Score Index* (IBS) e C-Index (Curva ROC generalizada) serão obtidas para avaliar os modelos, além de uma gama de análises gráficas de resíduos (Cox), predições e estimações sobre dados de *Out-Of-Bag* (OOB) de treino e predições sobre os dados de teste. Os resultados serão utilizados para tomadas de decisões frente a redução de mortalidade por câncer de cerébro, melhoria de qualidade de vida dos pacientes e alívio do impacto econômico nos sistemas de saúde.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Análise de sobrevivência

O termo "análise de sobrevivência" está muitas vezes ligado a situações médicas envolvendo dados censurados, que são informações parciais ou incompletas. No entanto, é uma área da estatística perfeitamente apropriada para os campos da engenharia, ciências sociais e economia. Nesta última área, por exemplo, pesquisadores trabalham em estudos de mudanças de empregos, desempregos, promoções e aposentadoria (Torres, 2024). O principal interesse da análise de sobrevivência, para Colosimo e Giolo (2021), é especificar a variável aleatória não negativa "tempo de falha" (tempo até a ocorrência de um evento de interesse).

2.1.1 Tempo de Falha

O tempo de falha, em análise de sobrevivência, pode ser determinado como o tempo até a ocorrência do evento de interesse, ou seja, é o evento a ser analisado no estudo. Aqui a falha é a morte por câncer de cérebro. Segundo Colosimo e Giolo (2021), a falha pode

ser definida como uma variável indicadora representada por:

$$\delta = \begin{cases} 1, & T \leq C, \\ 0, & T > C, \end{cases} \tag{1}$$

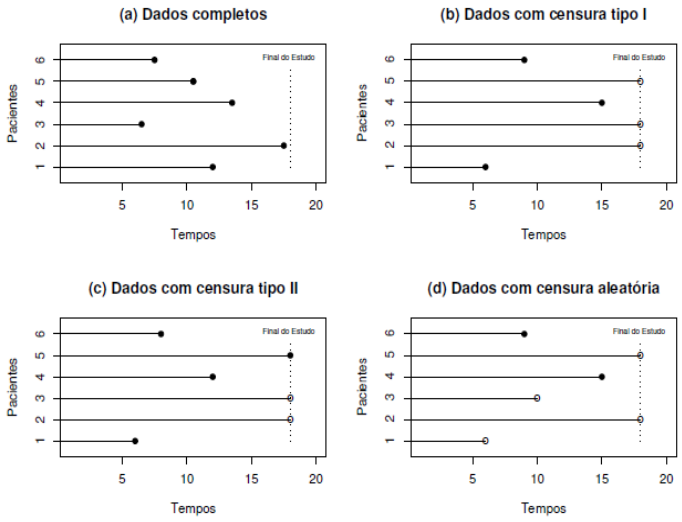
em que:
δ = variável indicadora de falha;
T = variável do tempo de falha;
C = variável independente de T que indica a censura.

2.1.2 *Censura e Dados Truncados*

A censura em análise de sobrevivência pode ser caracterizada como sendo a presença de observações incompletas ou parciais. Existe uma variedade de razões sobre a ocorrência dessas censuras. Por exemplo, a perda de acompanhamento do paciente no decorrer do estudo e o não desfecho do evento de interesse até o término do experimento. Porém, mesmo censurados, todos os resultados dos estudos de sobrevivência devem ser utilizados na análise estatística, devido às observações censuradas fornecerem informações sobre o tempo de vida dos indivíduos sob estudo. Além disso, ocultar as censuras no cálculo pode acarretar em conclusões viciadas. Para Colosimo e Giolo (2021), as censuras terão os seguintes tipos:

- Censura à direita ($T_i > t_i^-$): o tempo de ocorrência do evento de interesse está à direita do tempo registrado.
- Censura à esquerda ($T_i < t_i^+$): ocorre quando não conhecemos o momento da ocorrência do desfecho, mas sabemos que ele ocorreu antes do tempo observado.
- Censura intervalar ($t^- < T < t^+$): ocorre quando não se sabe o tempo exato em que o evento ocorreu, mas sabe-se que ele aconteceu dentro de um determinado intervalo de tempo, ou seja, entre duas observações consecutivas.

Figura 1 – Tipos de censura.



Fonte: Colosimo e Giolo (2021).

Na Figura 1 apresenta-se exemplos de censuras a direta, pois o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Observa-se, que o tipo I ocorre quando um estudo é finalizado após um tempo previamente determinado. Já a censura do tipo II acontece quando o estudo é encerrado depois que um número específico de indivíduos apresenta o evento de interesse. Existe ainda um terceiro tipo de censura, a aleatória, que é a mais comum na prática médica. Esse tipo de censura ocorre quando um paciente é retirado do estudo antes da ocorrência da falha, seja por razões externas ou por um falecimento causado por um fator não relacionado à pesquisa.

De acordo com Carvalho *et al.* (2005), o truncamento ocorre quando indivíduos são excluídos do estudo por motivo relacionado à ocorrência do evento.

2.1.3 Função de Sobrevivência

A função de sobrevivência é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver após esse tempo. Em termos probabilísticos, Kleinbaum e Klein (2012) define da seguinte forma:

$$S(t) = P(T \leq t). \tag{2}$$

Como resultado, a função de distribuição acumulada, ou seja,

$$F(t) = 1 - S(t), \tag{3}$$

é definida como a probabilidade de uma observação não sobreviver ao tempo t.

Devido a dificuldade em se estimar a função de sobrevivência e ter uma precisão confiável quando se contêm censura na base de dados, faz-se necessário a utilização de estimadores não-paramétricos que garantam a confiabilidade dos resultados. Os mais comumente utilizados para se estimar a função de sobrevivência são o estimador de Kaplan-Meier e o de Nelson-Aalen.

2.1.4 Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier, porposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é também chamado de estimador limite-produto, devido a forma como ele calcula a função de sobrevivência: por meio de um produto acumulado de probabilidades condicionais de sobrevivência. O estimador de Kaplan-Meier pode ser determinado como uma estatística que consegue englobar as censuras e, além de estimar a curva de sobrevida, esse estimador também determina a probabilidade de um evento acontecer em um tempo pré-definido. A sua função é dada por:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j} \right), \tag{4}$$

em que:

- t: é o tempo de falha;
- d_j: o número de falhas;
- n_j: é o número de indivíduos sob risco em t_j.

Nesse trabalho será aplicado o método para estimar as probabilidades de sobrevivência dos pacientes que sofreram câncer de cérebro, analisando a influência dos fatores de forma individual.

2.1.5 Teste de Log-Rank

Para verificar se há diferenças estatisticamente significativas nas estimativas das curvas de Kaplan-Meier entre os grupos, é preciso que um teste seja aplicado. Sendo assim, um teste bastante utilizado na análise de sobrevivência é o teste não-paramétrico de *Log-Rank* (Mantel; Haenszel, 1966), que mede a diferença entre o número observado de falhas em cada grupo, como descrito por (Carvalho *et al.*, 2005). Seu uso é apropriado quando o risco das populações a serem comparadas é aproximadamente constante. A estatística desse teste é dada por:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{V_i}. \tag{5}$$

Em que:

- O_i é o número observado de eventos no grupo i;
- E_i é o número esperado de eventos no grupo i;
- V_i é a variância da diferença $O_i - E_i$;
- k é o número de grupos.

2.1.6 Teste de Wilcoxon

Outro teste bastante conhecido e amplamente utilizado para comparar dois ou mais grupos é o teste generalizado de Wilcoxon (Gehan, 1965). Diferente do teste de *Log-Rank*, que atribui o mesmo peso a todas as observações ao longo do tempo e, consequentemente, dá maior destaque aos períodos mais tardios, o teste de Wilcoxon ajusta os pesos de acordo com o número de indivíduos ainda em risco. Isso significa que ele dá maior importância aos momentos iniciais do estudo. A estatística desse teste é definida por:

$$W = \frac{\left(\sum_j w_j(O_{kj} - E_{kj})\right)^2}{\sum_j w_j^2 V_j}, \tag{6}$$

em que:

- w_j é o peso associado ao tempo t_j, geralmente o número de indivíduos em risco no tempo t_j;
- O_{kj} é o número observado de eventos em k grupos no tempo t_j;
- E_{kj} é o número esperado de eventos em k grupos no tempo t_j (sob a hipótese nula de que as curvas de sobrevivência são iguais);
- V_j é a variância da diferença $O_{kj} - E_{kj}$.

Portanto, para realizar a comparação das curvas entre os grupos e verificar se há diferença significativa entre elas, são analisadas as seguintes hipóteses:

$$\begin{cases} H_0 : \mu_1 = \mu_2; \\ H_1 : \mu_1 \neq \mu_2. \end{cases} \tag{7}$$

A hipótese refere-se à função de risco em todo o tempo de observação e não às diferenças/semelhanças em trechos da curva. Sendo o p-valor menor que o valor da estatística de teste, rejeita-se a hipótese nula indicando que pelo menos uma curva difere das outras, significativamente, em algum momento do tempo.

2.1.7 Modelagem

Na análise de sobrevivência geralmente são utilizados dois tipos principais de modelos para ajustar os dados, considerando as diversas informações inerentes ao evento de interesse, isto é, as covariáveis.

2.1.8 Modelagem paramétrica

A modelagem paramétrica é uma das técnicas utilizadas na análise de sobrevivência, na qual se assume que os tempos de falha seguem uma distribuição de probabilidade específica, como Exponencial, Weibull, Log-normal ou Gama.

A modelagem paramétrica é amplamente utilizada em diversas áreas, incluindo engenharia, medicina e ciências sociais, permitindo análises preditivas robustas e fornecendo *insights* importantes sobre o comportamento dos dados, como aponta (Klein; Moeschberger, 2003).

2.1.9 Modelagem semi-paramétrica

Em análise de sobrevivência, o interesse muitas vezes é de se estudar a relação entre as covariáveis, ou variáveis explicativas, e o tempo de vida. Uma forma adequada de o fazer é através de modelos de regressão. No caso em que se ajusta aos dados uma distribuição conhecida, tem-se o que se conhece por modelo de regressão paramétrico, segundo (Cox, 1972). Porém, em várias situações, existem dúvidas sobre a distribuição adequada para os dados. Nesse contexto, o modelo mais utilizado na literatura é o chamado modelo de regressão semi-paramétrico de riscos proporcionais de Cox, sendo abreviadamente designado por modelo de regressão de Cox.

Os modelos de regressão de Cox geralmente são utilizados para analisar o tempo até ocorrência de um determinado evento de interesse. Estes modelos estimam a função de risco, em que esta depende de um conjunto de covariáveis. Cox (1972) propôs um modelo em que a função de risco no instante t é definida por:

$$\lambda(t) = \lambda_0(t) g(\mathbf{x}'\boldsymbol{\beta}), \tag{8}$$

sendo:

- $\lambda_0(t)$ componente não-paramétrico e não especificado, uma função não-negativa do tempo;
- $g(\mathbf{x}'\boldsymbol{\beta})$ componente paramétrico, geralmente utilizado em sua forma multiplicativa, garantindo que $\lambda(t)$ será sempre uma função positiva.

No qual

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta}) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p). \tag{9}$$

- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ são os coeficientes das covariáveis, estimados pelo modelo.

2.1.10 Estimação do Modelo de Cox

Para realizar a inferência nos dados, Cox ajusta a função de verossimilhança parcial ao modelo.

A função de verosimilhança parcial proposta por Cox (1972) é definida por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} \right)^{\delta_i}. \tag{10}$$

Sendo δ_i o indicador de falha. Os valores de β que maximizam a função de verossimilhança parcial $L(\boldsymbol{\beta})$ são determinados através da solução do sistema de equações, dado por $U(\boldsymbol{\beta}) = 0$, em que $U(\boldsymbol{\beta})$ é o vetor das primeiras derivadas da função $l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta}))$. Isto é,

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i - \frac{\sum_{j\in R(t_i)} \mathbf{x}_j \exp(\mathbf{x}_j'\beta)}{\sum_{j\in R(t_i)} \exp(\mathbf{x}_j'\beta)} \right] = 0. \tag{11}$$

A função de verossimilhança parcial assume que os tempos de sobrevivência são contínuos e supõe a não possibilidade de empates nos dados observados. Mas, na prática, se houver empates entre falhas e censuras, aceita-se que a censura ocorreu após a falha.

