Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estat´stica

Ajuste de modelos de Cox e de florestas aleatórias em análise de sobrevivência

Mathews Siqueira Melo

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estat´stica

Ajuste de modelos de Cox e de orestas aleatórias em análise de sobrevivência

Mathews Siqueira Melo
Orientadora: Teresa Cristina Martins Dias

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Resumo

Neste trabalho de conclusão de curso é explorada a aplicação e a comparação de métodos estatísticos clássicos e de aprendizado de máquina no contexto da análise de sobrevivência. Foi realizada uma revisão teórica detalhada dos fundamentos da análise de dados de tempo até o evento, abordando conceitos como censura, as funções de sobrevivência e de risco, o estimador não-paramétrico de Kaplan-Meier e o modelo semi-paramétrico de riscos proporcionais de Cox. Em paralelo, foram estudados os princípios do aprendizado de máquina, com foco na estrutura das árvores de decisão e na robustez dos métodos de *ensemble*, como as florestas aleatórias. Para avaliar a performance dos modelos, foram utilizadas métricas consolidadas, incluindo o índice C para discriminação, o escore de Brier para acurácia probabilística e métricas derivadas da matriz de confusão, como acurácia, sensibilidade e especificidade. A aplicação prática das metodologias foi realizada através de dois estudos de caso: um primeiro, com dados reais do estudo clínico `veteran`, ilustrou o fluxo de trabalho para o ajuste e validação de um modelo de Cox; um segundo, com dados simulados, permitiu uma comparação direta da performance preditiva entre uma árvore de decisão otimizada e uma floresta aleatória em um cenário controlado. Os resultados destacam o contraste entre a interpretabilidade dos modelos clássicos e o desempenho preditivo dos modelos de aprendizado de máquina, discutindo os cenários em que cada abordagem pode ser mais vantajosa.

Palavras-Chave: Aprendizado de máquina, árvore de decisão, floresta aleatória, função de risco, função de sobrevivência, Kaplan-Meier, modelo de Cox.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise de dados de tempo até o evento, mais conhecida como análise de sobrevivência, constitui um ramo fundamental da Estatística, com aplicações que se estendem por diversas áreas do conhecimento. Desde a avaliação da eficácia de novos tratamentos em ensaios clínicos na medicina, passando pela análise de confiabilidade de componentes na engenharia, até a modelagem de risco de crédito em finanças, a capacidade de modelar e prever o tempo até a ocorrência de um evento de interesse é de suma importância. A característica que define e distingue a área de análise de sobrevivência é a sua capacidade de lidar com dados censurados, observações para as quais o evento de interesse não ocorreu durante o período de acompanhamento, fornecendo apenas uma informação parcial sobre o tempo de vida do indivíduo ou unidade de estudo.

Historicamente, o campo da análise de sobrevivência foi consolidado sobre pilares metodológicos robustos. O estimador não-paramétrico de Kaplan-Meier (Kaplan e Meier, 1958) é uma das ferramentas mais utilizadas para estimar e visualizar a função de sobrevivência a partir de dados censurados à direita em análise de sobrevivência. Para investigar o efeito de covariáveis sobre o risco, o modelo de riscos proporcionais de Cox (Cox, 1972) tornou-se uma abordagem semiparamétrica bastante utilizada. Sua principal característica reside na capacidade de estimar o efeito de múltiplos preditores, expresso através de razões de risco (*hazard ratios*) e sem a necessidade de especificar a forma da função de risco base. A interpretabilidade e a solidez teórica do modelo de Cox garantiram sua posição como a técnica dominante na área por décadas. É importante notar que, embora sua forma semiparamétrica seja a mais difundida, o modelo de Cox também pode ser estendido para um contexto totalmente paramétrico, em que é assumida uma distribuição de probabilidade específica para a função de risco base.

Contudo, a crescente complexidade e dimensionalidade dos conjuntos de dados modernos trouxeram à tona as limitações inerentes a estes modelos clássicos. A suposição de riscos proporcionais, fundamental para a validade do modelo de Cox, é frequentemente violada na prática. Além disso, a estrutura linear do preditor no modelo de Cox pode não ser capaz de capturar relações não lineares complexas ou interações de alta ordem entre as covariáveis, que são cada vez mais comuns em dados genômicos ou de sistemas complexos.

Neste contexto, o campo de *machine learning* (aprendizado de máquina) surge como uma alternativa poderosa, oferecendo métodos flexíveis e com alto desempenho preditivo. Entre eles, os modelos baseados em árvores de decisão se destacam. Embora uma única árvore de decisão seja intuitiva e de fácil interpretação, ela é notória por sua instabilidade e tendência ao sobreajuste (Izbicki e dos Santos, 2020). Para mitigar essas desvantagens, foram desenvolvidos os métodos de *ensemble*, como as florestas aleatórias, propostas por Breiman (Breiman, 2001). Ao agregar as predições de um grande número de árvores não correlacionadas, as florestas aleatórias alcançam uma notável robustez e precisão, tornando-se uma das técnicas mais eficazes no arsenal do aprendizado de máquina supervisionado.

O objetivo do estudo está posicionado na interseção entre a análise de sobrevivência clássica e as abordagens modernas de aprendizado de máquina. O objetivo central é explorar e comparar o desempenho preditivo do tradicional modelo de riscos proporcionais de Cox com uma adaptação das florestas aleatórias para dados de tempo até o evento, conhecida como florestas aleatórias de sobrevivência (*random survival forests*). Assim, tendo como objetivo entender e investigar em quais condições os modelos de aprendizado de máquina podem oferecer uma alternativa superior aos modelos estatísticos clássicos para a predição de desfechos de sobrevivência, focando no equilíbrio entre a acurácia preditiva e a interpretabilidade dos modelos.

Para atingir este objetivo, neste trabalho foi seguida uma abordagem metodológica estruturada. Foram abordados de forma detalhada os fundamentos teóricos dos métodos clássicos de análise de sobrevivência e dos modelos de aprendizado de máquina, com foco no modelo de Cox e nas florestas aleatórias, respectivamente.

Este trabalho está estruturado de forma de que no capítulo 2 são apresentados os fundamentos teóricos da análise de sobrevivência, incluindo os conceitos de censura, as funções de risco e sobrevivência, o estimador de Kaplan-Meier e o modelo de riscos pro-

porcionais de Cox, culminando com uma aplicação prática. No capítulo 3 são introduzidos os conceitos de aprendizado de máquina, detalhando as árvores de decisão, as florestas aleatórias e as métricas de performance utilizadas, e finaliza com um estudo de simulação comparativo. Já nos dois últimos capítulos, introduzimos florestas aleatórias de sobrevivência, explicando seus conceitos detalhadamente, com exemplos e depois uma aplicação em dados reais. Por fim, são apresentadas as conclusões do estudo, sintetizando os resultados comparativos e discutindo as implicações das duas abordagens de modelagem.

Capítulo 2

Análise de sobrevivência

Neste capítulo será apresentada a teoria de análise de sobrevivência. Na seção 2.1 foram apresentados os conceitos básicos e as principais fórmulas a serem utilizados. Nas seções 2.2 e 2.3 foram apresentadas as técnicas que foram utilizadas no estudo, suas principais características e suas utilidades. Por fim, na seção 2.4 foi apresentado um exemplo aplicando as técnicas mostradas neste capítulo.

2.1 Conceitos básicos

A análise de sobrevivência é fundamentada em conceitos que permitem modelar o tempo até a ocorrência de um evento de interesse. Seja T uma variável aleatória contínua e não negativa. A distribuição de T pode ser caracterizada por sua função de densidade de probabilidade, denotada por $f(t)$ que descreve a probabilidade relativa da ocorrência do evento até um tempo específico t, de modo que a probabilidade de o evento ocorrer em um intervalo $[0, t]$ é dada pela integral:

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du.$$

A partir da função de densidade, define-se o conceito central da análise de sobrevivência: a função de sobrevivência, $S(t)$. Ela representa a probabilidade de o evento de interesse ainda não ter ocorrido até o tempo t, ou seja, a probabilidade do tempo até o evento ser maior que t.

$$S(t) = P(T > t) = 1 \quad F(T) = \int_t^\infty f(u)\, du. \tag{2.1}$$

Por definição, $S(t)$ é uma função não crescente, com $S(0) = 1$ e $S(\infty) = 0$.

Consequentemente, a relação entre as duas funções pode ser invertida. A função de densidade $f(t)$, pode ser obtida a partir da função de sobrevivência através da derivada negativa:

$$f(t) = \quad \frac{d}{dt} S(t). \tag{2.2}$$

Embora a função de densidade descreva o comportamento das ocorrências de eventos na população, muitas vezes o interesse está na probabilidade de um evento ocorrer em um instante t, dado que a unidade amostral não sofreu o evento até aquele momento (Colosimo e Giolo, 2006). Esse conceito é capturado pela função de risco (ou taxa de falha), denotada por $h(t)$. Esta é definida como a taxa instantânea de ocorrência do evento, condicionada à sobrevivência até o tempo t:

$$h(t) = \lim_{t \to 0} \frac{P(t \leq T < t + \quad t \mid T \geq t)}{t}. \tag{2.3}$$

A função de risco conecta os dois conceitos anteriores, pois pode ser expressa como a razão entre a função de densidade e a função de sobrevivência:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.4}$$

A relação expressa na Equação (2.4) é conceitualmente intuitiva, pois o risco instantâneo de um evento ocorrer, $\lambda(t)$, é definido pela taxa de ocorrência de eventos, $f(t)$ equação (2.2), normalizada pela proporção de unidades de estudo que ainda estão sob risco naquele instante, $S(t)$ equação (2.1).

Por fim, integrando a função de risco do início do estudo até um tempo t, obtemos a função de risco acumulada, $H(t)$:

$$H(t) = \int_0^t \lambda(u)\, du. \tag{2.5}$$

A função de risco acumulada quantifica o risco total acumulado até o tempo t. Ela possui uma relação direta e fundamental com a função de sobrevivência, que é:

$$S(t) = \exp\left(\quad H(t)\right). \tag{2.6}$$

Outro conceito importante para a análise de sobrevivência e distingue de outros

métodos estatísticos é a sua capacidade de lidar com dados incompletos, que é conhecido como censura. A censura ocorre em caso de não se observar o tempo exato até a ocorrência do evento de interesse para uma ou mais unidades de estudo, resultando em uma observação parcial do tempo. O tipo mais comum é a censura à direita, que acontece no instante em que o acompanhamento de uma unidade termina antes que o evento tenha ocorrido. Outras formas incluem a censura à esquerda, caso o evento já tenha ocorrido antes do início da observação, e a censura intervalar, em caso de que o evento ocorra dentro de um intervalo de tempo. Ignorar ou tratar incorretamente estas observações causaria um viés significativo nas estimativas. Portanto, os métodos de análise de sobrevivência foram e ainda são desenvolvidos para incorporar a informação parcial contida nos dados censurados. Para a censura não informativa a ideia central é de que o mecanismo que levou à censura seja independente do risco de ocorrência do evento. Se a censura for informativa, a probabilidade de censura está correlacionada com o desfecho, e técnicas mais avançadas são necessárias para evitar resultados enviesados.

Esses conceitos básicos são matematicamente interligados e formam a base para a construção e interpretação dos métodos utilizados na análise de sobrevivência, como o estimador de Kaplan-Meier (Kaplan e Meier, 1958) e o modelo de riscos proporcionais de Cox (Cox, 1972).

2.2 Estimador de Kaplan-Meier

O estimador introduzido por Kaplan-Meier (Kaplan e Meier, 1958), é um método não paramétrico fundamental para estimar a função de sobrevivência $S(t)$ na presença de dados censurados à direita. Este método é amplamente utilizado em estudos clínicos, engenharia de confiabilidade e demais áreas em que o tempo até a ocorrência de um evento é de interesse. A função de sobrevivência $S(t)$, definida pela equação (2.1), representa a probabilidade de um indivíduo ou unidade sobreviver além do tempo t, em que T é uma variável aleatória não negativa que representa o tempo até o evento de interesse.

Embora robusto, o estimador de Kaplan-Meier (Kaplan e Meier, 1958) possui limitações significativas. Primeiramente, ele não incorpora covariáveis, tornando-o inadequado para investigar efeitos de variáveis explicativas sobre o tempo de sobrevivência. Em contraste, modelos semi-paramétricos, como o de Cox (Cox, 1972), ou técnicas de aprendizado de máquina, como florestas aleatórias de sobrevivência, permitem a inclusão

de preditores.

2.2.1 Fundamento Matemático

O estimador de Kaplan-Meier é construído a partir de tempos de evento ordenados $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, em que $t_{(i)}$ denota o i-ésimo tempo distinto em que o evento ocorreu. Para cada $t_{(i)}$, definem-se duas quantidades críticas: n_i, o número de indivíduos em risco imediatamente antes de $t_{(i)}$, e d_i, o número de eventos observados no tempo $t_{(i)}$. O estimador é então calculado como um produto contínuo (produto-limite) das probabilidades condicionais de sobrevivência em cada intervalo entre eventos:

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \left(1 \quad \frac{d_i}{n_i}\right). \tag{2.7}$$

O produtório da equação (2.4) é realizado para todos os tempos $t_{(i)}$ menores ou iguais a t. Para $t < t_{(1)}$, assume-se $\hat{S}(t) = 1$, refletindo a ausência de eventos antes do primeiro tempo observado. A estrutura do produto reflete a natureza sequencial da sobrevivência: a probabilidade do evento não ter ocorrido até o tempo t depende da probabilidade condicional acumulada.

A lógica subjacente ao estimador reside na decomposição da probabilidade de sobrevivência em uma sequência de probabilidades condicionais. Em cada tempo $t_{(i)}$, a probabilidade do evento não ocorrer além desse ponto é condicionada à sobrevivência até $t_{(i)}$. Assim, $\hat{S}(t)$ pode ser interpretado como:

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} P(T > t_{(i)} \mid T \geq t_{(i)}),$$

o qual $P(T > t_{(i)} \mid T \geq t_{(i)}) = 1 \quad \frac{d_i}{n_i}$. Esse enfoque permite incorporar informações parciais de unidades de estudo censuradas, que contribuem para os cálculos apenas até o momento da censura. A censura é tratada implicitamente pela redução progressiva de n_i ao longo do tempo, excluindo unidades de estudo que já foram censuradas ou que experimentaram o evento.

O estimador de Kaplan-Meier é consistente e assintoticamente normal sob condições gerais (Colosimo e Giolo, 2006). Caso não haja censura, o estimador coincide com a função de sobrevivência empírica, dada por

$$\hat{S}(t) = \frac{\text{número de indivíduos com } T > t}{\text{total de indivíduos}}.$$

Na presença de censura, o estimador mantém propriedades de não viés caso a censura seja não informativa, ou seja, o mecanismo de censura é independente do risco de evento.

A variância do estimador pode ser aproximada pela fórmula de Greenwood (Greenwood, 1926):

$$\widehat{\text{Var}}(\hat{S}(t)) = \left(\hat{S}(t)\right)^2 \sum_{t_{i)} \leq t} \frac{d_i}{n_i(n_i \quad d_i)}, \tag{2.8}$$

que permite a construção de intervalos de confiança assimétricos para $S(t)$. Adicionalmente, o estimador é considerado um estimador de máxima verossimilhança não paramétrico para $S(t)$ sob censura não informativa, conforme citado em (Colosimo e Giolo, 2006). O intervalo de confiança que pode ser construído a partir da equação (2.8) é dado por:

$$\hat{S}(t) \pm z_{\ /2} \cdot \sqrt{\widehat{\text{Var}}(\hat{S}(t))},$$

em que $z_{\ /2}$ é o quantil da distribuição normal padrão correspondente ao nível de significância desejado.

2.2.2 Considerações

Uma das limitações é a sensibilidade a padrões complexos de censura. Se a censura for informativa, o estimador pode produzir viés. Além disso, é assumido para o método que os eventos são independentes e que os tempos de censura não carregam informação sobre o processo de falha. Em conjuntos de dados com alta dimensionalidade ou interações complexas entre variáveis, o Kaplan-Meier torna-se insuficiente, exigindo métodos mais flexíveis.

Na prática, o estimador de Kaplan-Meier frequentemente serve como ferramenta exploratória inicial, fornecendo uma visualização da curva de sobrevivência geral antes da aplicação de modelos mais sofisticados.

Posteriormente, para quantificar o efeito de covariáveis e gerar predições individuais, empregam-se modelos de regressão como o de Cox ou florestas aleatórias. A combinação entre o método não paramétrico de Kaplan-Meier, que oferece uma descrição intuitiva da sobrevivência agregada, e estas técnicas preditivas é particularmente valiosa, pois en-

quanto o primeiro serve como uma ferramenta exploratória, as últimas permitem capturar relações não lineares e interações complexas entre variáveis, melhorando a precisão preditiva em cenários mais desafiadores.

2.3 Modelo de Riscos Proporcionais de Cox

O modelo de riscos proporcionais de Cox, proposto em 1972 (Cox, 1972), é um dos métodos mais utilizados na análise de sobrevivência para investigar a relação entre a sobrevivência de um indivíduo e um conjunto de covariáveis. Ele possibilita o estudo dos efeitos das covariáveis sobre o tempo até a ocorrência de um evento de interesse, como morte, falha de um componente ou recorrência de uma doença. A principal vantagem do modelo de Cox é o fato de ser semiparamétrico, pois não é necessária a especificação da forma da função de risco base, uma vez que se concentra na ordem dos eventos em vez de seus tempos exatos, eliminando a necessidade de estimar a função de risco base.

O modelo de Cox é construído com a intenção de modelar a função de risco $\lambda(t|\mathbf{X})$, que representa a taxa instantânea de ocorrência do evento no tempo t, condicionada a um vetor de covariáveis $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. O modelo é dado por:

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}), \tag{2.9}$$

em que $\lambda_0(t)$ é a função de risco base, que depende apenas do tempo t e não das covariáveis. Ela representa o risco para um indivíduo com todas as covariáveis iguais a zero, servindo como uma referência basal para a comparação de riscos entre dois indivíduos diferentes. O termo $\exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ incorpora o efeito das covariáveis $\boldsymbol{X}$, tal que $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ são os coeficientes associados a cada covariável. Esses coeficientes quantificam o impacto das covariáveis sobre o risco de ocorrência do evento.

2.3.1 Interpretação dos Coe cientes

Para a interpretação dos coeficientes, vale ressaltar a importância da suposição fundamental de riscos proporcionais, o que implica que o efeito das covariáveis é multiplicativo ao longo do tempo. O que implica no risco relativo entre dois indivíduos com diferentes covariáveis ser constante ao longo do tempo. O termo risco relativo, no contexto do modelo de Cox, é a razão de riscos (*hazard ratio*). Ele quantifica o quanto o risco de

ocorrência de um evento é maior ou menor para um grupo de indivíduos em comparação com outro. Os coeficientes no modelo de Cox têm uma interpretação direta em termos de risco relativo. Para uma covariável X_j, o coeficiente β_j indica o logaritmo do risco relativo associado a um aumento unitário em X_j, mantendo todas as outras covariáveis constantes. O risco relativo é dado por:

$$\text{Risco relativo} = \exp(\beta_j).$$

Por exemplo, se $\beta_j = 0,5$, então $\exp(0,5) \approx 1,65$, indicando que um aumento unitário em X_j está associado a um aumento de 65% no risco de ocorrência do evento.

Estimativa dos Parâmetros

A estimativa dos coeficientes no modelo de Cox é feita por meio da maximização da função de verossimilhança parcial, que não depende da forma da função de risco base $\lambda_0(t)$. A função de verossimilhança parcial é dada por:

$$L(\quad \mathbf{x}, \boldsymbol{\delta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^\top \quad)}{\sum_{j \in R(t_{\ j})} \exp(\mathbf{x}_j^\top \quad)} \right)^{\delta_i}, \tag{2.10}$$

em que n é o número de indivíduos no estudo, δ_i é uma variável indicadora que assume o valor 1 se o evento foi observado para a unidade de estudo i e 0 se o tempo foi censurado, e $R(t_{(j)})$ é o conjunto de indivíduos em risco no tempo t_j, ou seja, aqueles que ainda não experimentaram o evento e não foram censurados antes de t_j.

Uma característica importante é que a função de verossimilhança parcial não depende explicitamente da função de risco $\lambda_0(t)$, e consequentemente, o tempo t não aparece diretamente na equação (2.10). Isso ocorre porque o modelo separa a contribuição do tempo, que está na função basal, da contribuição dos covariáveis, que são parametrizados pelos coeficientes $\boldsymbol{\beta}$. Dessa forma, a função de verossimilhança parcial é referente apenas à razão dos riscos entre indivíduos no momento do evento, eliminando a necessidade de especificar a forma funcional do risco. Para a estimação dos coeficientes $\boldsymbol{\beta}$, normalmente utiliza-se o método iterativo de Newton-Raphson (Akram e Ann, 2015), que utiliza a derivada primeira (*score*) e a matriz de informação de (Fisher, 1925) para maximizar a log-verossimilhança parcial, fornecendo estimativas eficientes e consistentes sem requerer o conhecimento explícito do tempo de sobrevivência.

Veri cação da Suposição de Riscos Proporcionais

A suposição de riscos proporcionais é importante para a validade do modelo de Cox. Se essa suposição não for atendida, é um indicativo de que o modelo não é válido, pois o modelo não se adequa bem aos dados. Existem várias técnicas para verificar a suposição de riscos proporcionais. Uma abordagem comum é a análise dos resíduos de Schoenfeld (Schoenfeld, 1982) dados por:

$$r_j^{(i)} = x_j^{(i)} \quad \frac{\sum_{l \in R(T_l)} x_j^{(l)} e^{\beta^T x^{\ l)}}}{\sum_{l \in R(T_l)} e^{\beta^T x^{\ l)}}},$$

que são calculados para cada covariável e tempo de evento. Se a suposição de riscos proporcionais for válida, esses resíduos não devem apresentar algum padrão de comportamento e nem heterocedasticidade. Outra abordagem é a utilização de testes estatísticos, como o teste de Grambsch e Therneau (Grambsch e Therneau, 1994), que avaliam a proporcionalidade dos riscos de forma quantitativa. Caso a suposição não seja válida, termos de interação entre as covariáveis e o tempo podem ser incluídos no modelo para capturar efeitos variáveis no tempo.

2.4 Exemplo de aplicaçação na área da saúde

Para aplicar num exemplo prático das técnicas apresentadas acima, foi utilizado um exemplo do livro de Kalbfleisch e Prentice (Kalbfleisch e Prentice, 1980), no qual a base é o banco de dados *veteran*. Vale ressaltar que os dados foram utilizados diretamente do próprio software R (R Core Team, 2024), no pacote *survival*.

2.4.1 Introdução aos dados

Os dados são originais de um ensaio clínico randomizado conduzido pela Administração de veteranos dos Estados Unidos, que investigou a eficácia de um novo regime de quimioterapia em comparação com a terapia padrão em pacientes do sexo masculino com carcinoma de pulmão avançado e inoperável.

O estudo original tinha como objetivo determinar se o tratamento quimioterápico experimental oferecia um benefício de sobrevida em relação ao tratamento padrão. Como objetivo secundário, buscou-se analisar o valor prognóstico de diversas covariáveis clínicas

e demográficas. O estudo é composto por 137 pacientes, para os quais o tempo até o óbito foi registrado. O evento de interesse é a morte, e o conjunto de dados inclui informações de censura à direita para 9 pacientes que saíram do estudo antes da ocorrência do evento.

As informações disponíveis para cada paciente no conjunto de dados são:

- Tempo: tempo de sobrevida observado, observado em dias do paciente no estudo.
- Status: Indicador de desfecho, em que o valor 1 indica que o evento (óbito) foi observado e 0 indica que o tempo de sobrevida foi censurado.
- Tipo de tratamento: Variável de tratamento, indicando o grupo ao qual o paciente foi alocado (1 = Terapia Padrão, 2 = Quimioterapia Experimental).
- Tipo histológico: Tipo histológico do tumor, uma variável categórica com quatro níveis (1 = Escamoso, 2 = Pequenas Células, 3 = Adenocarcinoma, 4 = Grandes Células).
- Escore Karno: Escore de performance de Karnofsky, uma medida padronizada da capacidade funcional do paciente no início do estudo, variando de 0 (morto) a 100 (saúde perfeita).
- Tempo Diagnóstico: Tempo, em meses, decorrido entre o diagnóstico inicial da doença e a inclusão do paciente no estudo.
- Idade:A idade do paciente em anos.
- Terapia Prévia: Indicador de terapia prévia, tal que 0 significa que o paciente não recebeu tratamento anterior e 1 significa que recebeu.

2.4.2 Análise descritiva

Fazendo uma análise descritiva inicial sobre as covariáveis, podemos observar na Tabela 2.4.1 os resultados obtidos:

Tabela 2.4.1: Estatísticas descritivas dos pacientes.

Descritivas para variáveis cont´nuas são apresentadas como Mediana (Quartil 1 - Quartil 3). Descritivas para variáveis categóricas são apresentadas como n (%).	
Variáveis Contínuas	
Idade ao início do estudo (anos)	60 (51 - 68)
Meses desde o diagnóstico	7 (3 - 15)
Escore de performance de Karnofsky	60 (40 - 75)
Variáveis Categóricas	
Tipo de Tratamento	
Padrão	69 (50%)
Teste	68 (50%)
Tipo Histológico	
Escamoso	35 (26%)
Pequenas Células	48 (35%)
Adeno	27 (20%)
Grandes Células	27 (20%)
Terapia Prévia	
Não	128 (93%)
Sim	9 (7%)

2.4.3 Análise Exploratória com Kaplan-Meier

Antes de proceder à modelagem multipla, uma análise exploratória univariada foi conduzida para avaliar o impacto individual de cada covariável na sobrevida dos pacientes. Para este fim, o estimador de Kaplan-Meier foi utilizado para gerar curvas de sobrevida estratificadas para cada variável categórica. A comparação formal entre estas curvas foi realizada através do teste de *log-rank* (Mantel, 1966), que é dado por:

$$Z_i = \frac{\sum_{j=1}^{J}(O_{i,j} \quad E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0,1),$$

na equação (2.4.3), os termos são definidos como Z_i sendo a estatística de teste padronizada para o grupo i; J o número total de tempos de falha distintos; $O_{i,j}$ o número de

eventos observados no grupo i no tempo j; $E_{i,j}$ sendo o número de eventos esperados no grupo i no tempo j sob a hipótese nula; $V_{i,j}$ como a variância da contagem de eventos no tempo j e $\mathcal{N}(0,1)$ a distribuição Normal Padrão.

O teste de *log-rank* é um teste de hipóteses não paramétrico cujo objetivo é determinar se existe uma diferença estatisticamente significativa entre as distribuições de sobrevida de dois ou mais grupos independentes. A hipótese nula (H_0) do teste é que não há diferença entre as curvas de sobrevida dos grupos, enquanto a hipótese alternativa (H_1) é que pelo menos uma das curvas difere das demais. Um p-valor baixo leva à rejeição da hipótese nula, sugerindo que a variável em questão é um fator prognóstico significativo.

A aplicação desta análise ao conjunto de dados veteran revelou que as covariáveis escore de performance e tipo histológico do tumor são os fatores prognósticos mais fortes. A Figura 2.1 apresenta as curvas de sobrevida para pacientes com escore de Karnofsky "Alto" (maior que 60) *versus* "Baixo" (menor ou igual a 60). Observa-se uma separação clara entre os grupos, e o teste de *log-rank* confirma esta observação visual com um resultado altamente significativo ($p < 0,0001$), indicando que um melhor estado funcional no início do estudo está fortemente associado a uma maior sobrevida.