2.1.11 Seleção de covariáveis no modelo de Cox

Para desenvolver modelos de regressão voltados à análise de sobrevivência, é fundamental construir modelos que sejam parcimoniosos — isto é, que utilizem o menor número possível de covariáveis sem comprometer o ajuste aos dados. Esse equilíbrio entre simplicidade e capacidade preditiva é alcançado por meio de critérios estatísticos que ajudam a decidir quais covariáveis devem ser incluídas no modelo.

No contexto dos modelos de Cox, a estimação dos parâmetros é realizada por meio do método da máxima verossimilhança, que busca os valores dos coeficientes que maximizam a probabilidade de observar os dados amostrais disponíveis. Uma vez obtida essa estimação, podemos realizar testes de hipótese para verificar a significância das covariáveis no modelo.

Dentre os testes baseados na verossimilhança, destacam-se o teste de Wald e o teste da razão de verossimilhança (LRT — *Likelihood Ratio Test*). Neste trabalho, adotamos o segundo, por ser amplamente utilizado na comparação entre modelos aninhados.

O *LRT* compara dois modelos: um modelo *reduzido*, que exclui a covariável em teste, e um modelo *saturado*, que a inclui. A estatística do teste é baseada na diferença entre os logaritmos das verossimilhanças dos dois modelos:

$$LRT = -2\log\left(\frac{L(\textit{modelo reduzido})}{L(\textit{modelo saturado})}\right) = -2\log(L_r) + 2\log(L_s), \tag{12}$$

em que o L_r é a verosimilhança do modelo reduzido e o L_s é a verossimilhança do modelo saturado. É importante salientar que se o modelo reduzido tem p parâmetros o modelo saturado tem $p + r$ parâmetros. A estatística de teste, LRT, segue uma distribuição Qui-quadrado, $T \sim \chi^2_r$, no qual r é o número de parâmetros estimados no modelo. De acordo com o teste de razão de verossimilhança, rejeita-se a hipótese nula se o valor da estatística for inferior ao nível de significância.

Para uma discussão aprofundada sobre a estimação por máxima verossimilhança e testes baseados em verossimilhança, ver os trabalhos (Cox; Oakes, 1984) e (Klein; Moeschberger, 2003).

2.1.12 Seleção do modelo adequado

A escolha de um modelo apropriado, do ponto de vista estatístico, é um tópico fundamental na análise de dados. Deve-se selecionar o modelo mais parcimonioso, ou seja,

o modelo que envolve o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento de uma variável resposta.

Para selecionar covariáveis de forma sequencial pode-se utilizar três métodos distintos: *Backward* (seleção regressiva), *Forward* (progressiva) e o *Stepwise*. Nesse estudo será utilizado o último método, pois ele combina os dois primeiros métodos, adicionando covariáveis uma a uma. Porém, em cada passo é feita uma análise das covariáveis já introduzidas até então, como forma de garantir que permanecem relevantes após a introdução de novas covariáveis, permimtindo assim uma melhor compreensão dos fatores que influenciam a sobrevida dos pacientes.

Dentro dos critérios para a seleção de modelos, os critérios baseados no máximo da função verosimilhança são os mais utilizados, como o teste da razão verosimilhança e o Critério de Informação de Akaike (AIC).

O Critério de Informação de Akaike (AIC) admite a existência de um modelo real que descreve os dados que é desconhecidos e tenta escolher entre um grupo de modelos avaliados (Akaike, 1974), o que minimiza a divergência de Kullback e Leibler (1951), ou seja, escolhe o melhor modelo como sendo aquele que apresenta o menor valor de AIC. A estimativa de AIC para um determinado modelo é dada por:

$$AIC = -2L + 2r, \tag{13}$$

em que, o **L** é o logaritmo da função de verosimilhanças com os parâmetros $\boldsymbol{\theta}$ e **r** é o número de parâmetro do modelo.

Entretanto, quando o tamanho da amostra (n) é pequeno em relação ao número de parâmetros, o AIC tende a favorecer modelos mais complexos. Para corrigir esse viés, Sugiura (1978) propôs o AIC corrigido (AICc), dado por:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}, \tag{14}$$

ou, substituindo o AIC:

$$AICc = -2\log(\hat{L}) + 2k + \frac{2k(k+1)}{n-k-1}. \tag{15}$$

Esse critério é particularmente importante quando $n/k < 40$, sendo mais confiável para amostras pequenas ou modelos com muitos parâmetros em relação ao tamanho da amostra.

Tem-se também, o Critério de Informação Bayesiano (BIC) e é definido como:

$$BIC = -2\log(\hat{L}) + k\log(n), \tag{16}$$

em que, os termos têm o mesmo significado que no AIC. A principal diferença está na penalização da complexidade do modelo: o BIC aplica uma penalização mais severa à medida que o tamanho da amostra aumenta, favorecendo modelos mais parcimoniosos. Para este estudo foi utilizado o AIC via método *stepwise*, pelo fato, que a escolha das variáveis é essencial para a eficiência do modelo e para se ter uma melhor interpretabilidade da sobrevida dos pacientes com câncer de cérebro.

2.2 Aprendizado de Máquina

2.2.1 Introdução ao Aprendizado de Máquina (Machine Learning)

Aprendizado de Máquina (*Machine Learning*) é considerado como um subcampo da Inteligência Artificial e está relacionado ao desenvolvimento de técnicas e métodos que permitem que o computador aprenda (Mitchell, 1997). Em termos simples, esta área do conhecimento dedica-se ao desenvolvimento de algoritmos que permitem que máquinas aprendam e executem tarefas e atividades. A aprendizagem automática apresenta grande intersecção com a área da estatística. A teoria da aprendizagem de máquina foi estabelecida no final da década de 60. Até meados dos anos 90, referia-se somente a uma análise teórica do problema de estimação de funções a partir de um conjunto de dados. Nas últimas décadas, com a crescente complexidade dos problemas a serem tratados computacionalmente, o processo de aprendizado de máquina tem sido marcado pela crescente utilização de *softwares*. Além disso, os desafios em aprender com os dados levaram à evolução do conjunto de ferramentas e técnicas estatísticas utilizadas. A aprendizagem estatística possui um papel fundamental em diversas áreas, tais como: ciência, finanças e indústria. Alguns exemplos são (Zhu; Hastie, 2001):

- Prever se um paciente hospitalizado, devido a um ataque cardíaco, terá um segundo ataque cardíaco. A previsão deve ser baseada em medidas clínicas e demográficas, para esse paciente;
- Identificar os números de um código postal manuscrito, a partir de uma imagem digitalizada, dentre inúmeros outros exemplos.

Em geral, suponha uma resposta quantitativa $\mathbf{Y} = (Y_1, \ldots, Y_n)$ (variável de saída - *output*) e p diferentes preditores e $(X_1, X_2, \ldots, X_p)$ (variáveis de entrada - *input*). Assuma que há relação entre $\mathbf{Y}$ e $\mathbf{X}$ é uma matriz $(X_{i1}, \ldots, X_{ip})$ com $i = 1, \ldots, n$ e que possa ser escrita na forma geral:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{17}$$

em que $\boldsymbol{\beta}$ é o vetor de coeficientes e $\boldsymbol{\epsilon}$ é o vetor de erros aleatórios, independente de $\mathbf{X}$ e com média 0. O $\boldsymbol{\beta}$ pode envolver várias variáveis de entrada e deve ser estimada com base nos dados observados. De acordo com James, Sandhya e Thomas (2013), a aprendizagem de máquina refere-se a um conjunto de ferramentas utilizadas para a compreensão dos dados. Tais ferramentas podem ser divididas em duas grandes áreas: métodos supervisionados e métodos não supervisionados.

2.2.2 Aprendizado Supervisionado versus Aprendizado não Supervisionado

O aprendizado supervisionado ocorre quando o modelo aprende a partir de resultados pré-definidos, utilizando os valores passados da variável *target* para aprender quais devem ser seus resultados de saída. Estes mesmos valores servem como "supervisão" destas previsões, permitindo o ajuste nas previsões com base nos erros. Isso é considerado aprendizado supervisionado, pois o modelo possui uma referência daquilo que está certo e daquilo que está errado.

Diversos métodos estatísticos fazem parte dessa abordagem, como regressão linear, Modelos Lineares Generalizados e Modelos Aditivos Generalizados (GAM) introduzidos por Hastie e Tibshirani (1990). Técnicas que usam reamostragem, tal como *boosting* para

reduzir viés e variância em classificação, como aponta Breiman (1986) e *Support Vector Machines (SVM)*, todos voltados para problemas de classificação binária e para regressão, conforme Vapnik (1995).

Ao contrário do aprendizado supervisionado, no aprendizado não supervisionado não existem resultados pré-definidos para o modelo utilizar como referência para aprender. Por exemplo, é apresentado para o modelo um conjunto de dados com várias informações sobre plantas (como comprimento do caule, cor da folha, espessura da raiz, etc.) e pede-se para o modelo separar esses dados em 5 categorias. Nesse caso, não foi especificado para o modelo o que caracteriza cada categoria, o modelo irá sozinho tentar encontrar semelhanças e diferenças entre os dados de maneira que essa separação em 5 categorias seja a melhor possível. Alguns exemplos de modelos de aprendizado não supervisionado seria *K-Means*, PCA (*Principal Component Analysis*), *Apriori Algorithm*, entre outros.

2.2.3 Procedimento para validação do processo de treinamento

Para validação da capacidade preditiva do modelo, é necessário calcular métricas da qualidade das estimativas. Dessa forma, geralmente divide-se os dados em dois subconjuntos, denominados, dados de treinamento e teste.

Os dados de treino correspondem à porção do conjunto de dados utilizada para ajustar o modelo de *Machine Learning*. Eles são responsáveis por ensinar ao algoritmo os padrões e relações existentes entre as variáveis e geralmente representam cerca de 70% da base total de dados.

Já os dados de teste são utilizados após o treinamento do modelo, com o objetivo de avaliar sua capacidade de generalização em situações não previamente observadas. Essa etapa simula o desempenho do modelo em condições reais de previsão e normalmente compreende os 30% restantes do conjunto de dados.

2.2.4 Erro de treino e erro de teste

Suponha que f seja estimado a partir das observações $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, em que y_i são qualitativos. A taxa de erro de treino é a proporção de erros cometidos pelo classificador no conjunto de treino:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i), \tag{18}$$

no qual $I(y_i \neq \hat{y}_i)$ é um indicador que assume 1 se houver erro $(y_i \neq \hat{y}_i)$ e 0 caso contrário. No aprendizado de máquina, o modelo busca uma função que relacione os valores x com y. O erro no teste é a média dos erros em novos dados não vistos no treino:

$$Média(I(y_0 \neq \hat{y}_0)), \tag{19}$$

em que y_0 é a classe prevista para a entrada x_0. Um bom classificador possui um baixo erro de teste.

2.2.5 Métricas de qualidade da estimativa

Para analisar o desempenho de um método de aprendizagem de máquina em um conjunto de dados é necessário medir o quanto as previsões para determinadas observações estão próximas de seus respectivos valores verdadeiros. A medida geralmente utilizada

no ajuste de regressão é o Erro Quadrático Médio - EQM (*Mean Squared Error* - MSE), dado por

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2, \tag{20}$$

$\hat{y}$ é a previsão de y para a i-ésima observação.

O EQM pode ser calculado nos dados de treino, sendo chamado de EQM_{TR}. No entanto, o mais relevante é o erro em dados de teste, pois avalia a generalização do modelo. No treino, conhece-se a resposta correta, enquanto no teste, deseja-se prever valores desconhecidos.