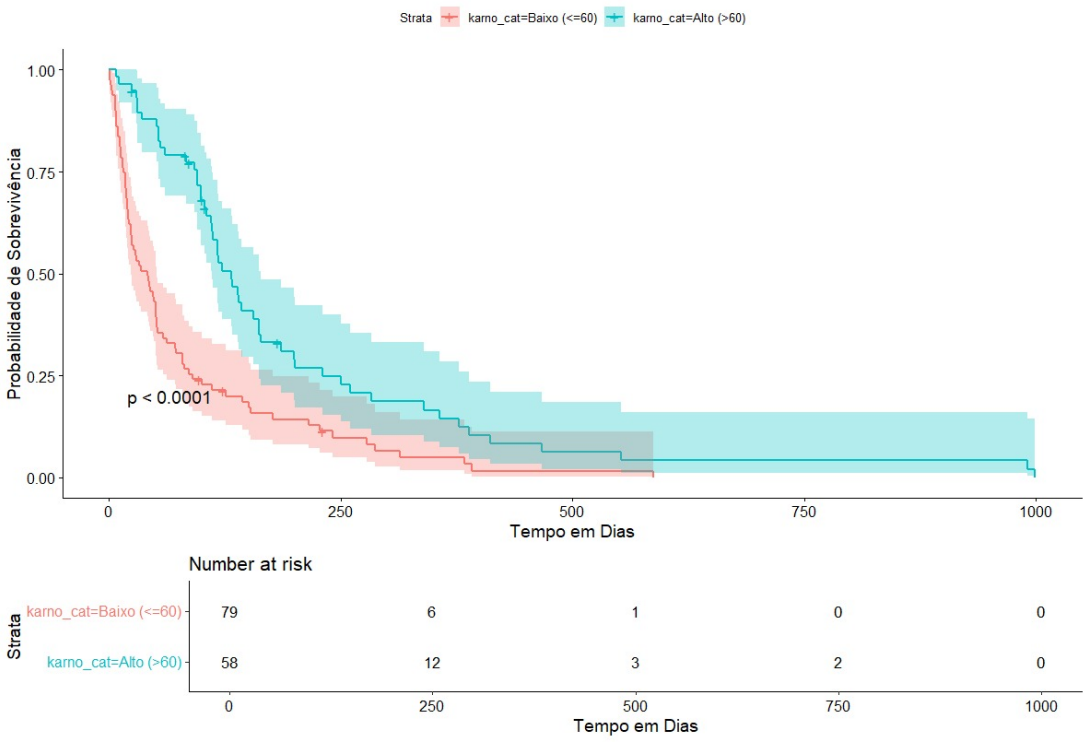


Figura 2.1: Curvas de Sobrevivência de Kaplan-Meier estratificadas pelo Escore de Performance de Karnofsky.

De forma semelhante, na Figura 2.2 foram ilustradas as diferenças na sobrevida entre

os quatro tipos histológicos de tumor. O teste de *log-rank* também apontou uma diferença altamente significativa ($p < 0,0001$) entre os grupos. Visualmente, pacientes com tumores do tipo "Escamoso" apresentam uma sobrevida notavelmente superior em comparação aos outros tipos, especialmente em relação ao de "Pequenas Células" que demonstrou o pior prognóstico.

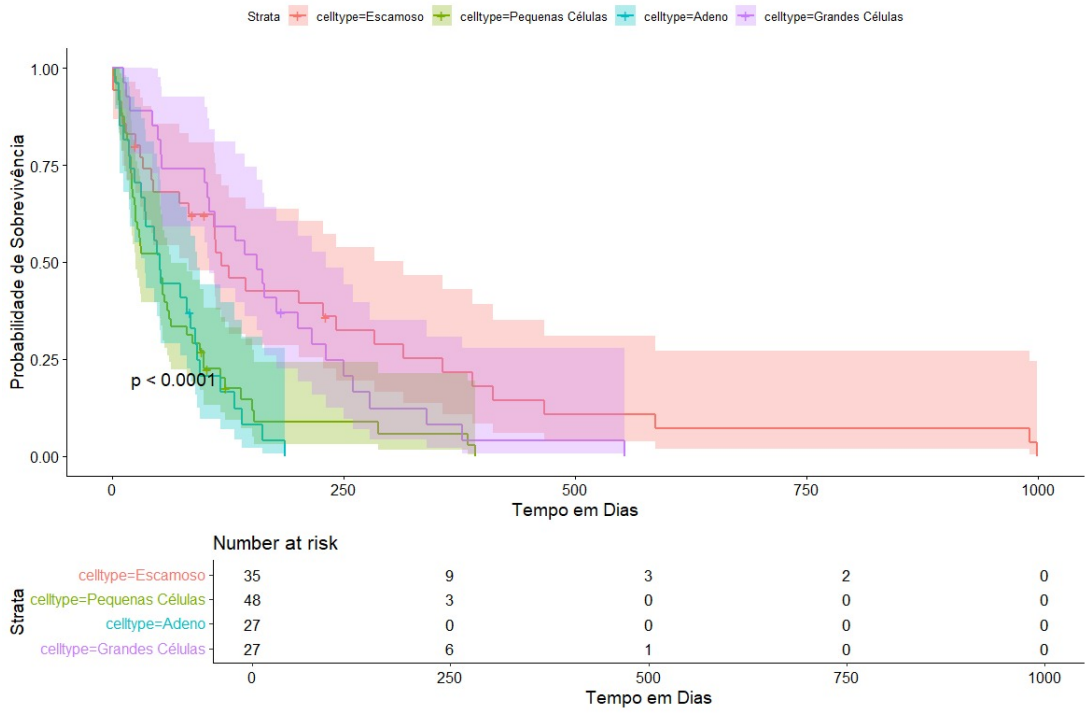


Figura 2.2: Curvas de Sobrevivência de Kaplan-Meier estratificadas pelo Tipo Histológico do Tumor.

Em contraste, as variáveis tipo de tratamento e terapia prévia não apresentaram diferenças estatisticamente significativas em suas respectivas análises univariadas (p-valores de 0,88 e 0,84, respectivamente). Este resultado, no entanto, não exclui a possibilidade de que estas variáveis tenham um efeito relevante no contexto de um modelo múltiplo, em que os efeitos são ajustados pela presença de outras covariáveis. A forte significância univariada do escore de Karno e tipo histológico, por outro lado, já sinaliza que estas foram, muito provavelmente, as variáveis mais importantes no modelo de Cox. A transição para a modelagem múltipla é, portanto, essencial para confirmar estes achados e para investigar o efeito ajustado de cada preditor na presença dos demais.

2.4.4 Ajuste e Diagnóstico do Modelo de Cox Inicial

O primeiro passo na modelagem múltipla foi o ajuste de um modelo de Cox completo, incluindo todas as covariáveis preditoras, assim, no software R, foi utilizado o pacote *survival* e a função *coxph* em sua forma *default*. Os resultados deste modelo inicial são apresentados na Tabela 2.4.2:

Tabela 2.4.2: Resultados do ajuste inicial do modelo de Cox com todas as covariáveis.

Variável	**coef**	**exp(coef)**	**se(coef)**	**z**	$Pr(Z > \|z\|)$
tratamento padrão	0,295	1,343	0,208	1,42	0,156
histológico Pequenas Celulas	0,862	2.367	0,275	3,13	0,002
histológico Adeno	1,196	3,307	0,301	3,98	$< 0,001$
histológico Grandes Celulas	0,401	1,494	0,283	1,42	0,156
Escore de Karno	-0,033	0,968	0,006	-5,96	$< 0,001$
Tempo diagnóstico	0,000	1,000	0,009	0,01	0,993
Idade	-0,009	0,991	0,009	-0,94	0,349
Terapia Prévia sim	0,072	1,074	0,232	0,31	0,758

Antes de uma análise detalhada dos resultados, vamos para a etapa seguinte que é a verificação da suposição de riscos proporcionais. O teste baseado nos resíduos de Schoenfeld foi aplicado, e seus resultados, exibidos na Tabela 2.4.3, revelaram uma violação significativa da suposição para duas covariáveis, além do modelo global. As violações foram impulsionadas principalmente pelas variáveis (p=0,0016) e `karno` (p=0,0003), indicando que seus efeitos sobre o risco não são constantes ao longo do tempo.

Tabela 2.4.3: Resultados do teste de suposição de riscos proporcionais para o modelo inicial.

Variável	**chisq**	**df**	**p-valor**
Tipo de tratamento	0,264	1	0,607
Tipo histológico	15,227	3	0,002
Escore de Karno	12,935	1	0,0003
Tempo diagnostico	0,013	1	0,910
Idade	1,829	1	0,176
Terapia Prévia	2,166	1	0,141
GLOBAL	34,553	8	$< 0,001$

2.4.5 Estratégia de Correção e Seleção do Modelo Final

Com a violação da suposição de riscos proporcionais, o modelo inicial não é válido. Uma abordagem padrão para simplificar o modelo é a seleção de variáveis. Decidimos usar uma abordagem de *stepwise backward* com critério AIC, assim chegando a um modelo mais parcimonioso contendo apenas tipo histológico e escore de karno. No entanto, ao reavaliar a suposição de riscos proporcionais neste modelo simplificado, a violação persistiu, demonstrando que a seleção de variáveis por si só não corrige problemas de não-proporcionalidade.

A estratégia metodológica adequada para lidar com uma covariável que viola a suposição de riscos proporcionais é a estratificação. Neste processo, o modelo é ajustado permitindo que a função de risco base, $\lambda_0(t)$, seja diferente para cada nível da variável que viola a suposição, relaxando a suposição de riscos proporcionais para aquela variável. Matematicamente, o modelo estratificado é definido como:

$$h(t|\boldsymbol{x}, g) = h0g(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta}), \tag{2.11}$$

em que g representa o estrato ao qual o indivíduo pertence (neste caso, "Baixo" ou "Alto" escore de Karnofsky), $\lambda_{0g}(t)$ é a função de risco base específica para o estrato g, e o efeito das demais covariáveis $\boldsymbol{x}$ é assumido como constante entre os estratos. Para implementar esta abordagem, a variável escore de karno foi discretizada em duas categorias com base em sua mediana. Em seguida, um novo modelo de Cox foi ajustado usando `strata(karno_cat)`, e o processo de seleção por AIC foi reaplicado.

O modelo final obtido, demonstrou resultados melhores aos outros modelos além de ser o único que atendeu as suposições de risco proporcional, estratificado por escore de Karno e ajustado por AIC, reteve apenas a variável tipo histológico. Os resultados deste modelo são apresentados na Tabela 2.4.4.

Tabela 2.4.4: Resultados do ajuste final de Cox, estratificado por escore de Karnofsky.

Variável	**coef**	**exp(coef)**	**se(coef)**	**z**	$Pr(Z > \|z\|)$
celltypePequenas Celulas	0,898	2,455	0,269	3,34	0,0008
celltypeAdeno	1,132	3,101	0,297	3,81	0,0001
celltypeGrandes Celulas	0,223	1,250	0,279	0,80	0,4226

2.4.6 Validação e Interpretação do Modelo Final

A validade do modelo final foi confirmada ao reaplicar o teste de suposição de PH (Tabela 2.4.5). Com um p-valor global de 0,085, não há mais evidências para rejeitar a hipótese nula, indicando que o modelo estratificado agora atende à suposição de riscos proporcionais.

Tabela 2.4.5: Resultados do teste de suposição de PH para o modelo final estratificado.

Variável	**chisq**	**df**	**p-valor**
Tipo histológico	6,62	3	0,085
Global	6,62	3	0,085

A interpretação do modelo final (Tabela 2.4.4) confirma as observações da análise exploratória. O tipo histológico do tumor é um preditor altamente significativo. Tomando o tipo "Escamoso" como referência, pacientes com tumores de "Pequenas Células" têm um risco de óbito 2,46 vezes maior (*Hazard Ratio* = 2,455), enquanto pacientes com "Adenocarcinoma" têm um risco 3,1 vezes maior (*Hazard Ratio* = 3,101). O tipo "Grandes Células" não apresentou um risco significativamente diferente do tipo "Escamoso". A variável escore de karno, agora como variável de estratificação, ajusta o modelo para o estado funcional do paciente sem estimar um *Hazard Ratio* único, reconhecendo que seu impacto na sobrevida varia ao longo do tempo.

Capítulo 3

Aprendizado de máquina

3.1 Árvores de decisão

As árvores de decisão representam uma classe de algoritmos de aprendizado supervisionado de importância teórica e prática. Aplicáveis tanto em tarefas de classificação quanto de regressão, sua popularidade deriva de uma estrutura que tem como objetivo imitar o raciocínio humano, baseada em uma sequência de decisões hierárquicas. Esta característica permite aos modelos uma alta interpretabilidade, uma propriedade valiosa em domínios em que a transparência do processo decisão é fundamental. Esta seção foi fundamentada na abordagem de Rafael e Tiago (Izbicki e dos Santos, 2020); a partir disso, foi feita uma análise detalhada da estrutura, do processo de construção e das propriedades desses modelos. Alguns conceitos de árvore de decisão estão detalhados no Apêndice A 7.

3.1.1 Estrutura e funcionamento

Uma árvore de decisão é um grafo acíclico direcionado que implementa o princípio da partição recursiva. A estrutura é composta por três tipos de nós: o nó raiz, que é o ponto de partida e abrange todo o conjunto de dados de treinamento; os nós internos, tal como se aplicam os testes sobre as covariáveis; e os nós folha, que não se dividem mais e contêm a predição final.

O funcionamento para uma nova observação $\mathbf{x} = (x_1, \ldots, x_p)^\top$ é direto, partindo do nó raiz, a observação percorre um único caminho na árvore de acordo com os resultados dos testes nos nós internos. Por exemplo, em uma divisão binária baseada na regra $X_j \leq c$, a observação é direcionada para o filho à esquerda se a condição for verdadeira, e para o

filho à direita, caso contrário. Esse processo continua até que um nó folha seja alcançado. O valor associado a esse nó é então atribuído como a predição para $\mathbf{x}$, sendo a classe majoritária para casos de classificação.

3.1.2 Construção da árvore: o algoritmo de crescimento

A construção de uma árvore de decisão é um processo iterativo que busca, a cada etapa, a melhor divisão possível dos dados.

O critério de divisão e o ganho de informação

O núcleo do algoritmo reside na escolha da melhor divisão. Uma divisão é definida por uma variável X_j e um ponto de corte c. O objetivo é encontrar o par (j, c) que maximize o ganho de informação, ou seja, que produza os nós filhos mais homogêneos possíveis. A medida de pureza ou impureza varia entre tarefas de classificação e regressão.

Para classificação, os critérios de impureza mais comuns são o índice de Gini, que mede a probabilidade de um elemento aleatoriamente escolhido no nó ser classificado incorretamente e para um nó t é definido por

$$G(t) = 1 \quad \sum_{k=1}^{K} p_k^2,$$

para qual p_k é a proporção de observações da classe k.A Entropia, originária da teoria da informação, que mede o grau de incerteza em um nó e é dada por

$$H(t) = \quad \sum_{k=1}^{K} p_k \log_2 p_k.$$

O ganho de informação para uma divisão s que particiona um nó t em filhos t_L (esquerda) e t_R (direita) é formalmente definido como:

$$\text{Ganho}(t, s) = \text{Impureza}(t) \quad (w_L \cdot \text{Impureza}(t_L) + w_R \cdot \text{Impureza}(t_R))$$

em que Impureza pode ser dado por Gini ou Entropia, e w_L, w_R são os pesos proporcionais ao número de observações em cada nó filho.

O processo de busca

O algoritmo de crescimento itera sobre todas as variáveis $X_j \in \{1, \ldots, p\}$ e, para cada variável, sobre todos os seus possíveis pontos de corte c. O par (j, c) que maximiza a função de ganho de informação é selecionado para realizar a divisão daquele nó.

Visualização de uma árvore de decisão

Como dois exemplos de formato de árvore de decisão, retirados do livro do Rafael e Tiago (Izbicki e dos Santos, 2020), podemos ver a seguir:

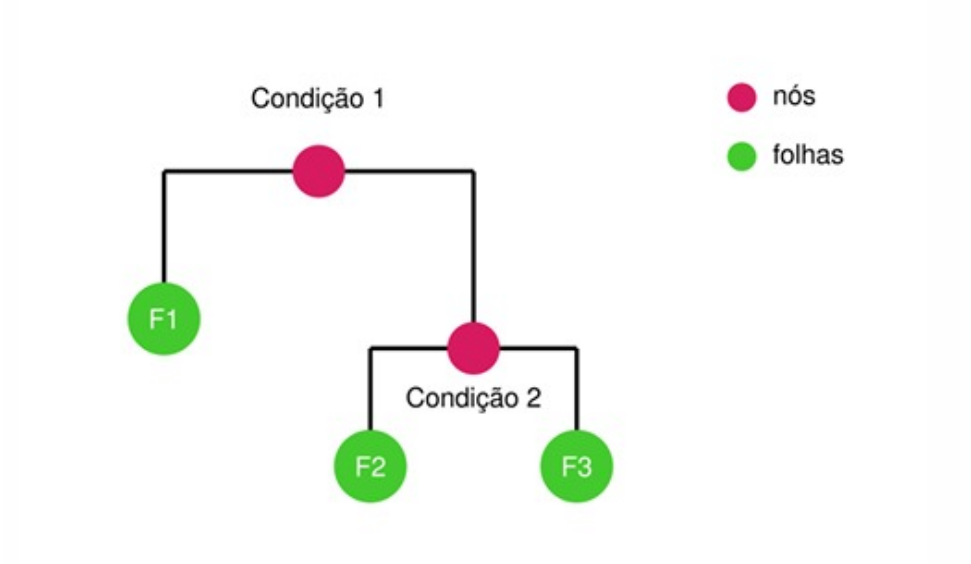


Figura 3.1: Exemplo de árvore de decisão.

na Figura 3.1 podemos ver uma forma mais genérica de como são gerados os nós e folhas de uma árvore, sem muita especificidade, já no exemplo da Figura 3.2, podemos ver de forma mais clara o corte tomado para cada variável.

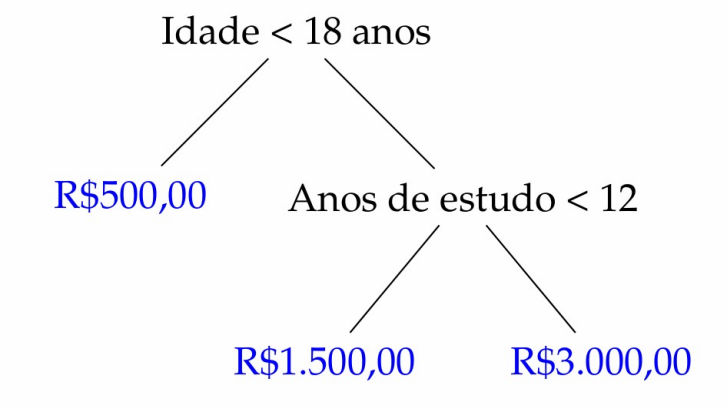


Figura 3.2: Exemplo com variáveis de árvore de decisão.

Controlando a Complexidade: poda e critérios de parada

Uma árvore totalmente crescida, sem restrições, se ajusta perfeitamente aos dados de treino, mas captura ruídos e particularidades que não generalizam para novos dados, um fenômeno conhecido como sobreajuste (*over tting*). Para construir um modelo com bom poder de generalização, é crucial controlar sua complexidade. Existem duas estratégias principais para isso. A primeira, conhecida como pré-poda (critérios de parada), consiste em parar o crescimento da árvore precocemente. Isso é feito estabelecendo-se regras de parada, como a profundidade máxima que a árvore pode atingir, o número mínimo de amostras que um nó folha deve conter, o número mínimo de amostras que um nó deve ter para poder ser dividido, ou um ganho de informação mínimo para que uma divisão seja efetuada.

A segunda estratégia é a pós-poda (*post-pruning*), que consiste em crescer a árvore até sua máxima extensão e, em seguida, "podar" ramos que não contribuem significativamente para o desempenho do modelo em um conjunto de validação. A técnica mais famosa é a poda por custo-complexidade (*cost-complexity pruning*), que busca um balanço entre o erro de ajuste da árvore e sua complexidade (número de folhas). Ela minimiza o critério:

$$C\ (\mathcal{T}) = \text{Erro}(\mathcal{T}) + \quad |\mathcal{T}|,$$

em que $\mathcal{T}$ é a árvore, $\text{Erro}(\mathcal{T})$ é o erro total nos nós terminais, $|\mathcal{T}|$ é o número de folhas, e ≥ 0 é o parâmetro de complexidade que penaliza árvores maiores, geralmente ajustado via validação cruzada.

Interpretabilidade e Extração de Regras

A estrutura hierárquica da árvore reflete a importância relativa das covariáveis: aquelas que aparecem mais próximas do nó raiz são, em geral, as que possuem maior poder discriminatório sobre o conjunto de dados.

Cada caminho percorrido da raiz até um nó folha representa uma sequência de condições que isola um subconjunto específico dos dados, caracterizando um perfil particular. A análise do conjunto de todos esses caminhos permite a formulação de um sistema de regras explícito que descreve as relações encontradas pelo modelo. Desta forma, a árvore não atua somente como um algoritmo preditivo, mas também como um método exploratório que pode revelar interações e padrões nos dados.

Vantagens e Limitações

As árvores de decisão possuem diversas vantagens, como sua alta interpretabilidade, já discutida, que as torna fáceis de visualizar e entender. Além disso, apresentam requisitos mínimos de preparação de dados, pois não exigem normalização de variáveis numéricas e lidam nativamente com variáveis categóricas. Também se destacam no manejo de dados ausentes, uma vez que algoritmos como o CART possuem estratégias sofisticadas, a exemplo do uso de divisões substitutas (*surrogate splits*), para lidar com valores ausentes durante o treino e a predição. Finalmente, são capazes de capturar automaticamente relações não-lineares e interações entre variáveis.

Contudo, os modelos também apresentam limitações importantes. A principal é sua instabilidade (alta variância), pois pequenas variações nos dados de treino podem levar a árvores drasticamente diferentes. Outra desvantagem é a tendência ao sobreajuste, o que requer um ajuste cuidadoso dos hiperparâmetros. Adicionalmente, as árvores possuem um viés de divisão, uma vez que estas são sempre ortogonais aos eixos das variáveis, tornando difícil a captura de relações lineares ou diagonais no espaço dos dados. Por fim, podem apresentar um viés por atributos categóricos, favorecendo variáveis com um grande número de níveis.

Essas limitações, especialmente a alta variância, foram o principal catalisador para o desenvolvimento de métodos de conjunto (*ensemble methods*). Tais métodos, ao invés de confiarem em um único modelo, combinam as predições de múltiplas árvores para produzir um resultado final mais robusto, estável e, geralmente, com maior acurácia preditiva. O passo lógico seguinte no aprimoramento de modelos baseados em árvores é, portanto, o estudo da técnica de Florestas Aleatórias, que é exatamente um *ensemble* de árvores de decisão.

3.2 Florestas Aleatórias

As florestas aleatórias (*random forests*) são um método de aprendizado de máquina supervisionado do tipo *ensemble*, introduzido por Breiman (Breiman, 2001). O principal ponto por trás de métodos de *ensemble* é que a combinação de múltiplos modelos, mesmo que individualmente fracos ou instáveis, pode gerar um modelo final único, mais robusto e com maior poder de generalização. As florestas aleatórias aplicam essa ideia utilizando árvores de decisão como seus modelos base, tornando-se uma das técnicas mais eficazes e

amplamente utilizadas tanto em problemas de classificação quanto de regressão.

3.2.1 Construção da Floresta

A ideia central das florestas aleatórias é mitigar a variância grande dos dados, uma característica conhecida das árvores de decisão individuais. Uma única árvore de decisão é muito sensível a pequenas variações nos dados de treinamento, podendo gerar estruturas completamente diferentes e, consequentemente, sofrer *over tting*. A floresta aleatória impede essa instabilidade introduzindo aleatoriedade em dois níveis distintos durante sua construção.

O primeiro nível de aleatoriedade vem do processo de *bagging* (agregação por *Bootstrap*). Para uma floresta com B árvores, são geradas B amostras de *bootstrap* a partir do conjunto de dados original. Uma amostra de *bootstrap* é uma nova amostra de mesmo tamanho da original, obtida através de amostragem com reposição. Isso significa que algumas observações originais podem aparecer múltiplas vezes na amostra de *bootstrap*, enquanto outras podem não aparecer. Cada uma das B árvores da floresta é então treinada em uma dessas amostras de *bootstrap* distintas, o que promove a diversidade entre elas.

O segundo nível de aleatoriedade é a seleção aleatória de covariáveis em cada nó. Ao construir uma árvore de decisão tradicional, o algoritmo busca a melhor divisão entre todas as p covariáveis disponíveis. Em uma floresta aleatória, para cada nó candidato a uma divisão, o algoritmo seleciona apenas um subconjunto aleatório de m covariáveis (tal que $m < p$). Apenas dentro desse subconjunto de candidatas a melhor covariável para a divisão é escolhida. Este passo é crucial, pois força as árvores a serem ainda mais diferentes umas das outras. Se uma ou duas covariáveis forem muito preditivas, sem essa seleção aleatória, a maioria das árvores as usaria no topo da sua estrutura, tornando-as muito semelhantes. Empiricamente, os valores recomendados para m são $m \approx \sqrt{p}$ para problemas de classificação e $m \approx p/3$ para problemas de regressão.

O processo de construção da floresta pode ser definido pelo seguinte algoritmo:

- Para cada uma das B árvores que irão compor a floresta:
  - Passo 1: uma amostra de *bootstrap*, Z_b, é gerada a partir do conjunto de treinamento original.

  - Passo 2: uma árvore de decisão, T_b, é treinada utilizando exclusivamente a amostra Z_b. Durante o crescimento desta árvore, o seguinte processo recursivo é aplicado em cada nó:
    * Um subconjunto de m covariáveis é selecionado aleatoriamente do total de p disponíveis.
    * A melhor covariável e o ponto de corte ótimo para a divisão são determinados considerando apenas as m covariáveis selecionadas.
    * A árvore cresce até que um critério de parada seja alcançado (como atingir a profundidade máxima ou um número mínimo de observações por nó).

- Repita os passos até a obtenção de B árvores: O resultado final é o modelo de *ensemble* composto pelo conjunto das B árvores treinadas, $\{T_b\}_{b=1}^{B}$.

3.2.2 Mecanismo de Agregação e Predição

Após a construção das B árvores independentes e sem correlação, a predição para uma nova observação $\mathbf{x}$ é feita pela agregação das predições de todas as árvores. A forma de agregação depende da natureza do problema, para o nosso caso de classificação, seguindo a teoria explicada no livro (Izbicki e dos Santos, 2020), temos que a predição final é a classe mais votada (moda) entre todas as árvores:

$$\hat{y}(\mathbf{x}) = \text{moda}\left\{T_1(\mathbf{x}), \ldots, T_B(\mathbf{x})\right\}.$$

Essa agregação explora o fato de que, embora cada árvore individual possa ter ruídos e ter alta variância, a média de muitas árvores independentes resulta em um preditor com variância muito menor, mantendo o baixo viés das árvores individuais profundas.

A eficácia desta redução de variância depende fundamentalmente da correlação entre as árvores. Se cada árvore tem uma variância σ^2 e a correlação média entre quaisquer duas árvores é ρ, a variância do estimador agregado $\hat{y}(\mathbf{x})$ é dada por:

$$\text{Var}(\hat{y}(\mathbf{x})) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{3.1}$$

A Equação (3.1) demonstra que, com o aumento do número de árvores B, o segundo termo da equação tende a zero. Contudo, a variância geral é limitada pelo primeiro termo, que depende diretamente da correlação ρ. É por isso que a seleção aleatória de covariáveis

é tão importante: ao reduzir diretamente a correlação ρ entre as árvores, ela diminui o limite inferior da variância do modelo final, tornando-o mais estável e preciso.

3.2.3 Importância das Variáveis

Uma das propriedades mais significativas das florestas aleatórias é sua capacidade de quantificar a importância das covariáveis preditoras, mesmo em cenários de alta dimensionalidade. Para isso, duas abordagens principais são comumente empregadas.