Matematicamente, considerando observações de treino $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, obtém-se uma estimativa $\hat{f}$, utilizada para calcular $\hat{f}(x_1), \ldots, \hat{f}(x_n)$. Se essas previsões forem próximas de $y_1, \ldots, y_n$, então o EQM_{TR} será pequeno. No entanto, o objetivo principal é minimizar o erro em novas observações, ou seja, o EQM_{TS}, dado por:

$$EQM_{TS} = \frac{1}{n_{TS}}\sum_{j=1}^{n_{TS}}(\hat{f}(x_{0j}) - y_{0j})^2. \tag{21}$$

Como os dados de teste podem não estar disponíveis, uma estratégia comum é escolher um modelo que minimize o EQM_{TR}, assumindo que também minimizará o EQM_{TS}. Entretanto, isso pode não ser verdadeiro, pois alguns métodos podem ajustar-se excessivamente aos dados de treino, resultando em um alto erro em novos dados.

Independentemente do conjunto de dados utilizado, quando um método produz pequeno EQM_{TR} mas grande EQM_{TS}, diz-se que está a ocorrer um problema de "sobreajustamento" (*overfitting*). Isto ocorre porque o procedimento de aprendizagem estatística está muito sensível a qualquer mudança, mesmo as que ocorrem meramente ao acaso, e as classifica incorretamente como se fossem "padrões".

Quando há sobreajustamento nos dados de treino, o EQM_{TS} será muito grande, porque os supostos "padrões" não existem nos dados de teste. De forma geral, independentemente de ter ocorrido sobreajustamento, quase sempre espera-se que o EQM_{TR} seja menor do que o EQM_{TS}, devido ao já mencionado fato dos métodos estatísticos em geral serem projetados visando minimizar o EQM_{TR}. O sobreajustamento ocorre frequentemente nos casos em que o modelo em estudo é pouco flexível.

Outra métrica é o Erro Percentual Absoluto Médio (MAPE, do inglês *Mean Absolute Percentage Error*) é uma métrica de desempenho comumente utilizada em problemas de regressão. Ele quantifica, em termos percentuais, o quão distantes estão as previsões dos valores reais representando o erro médio percentual cometido pelo modelo em relação aos valores observados. MAPE é dado por

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{22}$$

em que:

- y_i representa o valor real da i-ésima observação;
- $\hat{y}_i$ representa o valor predito pelo modelo;
- n é o número total de observações.

O MAPE fornece uma medida relativa do erro, o que o torna útil para comparar desempenho entre diferentes modelos ou conjuntos de dados. Contudo, a métrica apresenta

limitações quando existem valores reais $y_i = 0$, pois resulta em divisão por zero, ou quando os valores reais são muito próximos de zero, o que pode inflar o erro percentual.

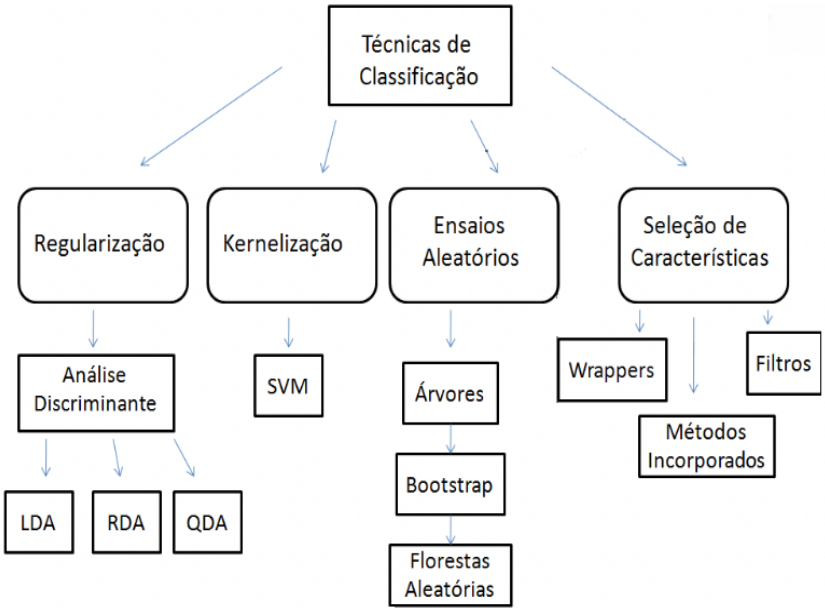
2.2.6 Métodos de classificação

Em muitas aplicações, a variável resposta é qualitativa, exigindo o uso de técnicas de classificação. O objetivo da classificação é alocar observações em categorias distintas com base em suas características. Um exemplo prático é a estratificação de pacientes em grupos de risco.

Nesse contexto, o problema pode envolver duas ou mais classes e busca-se, a partir de um conjunto de dados de treinamento, aprender uma função que permita classificar novas observações. Dentre os métodos mais utilizados estão a regressão logística, a análise discriminante, os vizinhos mais próximos, além de técnicas mais robustas como árvores de decisão, *Random Forest* e *Support Vector Machines (SVM)*.

Neste trabalho, o foco será nos métodos baseados em árvores, especialmente nas *Random Survival Forest*, que se mostraram promissoras em diversas aplicações biomédicas. A Figura 2 ilustra uma taxonomia simplificada das principais técnicas de classificação, destacando a família dos ensaios aleatórios, da qual as *Random Forest* fazem parte.

Figura 2 – Técnicas de Classificação e métodos específicos.



Fonte: Adaptado de Wu *et al.* (2015).

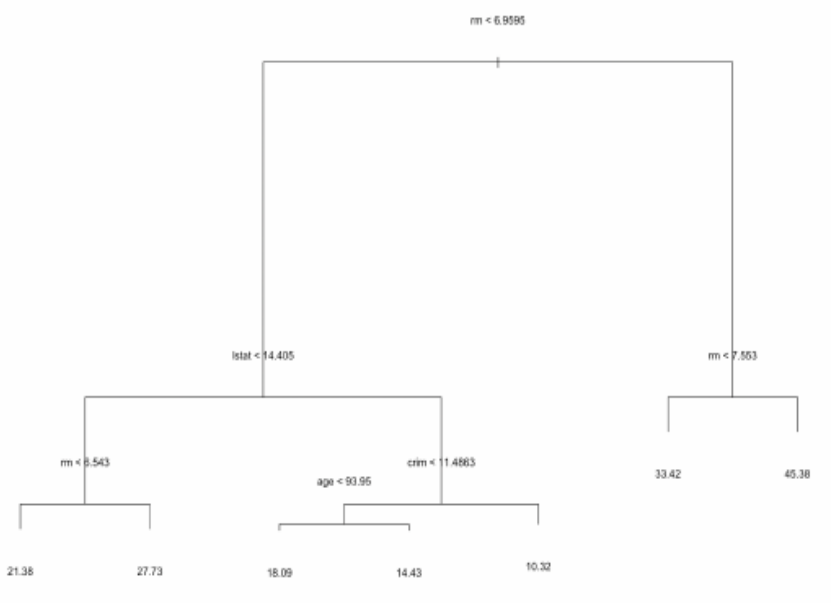
2.2.7 Árvores de Decisão para a Classificação e Regressão (CART)

As árvores de decisão são modelos preditivos de aprendizado supervisionado amplamente utilizados, tanto em problemas de classificação quanto de regressão. Neste trabalho, o enfoque recai sobre a regressão, na qual o objetivo é predizer o tempo de sobrevida do paciente relacionando com algumas variáveis explicativas. O método CART (*Classification and Regression Trees*), proposto por Breiman (1986), segmenta o espaço das covariáveis em regiões disjuntas por meio de regras binárias recursivas, atribuindo a cada região um valor médio da variável resposta observada.

Embora as árvores de decisão sejam intuitivas e de fácil interpretação, elas tendem a apresentar elevada variância e podem ser sensíveis a pequenas alterações nos dados, o que pode comprometer sua acurácia preditiva. Para mitigar essas limitações, surgiram métodos de ensemble como o *Random Forest*, que combina várias árvores de regressão construídas a partir de amostras *bootstrap* e subconjuntos aleatórios de variáveis, reduzindo a variância do modelo e aumentando sua capacidade de generalização.

Originalmente introduzidos por Morgan e Sonquist (1963), os métodos baseados em árvores ganharam destaque nos anos 1980 com a formalização do algoritmo CART. No contexto de regressão, a divisão do espaço preditor é guiada por critérios como a minimização do erro quadrático médio dentro dos nós, buscando particionar os dados em regiões homogêneas em relação à variável resposta. A Figura 3 ilustra um exemplo de estrutura de árvore de decisão para fins preditivos.

Figura 3 – Um exemplo de árvores de regressão.



Fonte: Technologies (2020).

A árvore prevê um preço médio de 45,38 para casas maiores em subúrbios em que os residentes tem alto status socioeconômico ($rm \geq 6,9595$) .

2.2.8 Random Forest

A técnica de *Random Forest* (RF) trata-se de uma abordagem poderosa para a exploração de dados, análise de dados e construção de modelos preditivos. Esta técnica foi desenvolvida por (Breiman, 1986) (criador do CART) na Universidade da Califórnia, Berkeley. Árvores de decisão individuais geralmente apresentam alta variabilidade ou grande viés. As RS's tentam mitigar tais problemas ao encontrar um equilíbrio natural entre estes dois extremos. Considerando que as RF's têm poucos parâmetros de sintonia (*tunning*) e podem ser utilizadas simplesmente com configurações padrão (*default*) de parâmetros, elas representam uma ferramenta simples de usar quando não se tem um modelo ou para produzir um modelo razoável rápido e eficiente. Segundo Technologies (2020) os maiores benefícios das RF's são:

- Acurácia;

- Eficiência - mesmo para grandes bases de dados;
- Capacidade de lidar com milhares de variáveis de entrada;
- Proporciona estimativas de quais variáveis são importantes em problemas de classificação;
- Gera uma estimativa não viciada do erro de generalização à medida que a floresta é construída;
- Proporciona métodos efetivos para a estimação de dados faltantes;
- Mantém a acurácia mesmo na presença de dados faltantes;
- Proporciona métodos que funcionam bem para dados não balanceados;

As desvantagens das Random Forest, incluem:

- Não são facilmente interpretáveis;
- Um grande número de árvores pode tornar o algoritmo lento e ineficaz para previsões em tempo real;
- Ao trabalhar com variáveis numéricas contínuas, a árvore de decisão perde informações quando categoriza variáveis em diferentes categorias.

2.2.9 Random Survival Forest

Uma *Random Survival Forest* (RSF) é um método de conjunto não paramétrico para a análise de dados de sobrevivência censurados à direita, construído como uma extensão de tempo para evento de *Random Forest* para classificação e regressão (Ishwaran; Kogalur, 2007). O método pode lidar com múltiplas covariáveis, covariáveis de ruído, bem como relações complexas e não lineares entre covariáveis sem necessidade de especificação prévia. Como tal, as RSFs frequentemente desempenham um papel na substituição da regressão de Cox clássica quando a suposição de riscos proporcionais está em questão, pois é livre de suposições (Bou-Hamad; Larocque; Ben-Ameur, 2011).

Neste sentido, existe um pacote o *randomForestSRC* do R Core Team (2025) para implementar a *Random Survival Forest*, sendo descrito pelos seguintes passos:

1. Obtenha n amostras de *bootstrap* formando n árvores a partir dos dados originais;
2. Cultive uma árvore para cada conjunto de dados amostrado. Em cada nó da árvore selecione aleatoriamente preditores (covariáveis) de divisão. Assim, utilizando um critério de divisão de sobrevivência onde o nó é dividido nesse preditor que maximiza diferenças de sobrevivência entre nós-filhas;
3. Cresça a árvore até o tamanho máximo sob a restrição que um nó terminal não deve ter um tamanho menor do que número de mortes únicas;
4. Calcule uma estimativa de risco acumulado combinando informações das n-árvores. Uma estimativa para cada indivíduo nos dados é calculada;

5. Calcule uma taxa de erro *Out-Of-Bag* (OOB) para o conjunto derivado usando as b árvores, onde $b = 1, \ldots, n$ árvores.

As regras de divisões de nó são fatores cruciais para o algoritmo. Diante disso, o pacote *randomForestSRC* fornece duas regras diferentes de divisão de sobrevivência para o utilizador. Estas são:

- Uma regra de divisão de logrank, a regra de divisão padrão, chamada pela opção `splitrule = "logrank"`;
- Uma regra de pontuação logrank, `splitrule = "logrankscore"`.