A primeira, conhecida como Redução Média de Impureza (*Mean Decrease in Impurity*), baseia-se na própria construção das árvores. Para cada árvore, toda vez que uma variável é utilizada para dividir um nó, a redução na impureza do nó (medida, por exemplo, pelo Índice de Gini em classificação ou pela variância em regressão) é calculada e atribuída àquela variável. A importância final da variável é a soma total dessas reduções ao longo de todas as árvores da floresta, devidamente normalizada. Embora seja uma métrica computacionalmente eficiente, ela pode apresentar um viés, favorecendo variáveis contínuas ou com um número maior de categorias.

Uma abordagem alternativa, considerada mais robusta e confiável, é a Redução Média da Acurácia (*Mean Decrease in Accuracy*). Este método, que é uma forma de importância por permutação, mede diretamente o impacto que uma variável tem no desempenho preditivo do modelo. Seu funcionamento baseia-se no erro *out-of-bag* (OOB). Para cada árvore, a acurácia é inicialmente calculada em sua amostra OOB. Em seguida, os valores de uma única covariável j são permutados aleatoriamente nesta mesma amostra, quebrando qualquer associação entre ela e a variável resposta. A acurácia é então recalculada com estes dados modificados. A importância da covariável j é a diferença média na acurácia do modelo, antes e depois da permutação, agregada para todas as árvores. Uma queda significativa na acurácia indica que a variável é crucial para as predições do modelo. Apesar de ser mais custosa computacionalmente, esta técnica não sofre do mesmo viés da "Redução Média de Impureza" e é, em geral, a mais recomendada para uma avaliação fidedigna da relevância das variáveis.

3.2.4 Validação e Controle de Sobreajuste

Ao contrário das árvores individuais, que tendem a sobreajustar os dados caso sejam muito profundas, as florestas aleatórias são notavelmente robustas ao *over tting*. O au-

mento do número de árvores (B) não leva ao *over tting*, na realidade ele converge o erro de predição da floresta. Isso se deve ao efeito de suavização do processo de agregação.

Uma das características de florestas aleatórias é o uso do erro *out-of-bag* (OOB) para validação interna. Como cada árvore é treinada em uma amostra de *bootstrap*, em média, cerca de um terço das observações originais não são incluídas em cada amostra. Essas observações formam o conjunto OOB para aquela árvore específica. Para cada observação i, pode-se obter uma predição agregando apenas os votos das árvores para as quais i estava no conjunto OOB. Comparando essas predições OOB com os valores reais, obtém-se uma estimativa de erro (o erro OOB) que serve como uma validação cruzada e computacionalmente eficiente.

De maneira formal, se $\hat{y}_{oob}^{(i)}$ é a predição OOB para a observação i, o erro OOB geral é a média do erro de predição em todas as observações:

$$\text{Erro OOB} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_{oob}^{(i)}),$$

em que $L(y, \hat{y})$ é uma função de perda apropriada, como o erro quadrático para regressão ou uma função indicadora de erro para classificação.

Os principais hiperparâmetros a serem ajustados são o número de árvores (B), o número de variáveis candidatas por divisão (m), e parâmetros que controlam a complexidade das árvores individuais, como a profundidade máxima ou o número mínimo de observações por nó folha.

3.2.5 Vantagens e Limitações

As florestas aleatórias destacam-se por seu notável desempenho preditivo em uma vasta gama de problemas. Sua robustez e precisão são tão consistentes que frequentemente servem como um modelo de referência de alta performance, contra o qual outras técnicas de aprendizado de máquina são comparadas. Elas lidam nativamente com cenários de alta dimensão, capturam interações complexas entre as covariáveis, são robustas a *outliers* e à inclusão de variáveis irrelevantes, e exigem relativamente poucos ajustes finos para alcançar um bom resultado.

Em contrapartida, a principal limitação do método reside na sua interpretabilidade reduzida se comparada a modelos mais simples. Ao contrário de uma única árvore de decisão, que pode ser facilmente visualizada e suas regras de decisão compreendidas, uma flo-

resta composta por centenas de árvores funciona essencialmente como uma “caixa-preta”. Embora as ferramentas para análise de importância de variáveis ofereçam informações sobre quais fatores mais influenciam a predição, o funcionamento interno do modelo como um todo permanece, de certa forma, desconhecido. Adicionalmente, seu custo computacional para treinamento e armazenamento é significativamente maior, o que pode se tornar um fator restritivo em aplicações com conjuntos de dados muito grandes ou com um número elevado de árvores.

3.3 Métricas de Performance

3.3.1 *C Index*

O Índice de Concordância (Índice C) é uma métrica que avalia a capacidade de discriminação de um modelo, ou seja, sua habilidade de ordenar corretamente os pacientes em função do risco (Harrell *et al.*, 1982). De forma intuitiva, ele mede a probabilidade de que, para um par de indivíduos, o modelo atribua um escore de risco maior àquele que de fato sofreu o evento primeiro. Em cenários de classificação binária, é matematicamente equivalente à Área Sob a Curva ROC (AUC).

Formalmente, o Índice C é definido como:

$$C = \frac{\sum_{i<j} \delta_{ij} \ (\hat{h}_i > \hat{h}_j) + 0,5 \times \delta_{ij} \ (\hat{h}_i = \hat{h}_j)}{\sum_{i<j} \delta_{ij}}$$

A fórmula calcula a proporção de pares concordantes entre todos os pares informativo” (onde o desfecho pode ser comparado sem ambiguidades de censura). Um par é considerado concordante se o indivíduo com o desfecho observado primeiro recebe um escore de risco ($\hat{h}$) maior. Casos de empate nos escores de risco recebem um crédito de 0,5.

O valor do Índice C varia de 0,5 a 1. Um valor de 1 indica uma discriminação perfeita, enquanto um valor de 0,5 significa que o desempenho do modelo não é melhor do que uma escolha aleatória.

3.3.2 *Brier Score*

O escore de Brier é uma regra de pontuação estritamente própria (*proper scoring rule*) que mede a performance de previsões probabilísticas, introduzida originalmente por

(Brier, 1950). Diferente de métricas que avaliam apenas a classificação final, o escore de Brier avalia a qualidade das probabilidades em si, penalizando previsões que são excessivamente confiantes e incorretas. Essencialmente, ele quantifica tanto a calibração quanto a discriminação do modelo.

Matematicamente, ele é definido como o Erro Quadrático Médio (EQM) entre as probabilidades previstas e os resultados binários observados:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i \quad o_i)^2,$$

para qual N é o número total de observações, f_i é a probabilidade do evento prevista para a observação i, e o_i é o desfecho observado (1 se o evento ocorreu, 0 caso contrário).

O Escore de Brier varia de 0 a 1, tal quais valores mais baixos indicam melhor performance. Um score de 0 representa uma previsão perfeita. Por ser sensível a erros grandes, ele penaliza fortemente um modelo que prevê 95% de chance para um evento que não ocorre, por exemplo. Em análise de sobrevivência, em que a censura impede a observação de todos os desfechos, o Escore de Brier é adaptado através da ponderação pelo inverso da probabilidade de censura, garantindo uma estimativa robusta da acurácia do modelo ao longo do tempo, conforme proposto por (Graf *et al.*, 1999).

3.3.3 Métricas de Classi cação: a tabela de confundimento

Enquanto o Índice C e o Brier Score avaliam a qualidade discriminativa e probabilística de um modelo, a avaliação de sua performance em tarefas de classificação categórica é classicamente realizada através de uma tabela de confundimento. Esta tabela é uma ferramenta essencial que sumariza o resultado das predições de um modelo ao compará-las com as classes reais observadas no conjunto de dados. A matriz quantifica quatro tipos de resultados:

- **Verdadeiro Positivo (VP):** observações da classe positiva que foram corretamente classificadas pelo modelo como positivas.
- **Verdadeiro Negativo (VN):** observações da classe negativa que foram corretamente classificadas como negativas.
- **Falso Positivo (FP):** observações da classe negativa que foram incorretamente classificadas como positivas. Também conhecido como Erro do Tipo I.

- **Falso Negativo (FN):** observações da classe positiva que foram incorretamente classificadas como negativas. Também conhecido como Erro do Tipo II.

A estrutura da tabela pode ser visualizada na Tabela 3.3.1.

Tabela 3.3.1: Estrutura de uma tabela de confundimento para classificação binária.

		Predito	
		Positivo	Negativo
Verdadeiro	Positivo	VP	FN
	Negativo	FP	VN

A partir desta tabela, algumas métricas de performance podem ser calculadas. As mais fundamentais são:

- **Acurácia:** mede a proporção de predições totais que o modelo acertou. Embora seja uma métrica intuitiva, pode ser enganosa em cenários com classes desbalanceadas.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.2}$$

- **Sensibilidade (Recall):** mede a proporção de positivos reais que foram corretamente identificados. É uma métrica crucial caso o custo de um falso negativo for alto.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{3.3}$$

- **Especi cidade:** mede a proporção de negativos reais que foram corretamente identificados. É importante caso o custo de um falso positivo é alto (e.g., direcionar uma custosa ação de retenção para um cliente que não iria cancelar).

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{3.4}$$

A análise conjunta destas métricas oferece uma visão muito mais completa da performance de um classificador do que a acurácia isoladamente.

3.4 Simulação de aplicação de Árvore de decisão e floresta aleatória

Com o objetivo de ilustrar a aplicação prática das técnicas discutidas, foi conduzido um estudo de caso computacional para modelar o risco de abandono (*churn*) de clientes, que seria a taxa de risco para o cliente abandonar o plano da empresa. O processo envolveu a simulação de um conjunto de dados sintético, seguido pela construção, otimização e avaliação comparativa de um modelo de Árvore de Decisão e um de Floresta Aleatória. Esta seção detalha cada etapa do processo, desde a geração dos dados até a interpretação dos resultados.

3.4.1 Geração do conjunto de dados

Para garantir um ambiente controlado e com relações preditivas conhecidas, foi gerado um conjunto de dados com 2.000 clientes e 4 variáveis preditoras. A reprodutibilidade da simulação foi assegurada fixando uma semente. As variáveis criadas foram:

- idade: idade do cliente, em anos, simulada a partir de uma distribuição uniforme: U[18,70];
- renda mensal: renda mensal do cliente, simulada a partir de uma distribuição Normal com média de R$ 4.500 e desvio padrão de R$ 1.500;
- meses cliente: tempo de permanência do cliente no plano da empresa, em meses, amostrado uniformemente de 1 a 84;
- plano fidelidade: indicador se o cliente participa de um plano de fidelidade.

A variável resposta, abandono, foi gerada a partir de um modelo logístico subjacente, para assegurar que as variáveis preditoras tivessem um sinal preditivo claro. Como queremos que a variável resposta seja uma probabilidade, a modelagem para a variável abandono requer uma transformação que mapeie a saída de uma combinação linear de preditores (cujo resultado pode assumir qualquer valor real) para o intervalo entre 0 e 1. Para este fim, a abordagem padrão é utilizar a transformação *logit*.

A função *logit* é o logaritmo natural da razão de chances (*odds*), tal que a chance é a razão entre a probabilidade de um evento ocorrer (p) e a probabilidade de ele não

ocorrer (1 p). Assim, o logito de p é $\ln(p/(1 \quad p))$. Esta função transforma um valor do intervalo de probabilidade $(0, 1)$ para o domínio dos números reais, permitindo que ele seja modelado por uma função linear. Desta forma, foi possível definir o logito da probabilidade de abandono como uma combinação linear das covariáveis. A variável resposta, *abandono*, foi gerada a partir de um modelo logístico subjacente para garantir um sinal preditivo claro. As covariáveis foram definidas da seguinte forma:

- meses sendo cliente = x_1;
- renda mensal = x_2;
- idade = x_3;
- plano fidelidade = x_4 (variável binaria, sim ou não).

A probabilidade de abandono foi então calculada aplicando-se a função logística inversa à seguinte combinação linear:

$$\text{logit}(p_{\text{abandono}}) = 1,0 \quad 0,045 \cdot x_1 \quad 0,7 \cdot x_2 + 0,01 \cdot x_3 \quad 1,6 \cdot \ (x_4 = \text{Sim}), \tag{3.5}$$

no qual $(\cdot)$ é a função indicadora. Os coeficientes foram escolhidos para simular um cenário realista, associando menor propensão ao abandono a clientes mais antigos, com maior renda ou fidelizados. O resultado binário final foi simulado a partir de uma distribuição de Bernoulli com a probabilidade individual calculada, resultando em uma taxa de abandono de 27,15%.

3.4.2 Preparação dos Dados e Divisão Amostral

Antes do treinamento dos modelos, o conjunto de dados foi dividido em amostras de treino (70%) e teste (30%). Para esta tarefa, utilizou-se a função ‘createDataPartition‘ do pacote `caret`. Uma característica importante desta função é que ela realiza uma divisão estratificada, ou seja, a proporção da variável resposta (abandono 1 *versus* 0) é preservada em ambas as amostras. Este procedimento é uma boa prática em problemas de classificação, pois garante que os conjuntos de treino e teste sejam representativos da população original, evitando desequilíbrios amostrais que poderiam enviesar a avaliação dos modelos.

3.4.3 Modelo 1: Árvore de Decisão com Poda

O primeiro modelo foi uma árvore de decisão, implementado com o pacote `rpart`. A construção seguiu um processo de duas etapas para encontrar um modelo ótimo: crescimento e poda.

Crescimento e Poda da Árvore

Inicialmente, uma árvore maximal, foi treinada no conjunto de dados de treino, utilizando um modelo completo, que indica que todas as outras variáveis foram consideradas como preditoras. O argumento de controle "cp" (parâmetro de complexidade) foi definido como zero. O "cp" serve como um limiar para a melhoria mínima exigida para que uma divisão seja feita; ao zerá-lo, permitimos que a árvore cresça até sua extensão máxima, capturando toda a complexidade e até mesmo o ruído dos dados de treino.

Para evitar o sobreajuste e melhorar a capacidade de generalização, aplicou-se um processo de poda baseado em validação cruzada. A função `rpart` armazena internamente uma tabela de complexidade, que contém o erro de validação cruzada para diferentes valores de "cp". O valor de "cp"ótimo foi identificado como aquele que minimizou o "xerror". Uma nova árvore, uma árvore podada, foi então gerada ao podar a árvore original com este "cp"ótimo, resultando em um modelo final mais simples e robusto.