A principal diferença entre os modelos *Random Forest* (RF) e *Random Survival Forest* (RSF) está no tipo de resposta que cada abordagem é projetada para prever. Enquanto o RF é utilizado para problemas de classificação ou regressão com variáveis resposta categóricas ou contínuas, o RSF é uma extensão desse modelo voltada para análise de sobrevivência, ou seja, problemas em que a variável de interesse é o tempo até a ocorrência de um evento, podendo estar sujeito à censura à direita. Assim, o RSF representa uma generalização do RF adequada para lidar com a censura nos dados e para modelar a distribuição de tempo até o evento, sendo particularmente útil em contextos biomédicos, como estudos clínicos de prognóstico e predição de sobrevivência.

2.2.10 Métricas de avaliação do modelo

A avaliação da precisão da previsão é um aspecto essencial em *machine learning* para validar modelos. No caso de dados de sobrevivência que possuem censura à direita, essa precisão é mensurada considerando métricas que levam em conta a dependência dos tempos, como a área sob a curva característica operacional e o *Brier score* (Graf *et al.*, 1999). Para analisar o desempenho da previsão em um modelo de *Random Forest*, pode-se utilizar os dados *Out-Of-Bag* (OOB). No entanto, quando a intenção é comparar modelos *Random Forest* com técnicas de previsão para análise de sobrevivência (RSF), é necessário realizar uma validação cruzada, o que permite estimar a acurácia da previsão. Dentro desse contexto, a estatística C-index é empregada nos dados *Out-Of-Bag* para medir a precisão dentro do conjunto de treinamento. Por outro lado, a estatística *Brier score* (IBS) é aplicada sobre os dados de teste, permitindo avaliar a qualidade da previsão feita pelo modelo.

2.2.11 C-Index

Considere os tempos de sobrevivência e os *status* de censura para os n indivíduos em um nó terminal h, representados por $(T_{1,h}, \delta_{1,h}), (T_{2,h}, \delta_{2,h}), \ldots, (T_{n,h}, \delta_{n,h})$. Além disso, seja $t_{1,h} < t_{2,h} < \ldots < t_{m,h}$ a sequência dos tempos de evento distintos no nó terminal h. Define-se então $d_{1,h}$ e $Y_{1,h}$ como o número de mortes e o número de indivíduos em risco no tempo $t_{1,h}$. A função de risco acumulada para estimar o nó terminal h é dada pelo estimador de Nelson-Aalen:

$$\hat{H}_h(t) = \sum_{t_{1,h} \leq t} \frac{d_{1,h}}{Y_{1,h}}. \tag{23}$$

Para um indivíduo i com uma covariável x_i de dimensão d, a função de risco acumulada é expressa como:

$$H(t|x_i) = \hat{H}_h(t), \quad se\ x_i \in h. \tag{24}$$

Nos procedimentos de *Random Survival Forest* (RSF), para estimar a função de risco acumulada para o indivíduo i, define-se $I_{i,b} = 1$ se i pertence ao conjunto *Out-Of-Bag* (OOB) na b-ésima amostra via *bootstrap*. Caso contrário, $I_{i,b} = 0$. Além disso, para a árvore gerada na b-ésima amostra via *bootstrap*, considera-se $H_b(t|x_i)$ como a função de risco acumulada para o indivíduo i.

Dessa maneira, a função de risco acumulada conjunta para i é obtida como:

$$H(t|x_i) = \frac{\sum_{b=1}^{B} I_{i;b} H_b(t|x_i)}{\sum_{b=1}^{B} I_{i;b}}. \tag{25}$$

Em árvores de sobrevivência, o indivíduo i será dito como tendo um risco predito mais alto do que o indivíduo j se

$$\sum_{l=1}^{m} H(t_l|x_i) > \sum_{l=1}^{m} H(t_l|x_j), \tag{26}$$

em que $t_1 < t_2 < \ldots < t_m$ são os tempos de eventos únicos no conjunto de dados. Neste sentido, em RSF, a função de risco acumulada $H(T|X)$ conjunta é mais utilizada do que $H(T|x)$. Um valor de $0,5$ para a estatística C-index não é considerado melhor do que, por exemplo, um resultado de cara ou coroa. Por outro lado, o valor 1 indica uma habilidade discriminatória completa.

2.2.12 Brier-Score

O *Brier Score* até o tempo t é definido como:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(0 - \hat{S}(t|x_i))^2}{\hat{G}(t_i)} I(T_i \leq t, \delta_i = 1) + \frac{(1 - \hat{S}(t|x_i))^2}{\hat{G}(t)} I(T_i > t) \right), \tag{27}$$

em que $\hat{G}(t)$ representa o estimador de Kaplan-Meier para a função de sobrevivência. A curva de erro de previsão é obtida a partir do *Brier score* ao longo do tempo. O *Brier score* integrado (IBS) corresponde ao erro acumulado ao longo do tempo e pode ser calculado como:

$$IBS = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} BS(t)\, dt. \tag{28}$$

Nesse contexto, um valor de IBS próximo de 1 indica uma capacidade preditiva fraca do modelo, enquanto valores próximos de 0 refletem uma melhor capacidade preditiva. Portanto, por meio dessas métricas pode-se predizer a sobrevida dos pacientes e ter um melhor prognóstico sobre o tempo médio de vida.

3 MATERIAIS E MÉTODOS

Os dados sob estudo são referentes às informações de 456 pacientes que foram coletados a partir do conjunto de dados disponíveis no Instituto Nacional de Câncer (2021) durante o intervalo de 748 dias, entre os anos de 2018 a 2021. Desse total, 185 pacientes foram censurados, ou seja, deixaram de ser acompanhados, ou simplesmente não sofreram o evento de interesse até o termino do estudo, enquanto os outros 271 pacientes que restaram foram acometidos pelo câncer de cérebro. Este banco de dados oferece dados extensos e detalhados do paciente, incluindo características demográficas, informações relacionadas ao tumor, gênero, idade, causa da morte, duração da sobrevida,

entre outras observações. Os pacientes atenderam aos critérios de elegibilidade para o estudo randomizado como diagnóstico patológico confirmado de câncer de cérebro, além disso, os pacientes precisavam ter um *status* e tempo de sobrevivência conhecidos.

Primeiramente, foi conduzida uma análise descritiva por meio da média ou frequências da situação dos pacientes, logo após, utilizou-se o método não-paramétrico para uma análise de sobrevivência introdutória dos dados, consequentemente ajustou-se um modelo semi-paramétrico de regressão de Cox com todas as variáveis da bases de dados e posteriormente uma analise dos resíduos de Schoenfeld, Deviance e Escore. E por fim, para a seleção da melhor configuração de covariáveis que explicassem o risco de morte decorrente do câncer de cérebro, aproveitou-se o Critério de Informação de Akaike (AIC). Após a definição do modelo de Cox mais adequado, foram ajustados diferentes modelos utilizando a técnica de *Random Survival Forest* (RSF), com base em distintos critérios de seleção. Os modelos avaliados foram:

- M1 (RF - Completo): contemplou todas as variáveis disponíveis no banco de dados;
- M2 (RF - Cox): replicou a configuração do modelo de Cox via AIC no qual utilizou-se as variáveis que se ajustaram melhor;
- M3 (RF - Corrente): incluiu as variáveis selecionadas com base na importância medida pelo fator de variância;
- M4 (Cox - Clássico): modelo de Cox tradicional, ajustado por verossimilhança parcial e análise diagnóstica dos resíduos.

Foram consideradas 1500 amostras *bootstrap* para compor as *random forest*, e utilizou-se a divisão do conjunto de dados em 70% para treino e 30% para teste, além do uso das amostras *Out-Of-Bag* para compor a análise de precisão do *Random Survival Forest*. As análises foram feitas por meio do *software* estatístico R na versão 4.4.1, de uso livre e amplamente utilizado em âmbito acadêmico, abrangendo também a modelagem via técnica de *Machine Learning* com o algoritmo do tipo *Random Forest*, nos pacotes `randomForestSRC`, e as análises clássicas pelos pacotes `survival`, `ggplot2` e `KMsurv`. As covariáveis contidas no banco de dados, e que serão analisadas, estão contidos na Tabela 1.

Tabela 1 – Covariáveis utilizadas na Análise de Sobrevivência.

Variáveis	**Descrição**	**Categorias**	**Tipo**
Tempo	Quantidade de dias do óbito ou censura	-	Quantitativo
Status	Indica se o paciente censurou ou falhou	0 = Censura 1 = Falha	Qualitativo
Gênero	Gênero dos pacientes que sofreram o infarto	0 = Masculino 1 = Feminino	Qualitativo
Idade	Idade dos pacientes que sofreram o câncer	-	Quantitativo
Escolaridade	Escolaridade do paciente	Fundamental I (1ª a 4ª série) Fundamental II (5ª a 8ª série) Médio Superior incompleto Superior completo Sem escolaridade Sem informação	Qualitativo
Morfologia	O estudo da forma do organismo ou de partes dele	Glioblastoma Astrocitoma Astrocitoma anaplastico Neoplasia Maligna Glioma maligno	Qualitativo
Meio de diagnóstico	Todos os testes que fornecem resultados que ajudam a estabelecer, confirmar ou excluir um diagnóstico	Citologia Clínico Histologia da metáfase Histologia do tumor primário Pesquisa SDO Sem informação	Qualitativo
Região	Regiões onde ocorreu ou não o evento	Centro-Oeste Nordeste Norte Sul Sudeste	Qualitativo
Raça/cor	Categorização social baseada na autodeclaração dos pacientes	Asiática/Branca Preta Parda Sem informação	Qualitativo

Fonte: Elaborada pelo autor, (2025).

4 RESULTADOS E DISCUSSÕES

A seguir, são apresentados os resultados obtidos a partir da aplicação do modelo *Random Survival Forest* (RSF) na predição da sobrevida de pacientes diagnosticados com câncer de cérebro, com o objetivo de avaliar seu desempenho frente ao modelo de Cox tradicional. Nesta seção, são discutidos os principais achados em termos de acurácia preditiva, importância das variáveis para o modelo e o comportamento dos diferentes métodos estatísticos utilizados, destacando as vantagens do RSF na análise de sobrevivência em contextos clínicos com dados complexos e censurados.

4.1 Análise exploratória

Inicialmente, em relação ao perfil sociodemográfico e clínico dos pacientes sob estudo, tem-se os resultados a seguir.

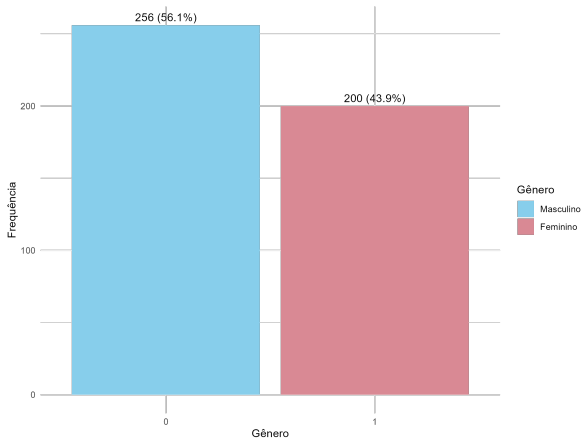
Tabela 2 – Estatísticas descritivas dos dados.

Variável	Mínimo	Média	Mediana	Máximo
Tempo	0,00	60,72	0,00	748,0
Idade	1,00	56,98	60,00	96,00

Fonte: Elaborada pelo autor, (2025).

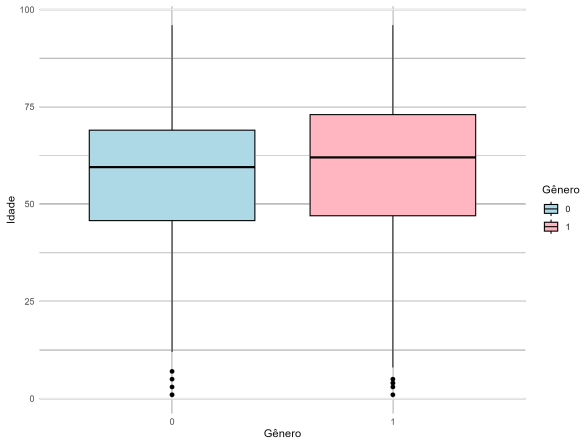
A partir da Tabela 2, nota-se que na idade dos pacientes há uma grande variabilidade entre a idade mínima e máxima em torno de uma idade média de quase 60 anos, com pacientes de 1 a 96 anos de idade, assim como no tempo de sobrevivência em que há pacientes que não resistiram sequer um dia, já outros tiveram um tempo de vida de 748 dias. Verifica-se também, que o primeiro paciente veio a óbito com pouco menos de 1 dia após o início do estudo, enquanto o último paciente que veio a óbito durante o estudo foi após o período de 2 anos.

Figura 4 – Quantidade de pacientes por gênero.



Fonte: Elaborado pelo autor, (2025).

Figura 5 – Faixa etária relacionada ao gênero dos pacientes.



Fonte: Elaborado pelo autor, (2025).

Por meio do Gráfico exibido na Figura 4, pode-se perceber que dos 456 pacientes incluídos no estudo, 256 são homens(aproximadamente 56%) e 200 são mulheres (cerca de 43%), resultando em um maior porcentual de pacientes do gênero masculino com casos de câncer de cérebro. Na Figura 5, observa-se que a demostração gráfica da idade mediana dos pacientes fica entre os 60 anos, independente do gênero.

Tabela 3 – Distribuição de frequências e tempo médio de sobrevivência das características sociodemográficas e clínicas dos pacientes com câncer de cérebro.

Variável	Categorias	Frequência (%)	Tempo médio (dias)
escolaridade	Sem informação	114 (25,00)	115
	Médio	103 (22,59)	49,6
	Fundamental I	86 (18,86)	41,9
	Fundamental II	68 (14,91)	53,3
	Superior Completo	57 (12,50)	77,7
	Sem Escolaridade	25 (5,48)	12,2
	Superior incompleto	3 (0,66)	Na
morfologia	Neoplasia Maligna	244 (53,51)	5,4
	Glioblastoma	120 (26,32)	189
	Glioma maligno	53 (11,62)	47,4
	Astrocitoma	24 (5,26)	187
	Astrocitoma anaplastico	15 (3,29)	243
status	1 = Óbito	271 (59,43)	57,5
	0 = Vivo/Censura	185 (40,57)	73,9
Meio.de.Diagnostico	Histologia	230 (50,44)	165
	SDO	188 (41,23)	0
	Outros métodos clínicos	37 (8,11)	25,5
	Sem informação	1 (0,22)	0
regiões	Sudeste	158 (34,65)	125
	Norte	117 (25,66)	37
	Centro-oeste	106 (23,25)	32,7
	Sul	40 (8,77)	30,9
	Nordeste	35 (7,68)	29,7
raça/cor	Parda	198 (43,42)	46
	Asiática/Branca	194 (42,54)	54
	Sem informação	42 (9,21)	190
	Preta	22 (4,82)	83,7

Fonte: Elaborado pelo autor, (2025).

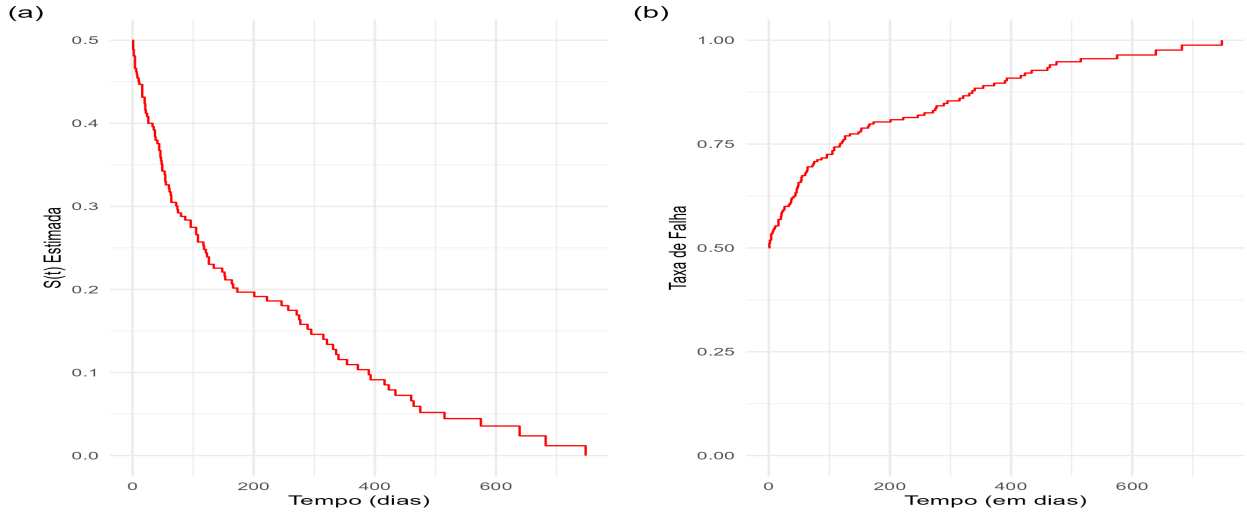
Na Tabela 3, observa-se que a maior categoria individualmente registrada em relação à escolaridade é “Sem informação” (25,0%), seguida pelo ensino médio completo (22,6%), e pelos ensinos fundamental I (18,9%) e fundamental II (14,9%). Essa elevada proporção de dados ausentes pode comprometer a análise do impacto da escolaridade na sobrevida dos pacientes. O tempo médio de sobrevivência varia significativamente entre as categorias, destacando-se a baixa sobrevivência dos pacientes sem escolaridade formal aproximadamente 12 dias, enquanto aqueles com ensino superior completo apresentam maior tempo médio perto dos 78 dias, sugerindo uma possível influência socioeconômica no prognóstico. Quanto à morfologia tumoral, a neoplasia maligna é o tipo mais frequente (53,5%), embora associada ao menor tempo médio de sobrevivência cerca de 5 dias, o que evidencia sua agressividade. O glioblastoma, o segundo tipo mais comum (26,3%), apresenta um tempo médio de sobrevida de 189 dias. Em contrapartida, o astrocitoma anaplásico, menos frequente (3,3%), possui o maior tempo médio de sobrevivência 243 dias, refletindo diferenças importantes nos perfis biológicos e prognósticos desses tumores. O status clínico indica uma alta taxa de óbitos, com 59,4% dos pacientes falecidos. O tempo médio de sobrevivência dos pacientes vivos ou censurados é de cerca 74 dias é superior ao dos pacientes óbitos de 57 dias, evidenciando a curta sobrevida geral da amostra. O principal meio de diagnóstico foi a histologia do tumor primário (50,4%), associada ao maior tempo médio de sobrevida de 165 dias. Em contrapartida, 41,2% dos pacientes foram diagnosticados por meio do Sistema de Dados Oncológicos (SDO), com tempo médio de sobrevivência igual a zero, o que pode indicar atrasos no diagnóstico, falhas na notificação ou dificuldades no acesso aos exames necessários. Em termos regionais, a maioria dos pacientes é oriunda das regiões Sudeste (34,7%), Norte (25,7%) e Centro-Oeste (23,3%). O tempo médio de sobrevivência é mais elevado no Sudeste com 125 dias em comparação

com as demais regiões, sendo menor no Nordeste de aproximadamente 30 dias, o que pode refletir desigualdades no acesso aos serviços de saúde e qualidade do atendimento. Quanto à raça/cor, predominam os pacientes pardos (43,4%) e asiáticos/brancos (42,5%). Interessantemente, os pacientes com dados ausentes para raça/cor apresentam o maior tempo médio de sobrevida de 190 dias, sugerindo possíveis vieses ou inconsistências na coleta desses dados. Pacientes pardos apresentam o menor tempo médio de sobrevivência com 46 dias, seguidos pelos asiáticos/brancos (54 dias) e pretos (83,7 dias), o que pode indicar disparidades raciais relacionadas ao prognóstico ou acesso ao tratamento. Por fim, a expressiva quantidade de dados ausentes em diversas variáveis reforça a necessidade de aprimoramento nos sistemas de registro e notificação, a fim de garantir análises mais robustas e fidedignas sobre o perfil e desfechos dos pacientes com câncer de cérebro.

4.1.1 *Estimativas de Kaplan-Meier*

Realizada a análise descritiva dos grupos, foram calculadas as estimativas de sobrevivência por meio do método Kaplan-Meier, os resultados obtidos foram os seguintes:

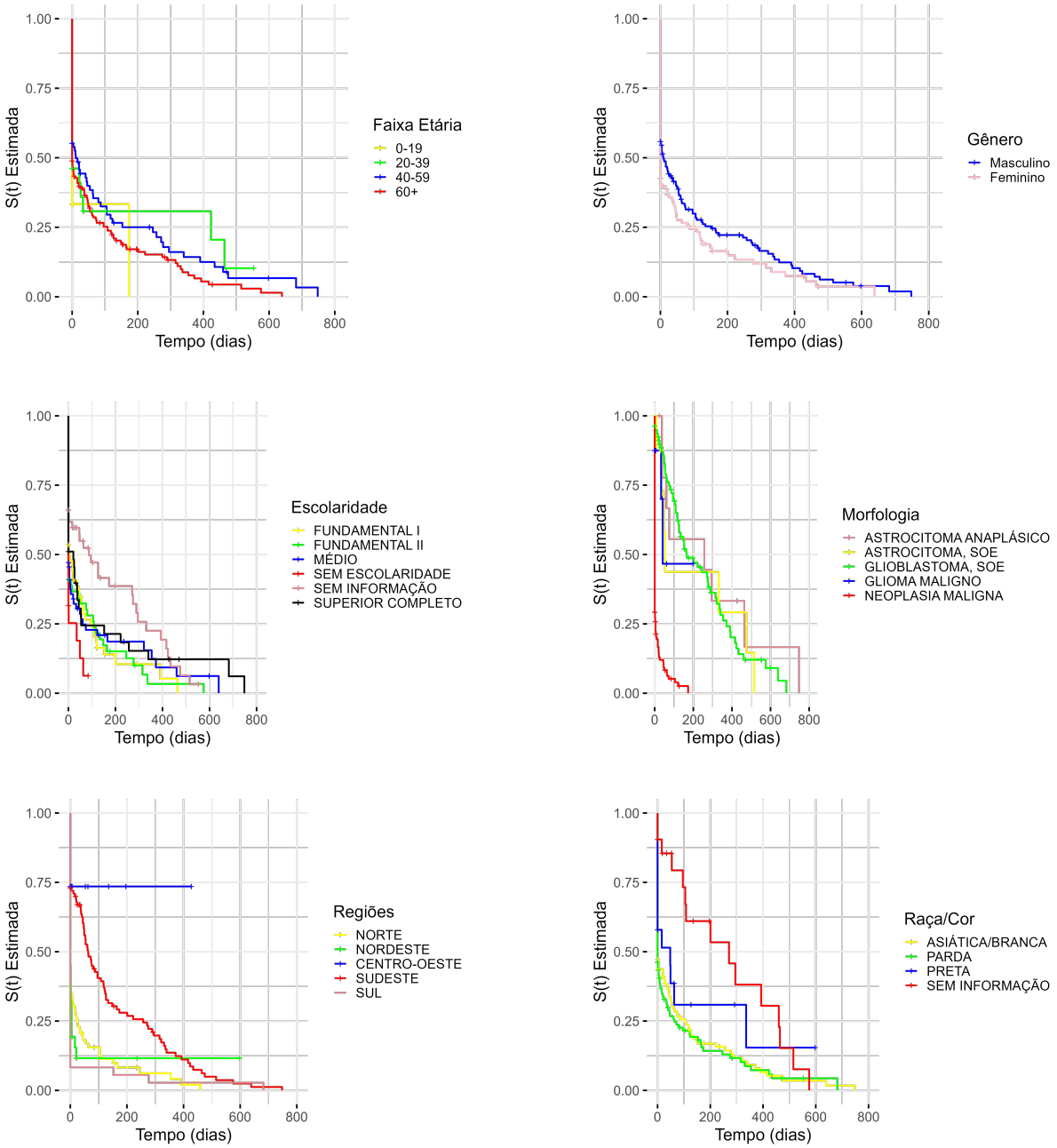
Figura 6 – Estimativas de Sobrevivência pelo método de Kaplan-Meier e a curva de risco de sobrevivência .