3.4.4 Modelo 2: Floresta Aleatória

O segundo modelo foi uma Floresta Aleatória, utilizando o pacote `randomForest`. O modelo foi treinado no mesmo conjunto de dados de treino com a função *rpart*, do pacote com o mesmo nome. Os principais hiperparâmetros configurados foram:

- `ntree = 500`: o modelo foi construído como um *ensemble* de 500 árvores de decisão.
- `importance = TRUE`: este argumento instrui o algoritmo a calcular e armazenar a importância de cada variável, permitindo uma análise posterior dos fatores mais influentes na predição.

3.4.5 Métricas de Avaliação de Performance

Para uma avaliação de ambos os modelos, foram utilizadas as métricas definidas na Seção 3.3, calculadas sobre o conjunto de teste:

- **Índice C (AUC):** a área sob a curva ROC, que mede o poder de discriminação do modelo.

- **Escore de Brier:** mede a acurácia das probabilidades previstas.

3.4.6 Análise e Resultado dos Modelos Ajustados

Resultados do modelo de árvore de decisão

O modelo de árvore de decisão podada alcançou uma acurácia geral de 81,64%. Uma análise mais detalhada, no entanto, revela um desequilíbrio entre as métricas de classe. O modelo demonstrou uma alta especificidade de 91,76%, o que indica uma excelente capacidade de identificar corretamente os clientes que não cancelaram. Por outro lado, a sensibilidade foi de 54,32%, mostrando que o modelo conseguiu identificar corretamente apenas pouco mais da metade dos clientes que de fato cancelaram o serviço.

Avaliando a qualidade das probabilidades, o modelo obteve um Índice C (AUC) de 0,788. Este valor, consideravelmente acima do limiar de 0,5, indica uma boa capacidade de discriminação. O escore de Brier foi de 0,146, um valor significativamente inferior ao de um modelo ingênuo, é um modelo que ignora todas as variáveis preditoras e atribui a mesma probabilidade de ocorrência do evento para todas as observações, que para uma taxa de abandono de 27,15% teria um Score de 0,198. Isso indica que as probabilidades geradas pelo modelo são bem calibradas e informativas. A análise visual da árvore (Figura 3.3), gerada com a função ‘rpart.plot‘, revela que a primeira e mais importante divisão foi feita com base na variável meses cliente, separando os clientes com menos de 22 meses dos demais.

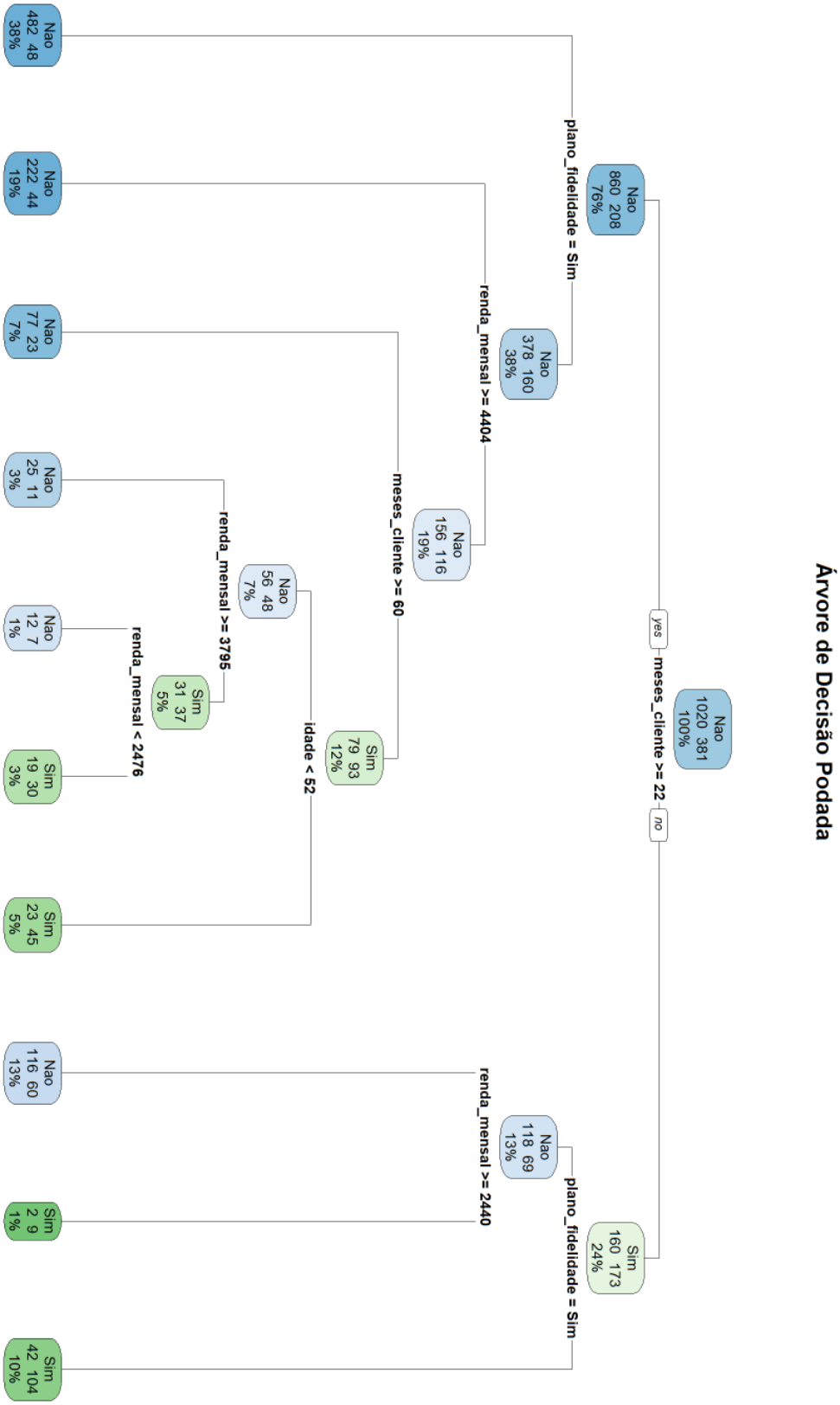


Figura 3.3: Estrutura final do modelo de Árvore de Decisão Podada.

Resultados do modelo de floresta aleatória

O modelo de floresta aleatória apresentou uma acurácia geral de 79,3%, inferior à da árvore de decisão. O desequilíbrio entre especificidade e sensibilidade também foi observado aqui, com uma especificidade de 90,39% e uma sensibilidade de 49,38%. Notavelmente, a capacidade do modelo de identificar os verdadeiros casos de abandono foi ainda menor que a do modelo de árvore única.

Em contrapartida, o índice C (AUC) foi de 0,802, ligeiramente superior ao da árvore de decisão, indicando um poder de discriminação marginalmente melhor. No entanto, seu escore de Brier foi de 0,153, um pouco maior (o que é pior) que o da árvore de decisão, sugerindo que suas previsões de probabilidade foram, em média, menos acuradas. A análise de importância de variáveis (Figura 3.4), mostrou que renda mensal e meses cliente foram os preditores mais influentes segundo as métricas de importância de variável.

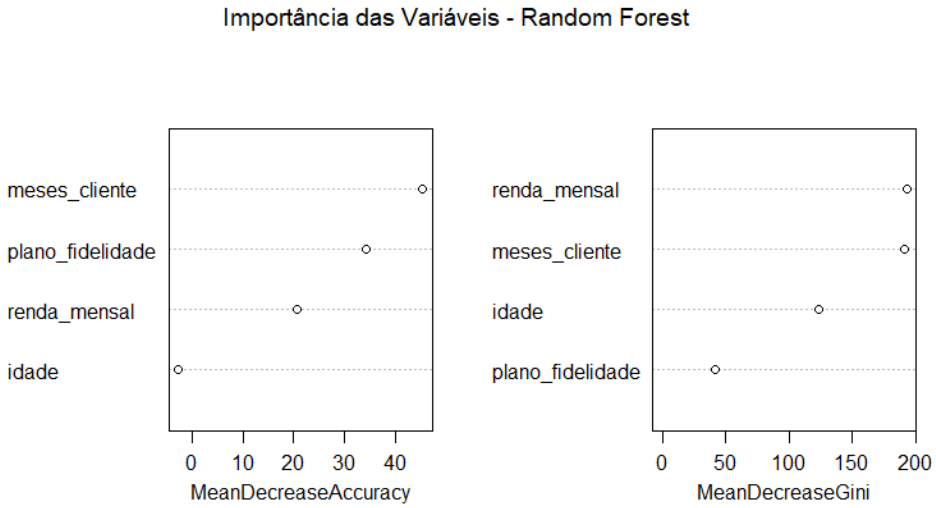


Figura 3.4: Importância das variáveis segundo o modelo de Floresta Aleatória.

Comparação e Conclusão

De forma geral, o modelo de árvore de decisão podada apresentou uma performance superior ao modelo de floresta aleatória na maioria das métricas de classificação para este conjunto de dados. A árvore única foi mais acurada (81,6% vs. 79,3%), mais sensível na detecção de abandono (54,3% vs. 49,4%) e gerou probabilidades mais bem calibradas (escore de Brier de 0,146 vs. 0,153).

A única vantagem da floresta aleatória é que o índice C marginalmente superior (0,802 vs. 0,788), indicando uma capacidade ligeiramente melhor de ranquear os clientes por

risco. Este resultado sugere que, para este problema simulado, as relações preditivas eram suficientemente claras para serem capturadas por um único conjunto de regras da árvore podada, e o processo de agregação das 500 árvores da floresta pode ter suavizado excessivamente as fronteiras de decisão, prejudicando a performance final em vez de melhorá-la.

A superioridade da árvore de decisão neste exemplo, no entanto, pode ser atribuída a uma combinação de fatores inerentes ao ambiente simulado. Primeiramente, por construção, os dados simulados continham um "sinal" muito forte e claro, com regras de decisão relativamente diretas, pois foi forçada uma relação entre eles, ou seja, a resposta final já teria influência das covariáveis criadas. Uma única árvore de decisão, otimizada via poda, é extremamente eficiente em capturar estas poucas regras dominantes e hierárquicas. Consequentemente, o principal benefício da floresta aleatória, que é a modelagem de interações complexas e a redução de ruído, não foi tão vantajoso, pois havia pouca complexidade e nenhum ruído para ser mitigado. Em um cenário com um sinal tão nítido, o processo de agregação e média da floresta pode, paradoxalmente, ter "suavizado" as fronteiras de decisão que a árvore única encontrou de forma tão eficaz. Adicionalmente, o sucesso do método de poda por validação cruzada foi notável, resultando em um modelo de árvore única que equilibrou perfeitamente a simplicidade e o poder preditivo para este problema específico.

Portanto, conclui-se que o resultado obtido, embora contra-intuitivo, não invalida a superioridade geral das Florestas Aleatórias, mas sim ilustra um cenário particular em que a simplicidade do problema favoreceu um modelo simples. O que vem de encontro com o sentido de utilizar modelos parcimoniosos, facilitando não apenas a interpretação, mas também o ajuste do modelo, sendo ele de forma menos complexa.

Capítulo 4

Florestas Aleatórias de Sobrevivência

Nos capítulos anteriores, foram explorados os fundamentos da Análise de Sobrevivência, com foco em modelos clássicos para dados de tempo até o evento. Além disso, introduzimos o conceito de florestas aleatórias (*random forests*) como uma poderosa técnica de *Machine Learning*. Este capítulo se aprofunda na interseção dessas duas áreas para apresentar as florestas aleatórias de Sobrevivência (*random survival forests*, RSF), uma metodologia moderna e robusta para modelagem de risco em cenários complexos.

O objetivo é detalhar como a estrutura de agregação de árvores das florestas aleatórias é adaptada para lidar com os desafios únicos dos dados de sobrevivência, como a presença de censura e a necessidade de estimar uma função de risco ao longo do tempo.

4.1 Unindo Machine Learning e Análise de Sobrevivência

As Florestas Aleatórias de Sobrevivência, introduzidas por Ishwaran (Ishwaran *et al.*, 2008), representam uma abordagem não paramétrica que flexibiliza as premissas rígidas dos modelos tradicionais, como o modelo de riscos proporcionais de Cox. Em vez de assumir uma relação linear entre as covariáveis e o logaritmo do risco, as RSF são capazes de capturar relações não lineares e interações complexas de forma automática.

A técnica serve para identificar variáveis prognósticas importantes e para obter predições individualizadas do risco ao longo do tempo, culminando na estimativa de uma curva de

sobrevivência para cada indivíduo.

A construção de uma Floresta de Sobrevivência integra pilares de duas grandes áreas. Da Análise de Sobrevivência, a metodologia emprega como seu principal insumo os dados de tempo até o evento; compostos pelo tempo observado e pelo indicador de estado (falha ou censura), e tem o critério para a divisão dos nós e a métrica de predição final intrinsecamente derivados de seus conceitos. A base da técnica, por sua vez, vem do aprendizado de máquina, utilizando o *bootstrap aggregating* (ou *bagging*) para gerar múltiplas amostras de treinamento e a seleção aleatória de um subconjunto de variáveis em cada nó, construindo assim um conjunto de árvores de decisão descorrelacionadas, que é a essência das Florestas Aleatórias.

4.1.1 Construção da Floresta

O algoritmo para construir uma RSF segue a estrutura geral das Florestas Aleatórias, mas com adaptações cruciais para dados de sobrevivência:

1. **Amostragem Bootstrap:** São criadas B amostras de *bootstrap* a partir do conjunto de dados de treino original. Cada amostra terá, em média, 63% das observações originais, que compõem os dados *in-bag*. As observações não selecionadas formam o conjunto *out-of-bag* (OOB), utilizado para avaliação do modelo.