Fonte: Elaborada pelo autor, (2025).

A curva apresentada na Figura 6 (a) demonstra que as probabilidades de sobrevivência vão diminuindo ao longo do tempo, além disso, um ponto importante observado é que nos dias iniciais, ou seja, entre 0 e 200 dias, a queda dessa curva é bem significativa, isso pode indicar que esse intervalo de tempo é o mais critico para a sobrevivência dos pacientes que sofreram o câncer e, após esse período, a curva estabiliza, mas continua decrescendo, indicando que ainda há chances visíveis de falecimento de pacientes ao decorrer do tempo. Na Figura 6 (b), nota-se que, de fato, o risco de falecer devido a doença nos primeiros dias do diagnostico é bem maior e a taxa de falha aumenta ao longo do tempo, chegando próximo a 1 no final do acompanhamento (700 dias), o que indica que quase todos os pacientes vieram a falecer até o final do estudo, o que demonstra a severidade da doença.

Figura 7 – Curvas de Kaplan-Meier por categorias sociodemográficas e clínicas.



Fonte: Elaborada pelo autor, (2025).

Na Figura 7 observam-se as curvas de sobrevivência obtidas por meio do estimador de Kaplan-Meier para os grupos em estudo, em que é possível examinar a diferença do tempo de sobrevivência entre esses grupos. Nota-se na faixa etária, que pacientes com 60 anos ou mais apresentam um quadro maior no risco de óbito comparado com as outras faixas etárias. Quanto ao gênero feminino, este possui probabilidades menores de sobrevivência quando comparado o gênero masculino, confirmando a informação de que o risco do óbito é ainda maior para as mulheres. Em relação aos níveis de escolaridade, tem-se os pacientes que não possuem um grau de escolaridade com um pior diagnóstico de sobrevida comparado aos outros níveis, podendo se refletir desigualdades no acesso ao sistema de saúde, diagnóstico tardio ou adesão ao tratamento, como aponta Oliveira Santos *et al.* (2023). Na morfologia tumoral observa-se que a neoplasia maligna apresenta menor

probabilidade de sobrevida dos pacientes, além disso essa morfologia representa casos avançados do câncer apresentando chances pequenas de sobrevida.

Em relação as regiões, observa-se que pacientes residentes do centro-oeste apresentaram as maiores probabilidades de sobrevivência, com curvas mais sustentadas, o que pode estar relacionado à melhor oferta de serviços de saúde especializados ou maior acesso ao diagnóstico precoce e tratamento adequado. Em contrapartida, no sul apresentou-se quedas mais acentuadas nas estimativas de sobrevivência, sugerindo menor tempo de vida após o diagnóstico.

Pacientes da raça preta (linha azul) e sem informação racial (linha vermelha) apresentaram maior probabilidade de sobrevivência ao longo do tempo, com curvas de sobrevivência mais elevadas e sustentadas. Em contraste, pacientes pardos (linha verde) e asiáticos/branco (linha amarela) demonstraram menor tempo de sobrevida, com quedas mais acentuadas na estimativa de $S(t)$. Isso pode indicar desigualdades no acesso ao diagnóstico precoce, tratamento adequado ou suporte oncológico entre os diferentes grupos. Além disso, nota-se que a queda mais abrupta da curva ocorre nos primeiros 200 dias para todos os grupos, sugerindo uma alta mortalidade no início do acompanhamento. A persistência da curva em níveis baixos após esse período apontaram para uma baixa taxa de sobrevivência a longo prazo.

Realizadas estimativas de Kaplan-Meier, serão aplicados os testes de Logrank e Wilcoxon para verificar se há diferença significativa entre os grupos observados e se as distintas categorias dos mesmos influenciam as probabilidades de sobrevivência dos pacientes que sofreram o câncer.

Tabela 4 – Comparação entre os grupos observados pelos testes de Logrank e Wilcoxon.

Covariáveis	**Log-Rank**	**Wilcoxon**
gênero	0,0636	(< 0,05)*
Idade	0,0985	0,3340
morfologia	(< 0,05)*	(< 0,05)*
escolaridade	(< 0,05)*	(< 0,05)*
regiões	(< 0,05)*	(< 0,05)*
raça/cor	(< 0,05)*	(< 0,05)*

* Significativo a 5%.

Fonte: Elaborada pelo autor, (2025).

Com a tabela 4 é possível afirmar que em ambos os testes a morfologia, escolaridade, estados e raça/cor obtiveram diferença significativa entre os grupos. Com isso, os teste apontam indícios de que a morfologia, escolaridade, estados e raça/cor interfere de modo significativo nas probabilidades de sobrevivência dos pacientes com câncer cerebral. No gênero, observa-se que as diferenças na sobrevivência entre eles ocorrem principalmente nos tempos iniciais, uma vez que o teste de Wilcoxon é mais sensível a eventos precoces e aponta significância.

4.1.2 Ajuste do modelo de Cox

Para verificar a influência das covariávies observadas, será aplicado o modelo semi-paramétrico de Cox. Considerando então este modelo, encontrou-se os resultados apresentados na Tabela 5.

Tabela 5 – Estimativas do modelo de Cox clássico (M4) para câncer de cérebro.

Covariáveis	**Coef.**	**Exp(coef)**	**Erro-Padrão**	**p-valor**
GêneroFeminino	0,0145	1,0146	0,1575	0,9265
Idade	-0,0005	0,9995	0,0051	0,9228
Astrocitoma	-0,7219	0,4858	0,7695	0,3482
Glioblastoma	-0,2408	0,7860	0,5563	0,6652
Glioma Maligno	0,5717	1,7713	0,8053	0,4778
Neoplasia Maligna	2,0674	7,9045	0,5797	(<0,05)*
Fundamental II	0,4700	1,6000	0,2684	0,0799.
Médio	0,2859	1,3309	0,2410	0,2355
SemEscolaridade	-0,0086	0,9914	0,3597	0,9808
SemInformação	0,0410	1,0418	0,2803	0,8838
Superior Completo	-0,0717	0,9308	0,2950	0,8080
Nordeste	0,6464	1,9086	0,3022	(<0,05)*
Norte	0,4238	1,5278	0,2343	0,0705.
Sudeste	0,6141	1,8479	0,2467	(<0,05)*
Sul	0,3734	1,4527	0,2971	0,2088
Parda	-0,2311	0,7937	0,1771	0,1921
Preta	-0,4497	0,6379	0,4592	0,3275
SemInformação	-0,6123	0,5421	0,3858	0,1125

* Significativo a 5%.

Fonte: Elaborada pelo autor, (2025).

A Tabela 5 apresenta o resultado do modelo de Cox preliminar, identificando a significância de alguns fatores e outros não apresentaram significância à 5% ou 10%, como por exemplo, ter diagnóstico de neoplasia maligna aumenta o risco de morte em 7 vezes comparado ao grupo de referência, não só isso, mas também, as regiões nordeste e sudeste também mostraram associação estatisticamente significativa com o risco de óbito, com razões de risco de 1,91 e 1,85, respectivamente, indicando que a localização geográfica pode influenciar o prognóstico dos pacientes. As demais variáveis não apresentam significância estatística ao nível de 5%, embora algumas tenham risco elevados que podem ser clinicamente relevantes, mas estatisticamente só foram significativas à 10%, então, o critério de seleção de variáveis será aplicado, a fim de encontrar a melhor modelagem possível de ser ajustada aos dados. Após a aplicação do Critério de Informação de Akaike (AIC) via *stepwise*, obteve-se o melhor modelo.

Tabela 6 – Modelo final de Cox via AIC para câncer de cérebro.

Idade	-0,0008	0,9991	0,0045	0,8475
Astrocitoma	-0,8362	0,4333	0,7060	0,2362
Glioblastoma	-0,2762	0,7587	0,5317	0,6035
Glioma Maligno	0,5075	1,6611	0,7771	0,5137
Neoplasia Maligna	1,8500	6,3598	0,5447	(<0,05)*
Nordeste	0,6601	1,9349	0,2950	(<0,05)*
Norte	0,3914	1,4791	0,2247	0,0815.
Sudeste	0,5681	1,7649	0,2377	(<0,05)*
Sul	0,5400	1,7160	0,2840	0,0572.

* Significativo a 5%.

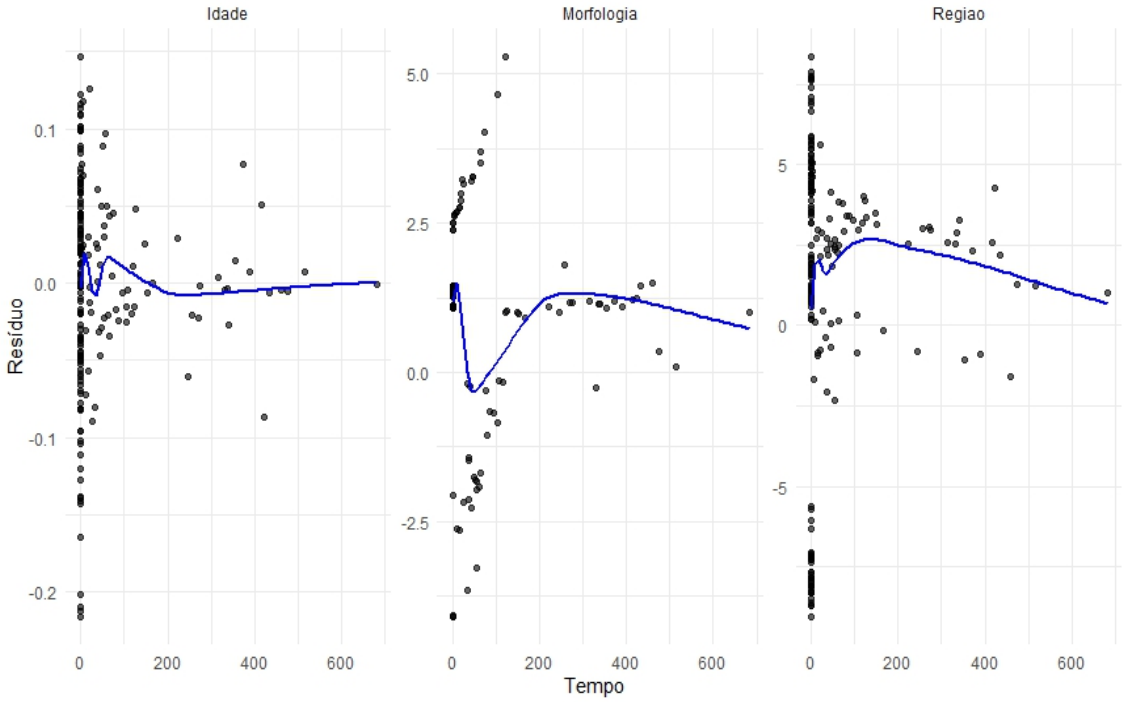
Fonte: Elaborada pelo autor, (2025).

Conforme apresentado na Tabela 6, percebe-se que três variáveis foram significativas ou marginalmente significativas para o modelo, ao nível de 5%, respectivamente, sendo a

neoplasia maligna, a região nordeste e sudeste abaixo de 0,05, pois conforme avança o tempo de vida do paciente mais aumenta o risco de morte relacionada a essas variáveis, como mostrou a Figura 7. Pacientes com neoplasia maligna apresentaram 6,36 vezes mais risco de morte em comparação com as outras morfologias. A variável como idade foi selecionada pelo Critério de Informação de Akaike (AIC) e por fazer parte do efeito de interação, optou-se assim por mantê-la.

Após o ajuste do modelo, seguiu para a etapa de análise dos resíduos, com o objetivo de investigar a consistência nas estimativas dos coeficientes do modelo ajustado. Dessa forma, na Figura 8 tem-se os resíduos de Schoenfeld obtidos para cada uma das covariáveis abordadas no modelo.

Figura 8 – Gráfico dos resíduos de Shoenfeld para os dados de câncer no cérebro.