2. **Crescimento das Árvores de Sobrevivência:** Para cada amostra *bootstrap*, uma árvore de sobrevivência é totalmente crescida, seguindo duas regras principais:

   - **Seleção Aleatória de Variáveis:** Em cada nó, um subconjunto aleatório de variáveis preditoras é selecionado como candidato para a divisão do nó.
   - **Critério de Divisão para Sobrevivência:** A variável e o ponto de corte escolhidos para dividir o nó são aqueles que maximizam a diferença de sobrevivência entre os nós filhos. A medida mais comum para quantificar essa diferença é a estatística do teste **Log-Rank**. Para uma candidata a divisão s em uma variável X, que separa um nó em dois filhos D_1 e D_2, a estatística do teste Log-Rank $L(X, s)$ é calculada. A divisão ótima (X^*, s^*) é aquela que maximiza essa estatística:

$$(X^*, s^*) = \arg\max_{X,s} |L(X, s)|$$

3. **Estimação nos Nós Terminais:** As árvores são crescidas até que um critério de parada seja atingido (ex: número mínimo de eventos em um nó terminal). Em cada nó folha h, a **Função de Risco Acumulada** (*Cumulative Hazard Function*, CHF) é estimada de forma não paramétrica. Sendo $t_{1,h} < t_{2,h} < \cdots < t_{N(h),h}$ os tempos de evento distintos no nó h, $d_{l,h}$ o número de eventos e $Y_{l,h}$ o número de indivíduos em risco no tempo $t_{l,h}$, a CHF é calculada pelo estimador de Nelson-Aalen:

$$\hat{H}_h(t) = \sum_{t_{l\ h} \leq t} \frac{d_{l,h}}{Y_{l,h}}$$

4. **Predição Agregada (Ensemble):** Para um novo indivíduo com vetor de covariáveis $\mathbf{x}$, a predição da floresta é obtida agregando-se as predições de todas as B árvores. O indivíduo é passado por cada árvore b, caindo em um nó terminal específico, resultando em uma estimativa de CHF, $\hat{H}_b(t|\mathbf{x})$. A CHF do *ensemble* é a média das CHFs de cada árvore:

$$\hat{H}_e(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{H}_b(t|\mathbf{x})$$

A partir da CHF agregada, a curva de sobrevivência individual, $\hat{S}_e(t|\mathbf{x})$, é diretamente calculada pela relação fundamental:

$$\hat{S}_e(t|\mathbf{x}) = \exp(\quad \hat{H}_e(t|\mathbf{x}))$$

4.2 Exemplo do algoritmo de criação da árvore

Para ilustrar o algoritmo de construção de árvores de sobrevivência, vamos construir manualmente uma única árvore a partir de um pequeno conjunto de dados. O objetivo é demonstrar como o critério de divisão é aplicado com cálculos detalhados e como as estimativas de sobrevivência são geradas nos nós terminais.

4.2.1 Conjunto de Dados de Exemplo

Considere o seguinte conjunto de dados com 10 pacientes, contendo o tempo de seguimento, o estado (1 para evento, 0 para censura), a idade do paciente e o tipo de tratamento recebido (A ou B).

Tabela 4.2.1: Conjunto de dados fictício para construção da árvore.

Paciente	*Tempo*	*Estado*	*Idade*	*Tratamento*
1	5	1	65	A
2	8	0	70	B
3	10	1	55	A
4	12	1	58	B
5	15	0	62	B
6	18	1	75	A
7	20	0	45	B
8	22	1	48	A
9	25	0	52	B
10	30	1	68	A

Nosso objetivo é encontrar a primeira divisão da árvore (no nó raiz) que melhor separe os pacientes em grupos com diferentes perfis de sobrevivência.

4.2.2 Passo 1: Avaliando a Divisão por Tratamento

Dividimos os dados pelos níveis da variável `Tratamento`: Grupo A e Grupo B. A lógica do teste Log-Rank é, para cada tempo em que ocorre um evento, comparar o número de eventos observados em um grupo com o número de eventos que seriam esperados se não houvesse diferença entre os grupos. A estatística do teste é dada por:

$$L = \frac{(O \quad E)^2}{V}, \quad \text{onde } O = \sum O_i, E = \sum E_i \text{ e } V = \sum V_i$$

Análise Tempo a Tempo para `Tratamento`

- *Tempo t=5:* 10 pacientes em risco ($n_i = 10$), 5 no Grupo A ($n_{A,i} = 5$). 1 evento ($d_i = 1$) no Grupo A ($O_{A,i} = 1$).

$$E_{A,5} = d_i \times \frac{n_{A,i}}{n_i} = 1 \times \frac{5}{10} = 0,500$$

$$V_5 = \frac{d_i(n_i \quad d_i)n_{A,i}(n_i \quad n_{A,i})}{n_i^2(n_i \quad 1)} = \frac{1(9)(5)(5)}{10^2(9)} = 0,250$$

- *Tempo t=10:* 8 pacientes em risco ($n_i = 8$), 4 no Grupo A ($n_{A,i} = 4$). 1 evento ($d_i = 1$) no Grupo A ($O_{A,i} = 1$).

$$E_{A,10} = 1 \times \frac{4}{8} = 0,500 \quad | \quad V_{10} = \frac{1(7)(4)(4)}{8^2(7)} = 0,250$$

- *Tempo t=12:* 7 pacientes em risco ($n_i = 7$), 3 no Grupo A ($n_{A,i} = 3$). 1 evento ($d_i = 1$) no Grupo B ($O_{A,i} = 0$).

$$E_{A,12} = 1 \times \frac{3}{7} \approx 0,429 \quad | \quad V_{12} = \frac{1(6)(3)(4)}{7^2(6)} \approx 0,245$$

- *Tempo t=18:* 5 pacientes em risco ($n_i = 5$), 3 no Grupo A ($n_{A,i} = 3$). 1 evento ($d_i = 1$) no Grupo A ($O_{A,i} = 1$).

$$E_{A,18} = 1 \times \frac{3}{5} = 0,600 \quad | \quad V_{18} = \frac{1(4)(3)(2)}{5^2(4)} = 0,240$$

- *Tempo t=22:* 3 pacientes em risco ($n_i = 3$), 2 no Grupo A ($n_{A,i} = 2$). 1 evento ($d_i = 1$) no Grupo A ($O_{A,i} = 1$).

$$E_{A,22} = 1 \times \frac{2}{3} \approx 0,667 \quad | \quad V_{22} = \frac{1(2)(2)(1)}{3^2(2)} \approx 0,222$$

- *Tempo t=30:* 1 paciente em risco ($n_i = 1$), 1 no Grupo A ($n_{A,i} = 1$). 1 evento ($d_i = 1$) no Grupo A ($O_{A,i} = 1$).

$$E_{A,30} = 1 \times \frac{1}{1} = 1,000 \quad | \quad V_{30} = 0$$

Tabela 4.2.2: Resumo do cálculo para a variável `Tratamento`.

Tempo (t_i)	d_i	n_i	$n_{A,i}$	$O_{A,i}$	$E_{A,i}$	V_i
5	1	10	5	1	0,500	0,250
10	1	8	4	1	0,500	0,250
12	1	7	3	0	0,429	0,245
18	1	5	3	1	0,600	0,240
22	1	3	2	1	0,667	0,222
30	1	1	1	1	1,000	0,000
Total	*6*			*O=5*	*E=3,696*	*V=1,207*

Cálculo final da estatística:

$$L(\text{Tratamento}) = \frac{(5 \quad 3,696)^2}{1,207} = \frac{1,700}{1,207} \approx 1,408$$

4.2.3 Passo 2: Avaliando a Divisão por Idade

O processo seria repetido para todos os pontos de corte possíveis na variável `Idade`. Como exemplo, o cálculo para o ponto de corte `Idade` ≤ 60 resultou em $L(\text{Idade}, 60) \approx 0,047$. Assumimos que nenhum outro ponto de corte gerou uma estatística maior que a da variável `Tratamento`.

4.2.4 Passo 3: Selecionando a Melhor Divisão (Nó Raiz)

Comparamos a maior estatística Log-Rank de cada variável:

- Melhor divisão para `Tratamento`: Estatística Log-Rank = **1,408**
- Melhor divisão para `Idade`: Estatística Log-Rank < 1,408

A primeira divisão da árvore será na variável `Tratamento`, criando os nós filhos:

- *Nó Filho 1 (Esquerda):* Pacientes com Tratamento = A. {1, 3, 6, 8, 10}.
- *Nó Filho 2 (Direita):* Pacientes com Tratamento = B. {2, 4, 5, 7, 9}.

4.2.5 Passo 4: Crescendo a Árvore - Segunda Camada

Com um critério de parada `nodesize=3`, o algoritmo continua.

Análise do Nó Filho 1 (Tratamento = A)

Testamos a variável `Idade` para dividir os 5 pacientes deste nó. Com o ponto de corte `Idade` ≤ 60, a estatística Log-Rank foi $L \approx 0,074$. Assumimos que este é o melhor corte e, relaxando o critério de `nodesize` para fins didáticos, a divisão é feita.

Análise do Nó Filho 2 (Tratamento = B)

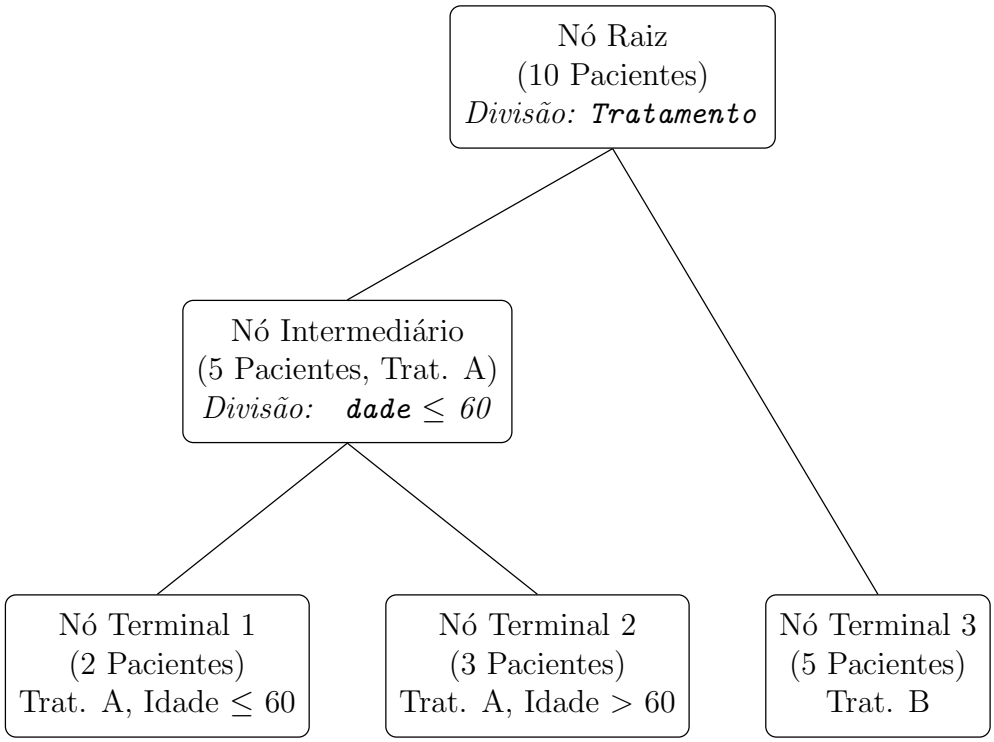
Neste nó, há apenas um evento. O algoritmo não consegue encontrar uma divisão válida, tornando-o um nó terminal.

4.2.6 Passo 5: Estrutura Final da Árvore

A estrutura final da árvore, detalhada abaixo, também pode ser visualizada na Figura 4.1.

- *Nó Raiz:* (10 pacientes) → Divisão por `Tratamento`
  - *Nó Intermediário (Tratamento=A):* (5 pacientes) → Divisão por `Idade <= 60`
    - *Nó Terminal 1:* (2 pacientes) {3, 8} - Trat. A, Idade ≤ 60
    - *Nó Terminal 2:* (3 pacientes) {1, 6, 10} - Trat. A, Idade ¿ 60
  - *Nó Terminal 3:* (5 pacientes) {2, 4, 5, 7, 9} - Trat. B

Figura 4.1: Representação visual da estrutura final da árvore.



4.2.7 Passo 6: Estimação da CHF nos Nós Terminais

Nó Terminal 1 (Trat. A, Idade ≤ 60)

Pacientes: {3(10,1), 8(22,1)}

- Tempo 10: 1 evento, 2 em risco. Risco = 1/2. CHF = 0,500.
- Tempo 22: 1 evento, 1 em risco. Risco = 1/1. CHF = 0,500 + 1,000 = 1,500.

Nó Terminal 2 (Trat. A, Idade > 60)

Pacientes: {1(5,1), 6(18,1), 10(30,1)}

- Tempo 5: 1 evento, 3 em risco. Risco = 1/3. CHF = 0,333.

- Tempo 18: 1 evento, 2 em risco. Risco = 1/2. CHF = 0,333 + 0,500 = 0,833.
- Tempo 30: 1 evento, 1 em risco. Risco = 1/1. CHF = 0,833 + 1,000 = 1,833.

Nó Terminal 3 (Trat. B)

Pacientes: {2(8,0), 4(12,1), 5(15,0), 7(20,0), 9(25,0)}

- Tempo 12: 1 evento, 4 em risco. Risco = 1/4. CHF = 0,250.

4.2.8 Passo 7: Prevendo um Novo Paciente (Paciente 11)

Usamos a árvore para estimar o risco de um novo paciente com `Idade = 58` e `Tratamento = 'A'`.

1. *Nó Raiz:* A divisão é por `Tratamento`. Como o paciente é 'A', ele segue para o nó da esquerda.
2. *Nó Intermediário:* A divisão é por `Idade <= 60`. Como 58 $\leq$ 60, ele segue para o nó da esquerda, chegando ao *Nó Terminal 1.*

Conclusão: O Paciente 11 pertence ao Nó Terminal 1. Sua predição de risco individual é a CHF daquele nó (CHF = 0,500 até t=10; CHF = 1,500 após t=22).