Fonte: Elaborada pelo autor, (2025).

Os resíduos de Shoenfeld servem para verificar a proporcionalidade dos riscos entre os grupos. Por meio da análise da Figura 8 nota-se que, para a variável Idade, os resíduos se distribuem de forma relativamente constante ao longo do tempo, o que sugere que a suposição de proporcionalidade é atendida. No entanto, para as variáveis Morfologia e Região, há variações mais acentuadas nos resíduos, especialmente nos períodos iniciais e finais, o que indica uma possível violação da suposição de proporcionalidade dos riscos. Assim, enquanto a variável Idade parece estar de acordo com os pressupostos do modelo, Morfologia e Região. Espera-se que haja uma linha reta em torno do zero indicando proporcionalidade entre as curvas de risco, conclui-se que através dessa complexidade o modelo de RSF pode-se ser um aliado forte, pelo fato que ele faz essas interações mais precisas.

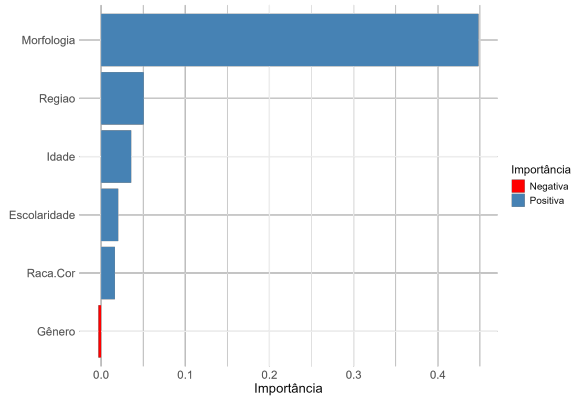
4.1.3 Ajuste do Random Survival Forest (RSF)

Para realizar a predição selecionou-se o melhor modelo ajustado via riscos proporcionais e denominou-se de Modelo 2, e realizou-se o ajuste via *Random Forest* para análise de sobrevivência (RSF). Os principais resultados são vistos nas Figuras de 9, bem como pelo resultado da comparação entre os modelos na Tabela 7. De inicio, foi ajustado o RSF, com todas as covariáveis para 1500 amostras *bootstrap* e calculou-se a taxa de erro e o fator de importância de variância (*Erro Rate and Variable importance*; Figuras 9 e 10).

Figura 9 – Taxa de erro.

Fonte: Elaborada pelo autor, (2025).

Figura 10 – Fator de importância da variável.



Fonte: Elaborada pelo autor, (2025).

Foi verificado, por meio da Figura 9, que há uma estabilização na taxa de erro de previsão com os dados OOB por volta de 21% a partir de 500 arvores de decisão. Nesse conjunto de dados foi realizado o crescimento de 1500 arvores na floresta aleatória, de modo que obteve-se convergência nessa simulação. Na Figura 10, percebe-se também que o gênero apresenta contribuição baixa e negativa para a discriminação dos sujeitos dentro da árvore, fazendo com que sejam geradas diferentes configurações de RSF: com esta variável e sem esta variável.

Tabela 7 – Desempenho de diferentes configurações RFS e modelo de Cox clássico utilizando o conjunto de dados de treino e teste.

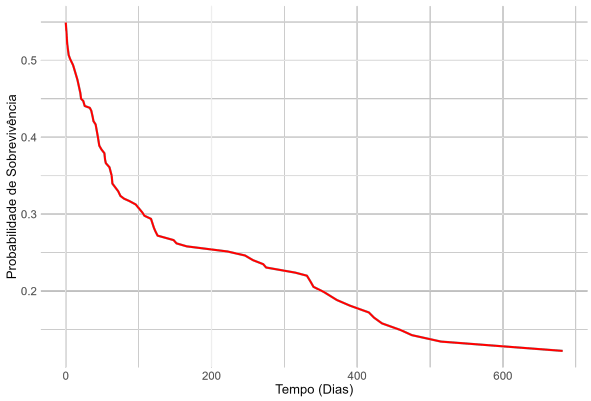
Modelos Analisados	**C-index**		
	Treino	**Teste**	% de variação
M1: RSF - Completo	0,7878	0,7527	-4,46
M2: RSF - COX	0,8198	0,8312	+1,39
M3: RSF - Corrente	0,7781	0,7556	-2,89
M4: Cox - Clássico	0,7967	0,8012	+0,56

Fonte: Elaborada pelo autor, (2025).

A Tabela 7 apresenta os valores do índice de concordância (C-index) obtidos pelos modelos avaliados nos conjuntos de treino e teste. Observa-se que o modelo RSF com seleção de variáveis via modelo de Cox (M2) alcançou o melhor desempenho preditivo, com C-index de 0,8312 no conjunto de teste, superando inclusive o valor obtido no treino

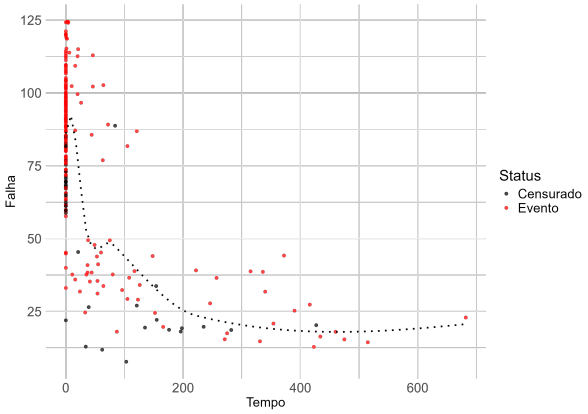
(0,8198), o que pode indicar uma boa capacidade de generalização e ausência de sobreajuste. Por outro lado, o modelo clássico de Cox (M4) também demonstrou desempenho bastante satisfatório, com C-index de 0,8012 no teste. Em contraste, os modelos M1 (RSF completo) e M3 (RSF com covariáveis correntes) apresentaram quedas acentuadas de desempenho no conjunto de teste, o que sugere possível *overfitting*. Com isso, pode-se concluir que, embora o modelo de Cox clássico tenha apresentado uma proximidade entre o M2 no teste e tenha tido uma pequena variação, o modelo RSF com seleção baseada em Cox via AIC se mostrou uma alternativa robusta e com uma melhor capacidade de generalização.

Figura 11 – Curva de sobrevivência predita.



Fonte: Elaborada pelo autor, (2025).

Figura 12 – Gráfico da falha *versus* tempo.



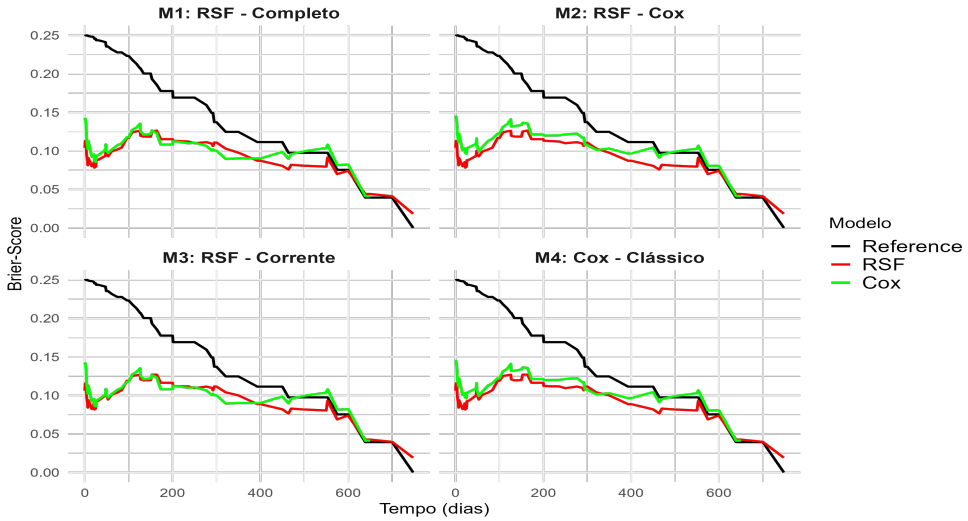
Fonte: Elaborada pelo autor, (2025).

As Figuras 11 e 12 apresentam, respectivamente, a curva de sobrevivência predita com base no conjunto de teste e a taxa de falha ao longo do tempo. Observa-se, por meio dessas representações, o comportamento individual de cada paciente em teste e sua respectiva estimativa de sobrevida ao longo do tempo. Nota-se uma maior concentração de predições de falha no intervalo entre 0 e 200 dias, o que evidencia a boa capacidade do modelo RSF em discriminar entre pacientes censurados e aqueles que evoluíram para a morte, especialmente nesse período.

4.1.4 Comparação dos modelos

A seguir, foi realizada uma análise comparativa entre diferentes métodos de modelagem de sobrevivência, sendo uma etapa essencial para avaliar o desempenho preditivo e a capacidade de cada abordagem em representar a realidade clínica dos pacientes. Aqui, são comparados quatro modelos: de *Random Survival Forest* (RSF) (M1,M2,M3) *versus* o modelo de Cox (N4). Cada um desses métodos possui características distintas quanto à forma de lidar com variáveis explicativas, suposições estatísticas e capacidade de capturar relações não lineares ou interações complexas.

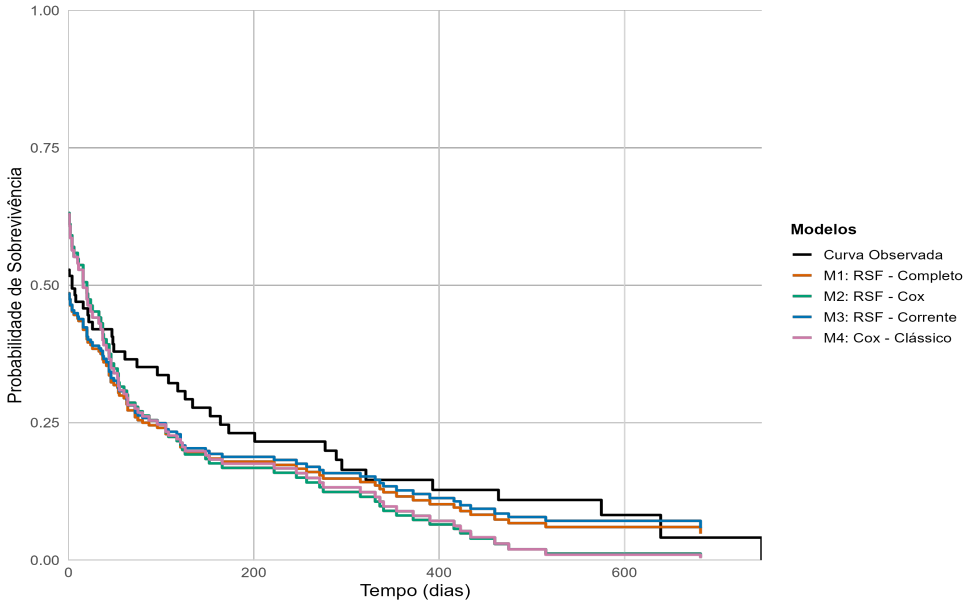
Figura 13 – Curvas do erro de previsão ao longo do tempo.



Fonte: Elaborada pelo autor, (2025).

A Figura 13 apresenta as curvas do erro de previsão ao longo do tempo, obtidas por meio do *Brier Score* para o modelo de Cox (M4) e *Random Survival Forest* (RSF) (M1,M2,M3), bem como a curva de referência (modelo nulo). Observa-se que ambas as abordagens — Cox e RSF — apresentaram desempenho preditivo superior ao da referência, que reflete um modelo sem informações preditoras (linha preta). Comparando os dois modelos, nota-se que a curva do RSF (linha vermelha) se mantém consistentemente abaixo ou muito próxima da curva do modelo de Cox (linha verde) ao longo de todo o período analisado, indicando menor erro de previsão. Isso sugere que o RSF apresenta ligeira vantagem na acurácia preditiva, sobretudo nos períodos iniciais, onde os erros são menores. De forma geral, ambos os modelos demonstram boa performance preditiva, porém, o RSF destaca-se por apresentar erros ligeiramente inferiores, corroborando os resultados obtidos nas métricas de C-index.

Figura 14 – Comparação de Curvas de Sobrevivência entre modelos.



Fonte: Elaborada pelo autor, (2025).

A Figura 14 apresenta as curvas de sobrevivência estimadas pelos quatro modelos e a sobrevivência da base de teste estimada pelo Kaplan-Meier para os pacientes com câncer de cérebro. Observa-se que os modelos baseados em *Random Survival Forest* (M1, M2 e M3) apresentam maior aderência à curva observada ao longo de todo o período analisado, especialmente nos primeiros 400 dias, indicando melhor capacidade preditiva em relação ao modelo de Cox. Entre os modelos RSF, as curvas são bastante próximas entre si, sugerindo robustez do algoritmo mesmo sob diferentes estratégias de seleção de variáveis. Por outro lado, o modelo de regressão de Cox clássico (M4) tende a subestimar a probabilidade de sobrevivência nos períodos mais longos, o que indica possível violação da suposição de proporcionalidade dos riscos. Assim, conclui-se que os modelos baseados em RSF apresentam desempenho superior na predição da sobrevida de pacientes com câncer cerebral, mostrando-se mais adequados para lidar com a complexidade dos dados e com possíveis interações e não linearidades presentes nas covariáveis.

5 CONCLUSÃO

Este estudo teve como principal objetivo comparar o desempenho de três abordagens distintas para análise de sobrevivência — Kaplan-Meier, modelo de Cox e *Random Survival Forest* (RSF) do tempo até o óbito em pacientes diagnosticados com câncer de cérebro. Utilizando um conjunto de dados composto por 456 pacientes oriundos de diferentes regiões do Brasil, a análise buscou identificar qual metodologia apresenta maior capacidade preditiva e utilidade prática no auxílio ao prognóstico clínico.

Inicialmente, o método Kaplan-Meier permitiu a visualização não paramétrica das curvas de sobrevivência globais e estratificadas, fornecendo uma visão geral da evolução do risco ao longo do tempo. Esta abordagem demonstrou sua limitação na incorporação de múltiplas covariáveis e na modelagem de interações entre variáveis, servindo, contudo, como uma ferramenta exploratória inicial útil.

O modelo semi-paramétrico de Cox, possibilitou a incorporação de covariáveis como idade, gênero, morfologia do tumor, escolaridade, estado e raça/cor. A análise evidenciou que a morfologia tumoral (com destaque para neoplasias malignas e gliomas malignos) foram fatores significativamente associados ao risco de óbito. A performance do modelo de Cox clássico foi inferior ao RSF no conjunto de teste, apresentando um C-index menor, entretanto, ele indicou que o modelo aprendeu padrões reais dos dados, sem se prender excessivamente às particularidades do conjunto de treino. Visto que, aparentemente ele tenha violado os pressupostos da proporcionalidade dos riscos.

Já o RSF, demonstrou superioridade com a seleção de variáveis via modelo de Cox. Por meio do uso de múltiplas árvores de decisão e validação com dados *Out-Of-Bag*, o RSF apresentou maior capacidade de capturar relações não lineares e interações complexas entre variáveis. O modelo M2 alcançou o melhor desempenho preditivo, com C-index superior tanto no conjunto de treino quanto no de teste, além de uma variação percentual de erro baixa (1,39%). Tais resultados indicam que o RSF não só fornece uma previsão mais acurada, como também maior estabilidade ao ser aplicado em novos conjuntos de dados.

A análise da importância das variáveis no RSF destacou que variáveis como morfologia, estado e idade foram determinantes na predição da sobrevida, enquanto outras, como o gênero, apresentaram baixa ou até mesmo contribuição negativa, reforçando a necessidade de um processo seletivo rigoroso das covariáveis em modelos clínicos preditivos.

Portanto, pode-se concluir que, embora os métodos tradicionais (como Cox e

Kaplan-Meier) mantenham relevância e utilidade prática na análise de sobrevivência, o RSF, oferecem vantagens consideráveis quando o objetivo é maximizar o poder preditivo e a acurácia, especialmente em bases de dados com múltiplas covariáveis e complexidade estrutural. Os resultados deste estudo sugerem que o RSF representa uma alternativa promissora na prática médica, podendo auxiliar de forma mais eficaz na definição de estratégias terapêuticas e no acompanhamento clínico personalizado de pacientes com câncer de cérebro.

REFERÊNCIAS

AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974. DOI: 10.1109/TAC.1974.1100705. Citado na página 17.

BOU-HAMAD, Imad; LAROCQUE, Denis; BEN-AMEUR, Hatem. A review of survival trees, 2011. Citado na página 23.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, p. 5–32, 1986. Citado nas páginas 11, 19, 21, 22.

CARVALHO, Marilia Sá; ANDREOZZI, Valeska Lima; CODEÇO, Claudia Torres; BARBOSA, Maria Tereza Serrano; SHIMAKURA, Silvia Emiko. Análise de sobrevida: teoria e aplicações em saúde. *In:* ANÁLISE de sobrevida: teoria e aplicações em saúde. [*S. l.*]: Editora Fiocruz, 2005. p. 395–395. Citado nas páginas 13, 14.

CELESTINO, Matheus Jubini; PEREIRA, Paula Eduarda Mercier; MELLO, Raissa Santos; JANUÁRIO, Pedro Ramos; RONCHI, Silas Nascimento. NEOPLASIA MALIGNA DO ENCÉFALO: PERFIL EPIDEMIOLÓGICO NO BRASIL EM 2023. **RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218**, v. 5, n. 10, e5105822–e5105822, 2024. Citado na página 10.

COLOSIMO, Enrico Antonio; GIOLO, Suely Ruiz. **Análise de sobrevivência aplicada**. 2. ed. [*S. l.*]: Editora Blucher, 2021. Citado nas páginas 11, 12.

COX, David R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado nas páginas 15, 16.

COX, David R.; OAKES, David. **Analysis of Survival Data**. [*S. l.*]: Chapman e Hall, 1984. Citado na página 16.

EHRLINGER, Justin. ggRandomForests: Exploring random forest survival. **The Annals of Applied Statistics**, v. 1, p. 1–41, 2016.

GEHAN, Edmund A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. **Biometrika**, Oxford University Press, v. 52, n. 1-2, p. 203–224, 1965. Citado na página 14.

GRAF, Erika; SCHMOOR, Claudia; SAUERBREI, Willi; SCHUMACHER, Martin. Assessment and comparison of prognostic classification schemes for survival data. **Statistics in medicine**, Wiley Online Library, v. 18, n. 17-18, p. 2529–2545, 1999. Citado na página 24.

HASTIE, Trevor; TIBSHIRANI, Robert. **Generalized Additive Models**. London: Chapman e Hall/CRC, 1990. Citado na página 18.

INSTITUTO NACIONAL DE CÂNCER. **Estimativa 2021: Incidência de Câncer no Brasil**. [*S. l.: s. n.*], 2021. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros. Citado na página 25.

INSTITUTO VENCER O CÂNCER. **Câncer de cérebro | O que é?** [*S. l.: s. n.*], 2025. Acesso em: 5 maio 2025. Disponível em: https://vencerocancer.org.br/tipos-de-cancer/cancer-de-cerebro-o-que-e/. Citado na página 10.

ISHWARAN, Hemant; KOGALUR, Uday B. Random survival forests for R. **R News**, v. 7, n. 2, p. 25–31, 2007. Citado nas páginas 11, 23.

JAMES, Joby; SANDHYA, L.; THOMAS, Ciza. Detection of phishing URLs using machine learning techniques. *In:* 2013 International Conference on Control Communication and Computing (ICCC). [*S. l.*]: IEEE, 2013. p. 304–309. Citado na página 18.

JIANG, Chen; WANG, Kan; YAN, Lizhao; YAO, Hailing; SHI, Huiying; LIN, Rong. Predicting the survival of patients with pancreatic neuroendocrine neoplasms using deep learning: A study based on Surveillance, Epidemiology, and End Results database. **Cancer Medicine**, Wiley Online Library, v. 12, n. 11, p. 12413–12424, 2023. Citado na página 10.

KAPLAN, Edward L.; MEIER, Paul. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 13.

KLEIN, John P.; MOESCHBERGER, Melvin L. **Survival Analysis: Techniques for Censored and Truncated Data**. 2. ed. [*S. l.*]: Springer, 2003. Citado nas páginas 15, 16.

KLEINBAUM, David G.; KLEIN, Mitchel. **Survival Analysis: A Self-Learning Text**. 3rd. [*S. l.*]: Springer, 2012. DOI: 10.1007/978-1-4419-6646-9. Citado nas páginas 11, 13.

KULLBACK, Solomon; LEIBLER, Richard A. On Information and Sufficiency. **The Annals of Mathematical Statistics**, v. 22, n. 1, p. 79–86, 1951. DOI: 10.1214/aoms/1177729694. Citado na página 17.

LUNETTA, Philippe; SMITH, Gordon S. The Role of Alcohol in Injury Deaths. *In:* CHERPITEL, M. G.; BORGES, G.; GIESBRECHT, N. (ed.). **Alcohol and Injuries: Emergency Department Studies in an International Perspective**. [*S. l.*]: World Health Organization, 2004. p. 13–26. Citado na página 11.

MACHADO, Angelo; HAERTEL, Lucia Machado. **Neuroanatomia funcional**. [*S. l.*]: Atheneu, 2013. Citado na página 10.

MANTEL, Nathan; HAENSZEL, William. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemother Rep**, v. 50, n. 3, p. 163–170, 1966. Citado na página 14.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997. Citado na página 18.

MORGAN, J. N.; SONQUIST, J. A. Problems in the Analysis of Survey Data, and a Proposal. **Journal of the American Statistical Association**, v. 58, n. 302, p. 415–434, 1963. DOI: 10.1080/01621459.1963.10500855. Citado na página 22.

OLIVEIRA SANTOS, Marceli de; LIMA, Fernanda Cristina da Silva de; MARTINS, Luís Felipe Leite; OLIVEIRA, Julio Fernando Pinto; ALMEIDA, Liz Maria de; CAMARGO CANCELA, Marianna de. Estimativa de incidência de câncer no Brasil, 2023-2025. **Revista Brasileira de Cancerologia**, v. 69, n. 1, 2023. Citado na página 31.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2025. http://www.R-project.org/. ISBN 3-900051-07-0. Citado na página 23.

SUGIURA, Nariaki. Further analysts of the data by Akaike's information criterion and the finite corrections. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 7, n. 1, p. 13–26, 1978. Citado na página 17.

TECHNOLOGIES, Dimensionless. **Introduction to Random Forest**. Acesso em: 2 jun. 2025. 2020. Disponível em: https://dimensionless.in/introduction-to-random-forest/. Acesso em: 2 jun. 2025. Citado na página 22.

TORRES, Vagner dos Santos. **Resiliência no mercado de trabalho formal no Nordeste brasileiro: o uso da análise de sobrevivência para estimar a permanência nos postos de trabalho nos anos de 2019 e 2021**. 2024. Dissertação de Mestrado – Universidade Federal do Rio Grande do Norte, Natal, Brasil. Acesso em: 5 maio 2025. Disponível em: https://repositorio.ufrn.br/handle/123456789/60076. Citado na página 11.

VAPNIK, Vladimir N. **The Nature of Statistical Learning Theory**. New York: Springer, 1995. Citado na página 19.

WELLER, Michael; BENT, Martin van den; HOPKINS, Kelly; TONN, Joerg-Christian; STUPP, Roger; FALINI, Andrea; REIFENBERGER, Guido. EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. **The Lancet Oncology**, v. 15, n. 9, e395–e403, 2015. DOI: 10.1016/S1470-2045(14)70379-1. Citado na página 10.

WU, Jiajun; YILDIRIM, Ilker; LIM, Joseph J; FREEMAN, Bill; TENENBAUM, Josh. Galileo: Perceiving physical object properties by integrating a physics engine with deep

learning. **Advances in neural information processing systems**, v. 28, 2015. Citado na página 21.

ZHU, Ji; HASTIE, Trevor. Kernel logistic regression and the import vector machine. *In:* ADVANCES in Neural Information Processing Systems 14. [*S. l.: s. n.*], 2001. Citado na página 18.