4.3 Avaliação de Desempenho e Importância de Variáveis

Uma das grandes vantagens das Florestas Aleatórias é sua capacidade de gerar uma estimativa de erro de predição imparcial durante o próprio treinamento, utilizando os dados *out-of-bag* (OOB). Para cada observação i, sua predição é calculada utilizando apenas as árvores para as quais i estavam no conjunto OOB. O erro de predição do *ensemble* é então calculado comparando os resultados previstos com os observados para todo o conjunto de dados, sendo o Índice de Concordância (Índice C) a métrica mais comum na análise de sobrevivência.

Outro subproduto valioso é a medida de **Importância de Variáveis** (*Variable Importance*, VIMP). Conforme proposto por Ishwaran *et al.* (2008), a importância de uma variável X_j é calculada da seguinte forma:

1. O erro de predição OOB da floresta é calculado (ex: 1 - Índice C).

2. Em seguida, os valores da variável X_j são permutados aleatoriamente no conjunto de dados OOB.

3. O erro de predição OOB é recalculado com os dados permutados.

4. A VIMP para X_j é a diferença entre o erro após a permutação e o erro original.

Uma queda acentuada no desempenho (ou seja, um valor de VIMP positivo e alto) indica que a variável é altamente preditiva. Isso oferece uma maneira robusta e eficaz de classificar as variáveis segundo sua contribuição para o modelo.

Dessa forma, a Floresta Aleatória de Sobrevivência combina a robustez e a capacidade preditiva do *ensemble* de árvores com as ferramentas estatísticas consagradas da análise de sobrevivência, oferecendo um modelo poderoso e flexível para a análise de dados de tempo até o evento.

Capítulo 5

Aplicação de Florestas de Sobrevivência

Este capítulo dedica-se à aplicação prática e à comparação de desempenho entre a modelagem clássica, via modelo de riscos proporcionais de Cox, e a abordagem de aprendizado de máquina, utilizando florestas aleatórias de sobrevivência. Para tal, foi utilizado um conjunto de dados (reais) do setor de telecomunicações, visando prever o tempo até o cancelamento de serviço (*churn*).

5.1 O Conjunto de Dados: Telco Customer Churn

O estudo baseia-se no conjunto de dados *Telco Customer Churn*, disponibilizado publicamente na plataforma Kaggle (Blastchar, 2018). A base original compreende 7.043 observações, em que cada linha representa um cliente único e as colunas descrevem atributos demográficos, serviços contratados e o *status* de permanência na empresa.

5.1.1 Adaptação para Análise de Sobrevivência

Originalmente estruturado para classificação binária, o conjunto de dados exigiu adaptações fundamentais para o contexto de análise de sobrevivência, onde o interesse reside não apenas na ocorrência do evento, mas no tempo até sua ocorrência. O par de variáveis que definem a resposta foram definidas como:

- **Tempo (*tenure*):** Representa a variável de tempo T, medida em meses de permanência do cliente na empresa. Observações com $T = 0$ foram removidas, dado

que não contribuem com tempo de exposição ao risco.

- **Status (*Churn*):** Define a variável indicadora de falha δ. Clientes que cancelaram o serviço (`Yes`) foram codificados como falha ($\delta = 1$). Clientes ativos ao final do período de coleta (`No`) foram tratados como censura à direita ($\delta = 0$).

5.1.2 Covariáveis e Pré-processamento

Após a exclusão do identificador único e o tratamento de dados faltantes (exclusão de registros com valores nulos na variável), o conjunto final foi composto por 18 covariáveis preditoras. Estas incluem dados demográficos (gênero, idade, dependentes), características dos serviços contratados (telefone, internet, segurança, *backup*) e informações financeiras (tipo de contrato, método de pagamento).

Para garantir a comparabilidade entre as abordagens, aplicou-se um protocolo comum de pré-processamento:

1. **Tratamento de Colinearidade:** Níveis redundantes em variáveis categóricas, especificamente nos serviços de internet, foram unificados para reduzir a multicolinearidade.

2. **Particionamento:** Os dados foram divididos aleatoriamente em conjuntos de treino (70%, $n = 4922$) e teste (30%, $n = 2110$), preservando a proporção de eventos.

5.2 Abordagem 1: Modelo de Riscos Proporcionais de Cox

A modelagem semiparamétrica de Cox foi conduzida de maneira iterativa, priorizando a parcimônia e a validade das suposições estatísticas.

5.2.1 Seleção de Variáveis e Diagnóstico

Inicialmente, ajustou-se um modelo contendo todas as covariáveis disponíveis. Procedeu-se com a seleção de variáveis via (*backward elimination*), removendo sequencialmente preditores não significativos ($p > 0,05$) até a obtenção de um modelo reduzido.

A validação da premissa fundamental de riscos proporcionais foi realizada através do teste de resíduos de Schoenfeld. O diagnóstico revelou violações globais e específicas

significativas ($p < 0,05$) para as variáveis: Contrato, Múltiplas Linhas, Serviço de Internet e Método de Pagamento.

5.2.2 Estrati cação e Modelo Final

Para contornar a violação da proporcionalidade sem descartar preditores relevantes, adotou-se a estratégia de **Cox Estrati cado**. Nesta abordagem, permite-se que a função de risco base, $\lambda_0(t)$, varie para cada estrato k formado pelas variáveis problemáticas, enquanto se assume que o efeito das demais covariáveis ( ) permanece constante entre os estratos. A função de risco para um indivíduo no estrato k é dada por:

$$h(t|\mathbf{X}) = h_{0k}(t) \exp \quad ^T\mathbf{X})$$

O modelo final resultou na estratificação pelas variáveis Contrato, Internet, Pagamento e Múltiplas Linhas, mantendo como preditores diretos apenas: Ter Parceiro, Segurança Online e *Online Backup*.

5.2.3 Resultados do Modelo de Cox

A Tabela 5.2.1 apresenta as estimativas para o modelo final estratificado. Observa-se que as variáveis remanescentes atuam como fatores de proteção.

Tabela 5.2.1: Estimativas do Modelo de Cox Estratificado (Dados de Treino).

Variável	**Coef**	**HR**	**IC 95% (HR)**	**p-valor**
Tem Parceiro (Sim)	-0,547	0,579	(0,51 - 0,65)	$< 0,001$
Segurança Online (Sim)	-0,732	0,481	(0,41 - 0,57)	$< 0,001$
Online Backup (Sim)	-0,730	0,482	(0,42 - 0,55)	$< 0,001$

A interpretação direta sugere, por exemplo, que a contratação de Segurança Online reduz o risco de cancelamento em aproximadamente 52% (1 $0,481$), mantendo constantes os estratos e demais variáveis. O desempenho preditivo deste modelo será discutido na seção comparativa.

5.3 Abordagem 2: Floresta Aleatória de Sobrevivência

Diferentemente do modelo de Cox, a floresta aleatória de sobrevivência é um método de aprendizado de conjunto (*ensemble*) totalmente não-paramétrico. Esta abordagem

constrói centenas de árvores de sobrevivência independentes e agrega suas previsões, permitindo capturar interações complexas e efeitos não-lineares sem a necessidade de suposições rígidas sobre a distribuição dos dados ou a proporcionalidade dos riscos.

Foram ajustados dois modelos RSF para fins de comparação:

1. **RSF Completo:** Utilizando todas as 18 covariáveis disponíveis.

2. **RSF Seletivo:** Utilizando apenas as 7 covariáveis presentes na fórmula do modelo final de Cox (estratificadoras + preditoras).

5.3.1 Mecânica da Floresta e o Critério de Divisão Log-Rank

O algoritmo opera através do crescimento de B árvores (neste estudo, $ntree = 1000$). Para cada árvore, o processo de divisão dos nós é o motor principal da aprendizagem.

Em cada nó da árvore, o algoritmo seleciona aleatoriamente um subconjunto de variáveis candidatas e busca o melhor ponto de corte c para uma variável x, de modo a maximizar a diferença de sobrevivência entre os dois nós filhos resultantes (esquerdo e direito).

A regra de divisão utilizada foi o **teste de Log-Rank**. Matematicamente, para um determinado ponto de corte que divide os dados do nó pai em dois grupos, o algoritmo calcula a estatística de log-rank (L), que compara o número de eventos observados (O_i) com o número de eventos esperados (E_i) sob a hipótese nula de que as curvas de sobrevivência são idênticas:

$$L(x, c) = \frac{\sum_{j=1}^{k} (O_{j,1} \quad E_{j,1})}{\sqrt{Var(\sum (O_{j,1} \quad E_{j,1}))}}$$

O algoritmo itera sobre possíveis divisões e seleciona aquela que maximiza o valor absoluto $|L(x, c)|$. Maximizar essa estatística significa maximizar a heterogeneidade entre os nós filhos, garantindo que indivíduos com perfis de sobrevivência distintos sejam segregados em ramos diferentes da árvore. Esse processo continua recursivamente até que um critério de parada (como o tamanho mínimo do nó terminal, definido aqui como 15) seja atingido.

5.3.2 Estimação e Agregação do Ensemble

Ao atingir um nó terminal (folha) h, a árvore fornece uma estimativa da Função de Risco Acumulado (*Cumulative Hazard Function*, CHF), denotada por $\hat{H}_h(t)$, calculada usualmente pelo estimador de Nelson-Aalen baseado nos indivíduos que caíram naquela folha.

A predição final do modelo de floresta para um novo indivíduo com covariáveis $\mathbf{x}_i$ é obtida fazendo-o percorrer todas as árvores do conjunto. A CHF do ensemble é a média das CHFs de todas as árvores:

$$\hat{H}_e(t|\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{H}_b(t|\mathbf{x}_i)$$

A partir desta função agregada, obtém-se a probabilidade de sobrevivência estimada $\hat{S}_e(t|\mathbf{x}_i) = \exp(\quad \hat{H}_e(t|\mathbf{x}_i))$. É essa capacidade de gerar curvas de sobrevivência individualizadas, baseadas na média de estruturas complexas de árvore, que confere robustez ao método.

5.3.3 Resultados dos Modelos RSF

A Tabela 5.3.1 resume as características estruturais e o desempenho interno (OOB - *Out of Bag*) dos modelos gerados.

Tabela 5.3.1: Resumo dos Modelos RSF Ajustados (Treino, n=4922).

Parâmetro / Métrica	**RSF Completo**	**RSF Seletivo**
Covariáveis Disponíveis	18	7
Variáveis testadas por nó (*mtry*)	5	3
Média de Nós Terminais	175,27	120,56
C-Index OOB (Estimado)	**0,8876**	**0,8377**

A maior complexidade do modelo completo (média de 175 nós terminais) reflete a utilização de mais variáveis informativas para particionar o espaço de dados. O *C-Index* OOB de 0,8876 sugere uma capacidade discriminatória muito alta, superior à versão seletiva (0,8377), indicando que as variáveis descartadas na análise de Cox continham informações preditivas valiosas para a estrutura não-linear da floresta.

5.4 Comparação de Desempenho: Cox vs. RSF

A avaliação final compara a capacidade de generalização dos modelos no conjunto de teste ($n = 2110$), utilizando duas métricas complementares: o Índice de Concordância (*C-Index*), que mede a discriminação (capacidade de ordenação correta de risco), e o Escore de Brier (*Brier Score*), que mede a calibração (precisão da probabilidade prevista).

Tabela 5.4.1: Comparação de Desempenho Preditivo (Conjunto de Teste).

Modelo	**C-Index**	**Brier Score (t=72)**
Referência (Kaplan-Meier)	0,500	0,145
Cox (Estratificado)	0,652	0,097
RSF Seletivo	0,826	0,089
RSF Completo	**0,879**	**0,078**

5.4.1 Discussão dos Resultados

Os resultados evidenciam uma superioridade expressiva da abordagem de aprendizado de máquina sobre a modelagem estatística clássica para este conjunto de dados.

1. **Discriminação Superior:** O modelo RSF Completo atingiu um *C-Index* de 0,879, representando um ganho de desempenho de aproximadamente 35% sobre o modelo de Cox (0,652). Mesmo quando restrito às mesmas variáveis do Cox (RSF Seletivo), a floresta obteve 0,826. Isso demonstra que a estrutura hierárquica das árvores consegue capturar nuances nos dados que a estrutura linear (nos coeficientes) do Cox não consegue, mesmo com estratificação.

2. **Calibração e Erro:** O *Brier Score* avaliado em $t = 72$ meses confirma a robustez das probabilidades estimadas. O erro quadrático médio de predição da RSF Completa (0,078) foi inferior ao do Cox (0,097) e quase metade do erro do modelo nulo (0,145).

3. **Flexibilidade vs. Suposições:** O desempenho inferior do Cox pode ser atribuído à rigidez de suas suposições. A necessidade de estratificação para cumprir a premissa de riscos proporcionais fragmentou a amostra, reduzindo o poder estatístico. Em contraste, a RSF, através do critério de divisão *log-rank*, adaptou-se automaticamente às mudanças nas taxas de risco ao longo do tempo e entre subgrupos, sem exigir intervenção manual ou verificação de premissas.

Em conclusão, para o problema de *churn* na telecomunicação, onde as interações entre serviços e perfil do cliente são complexas e multifacetadas, a Floresta de Sobrevivência provou ser a ferramenta mais adequada para a predição individualizada de risco. A floresta aleatória de sobrevivência revelou-se uma ferramenta robusta e suficiente para lidar com a complexidade inerente a dados reais. Em suma, embora o Modelo de Cox permaneça valioso para a interpretação causal e quantificação de risco relativo, a floresta aleatória de sobrevivência estabelece-se como a escolha preferencial quando o objetivo primário é a precisão da predição individualizada em cenários de dados modernos e multidimensionais.

Capítulo 6

Conclusão

Por fim, podemos concluir este trabalho de graduação, interligando todos os temas estudados. O objetivo inicial do trabalho foi alcançado, fazendo uma comparação entre um modelo clássico para análise de dados de sobrevivência e um modelo de *machine learning*, florestas aleatórias, que tem o foco em análise de dados complexos. Foi revisitado vários conceitos básicos sobre ambas as metodologias e abordado também um exemplo para cada uma delas, foram desenvolvidas suas em suas áreas teóricas e práticas, mostrando como os modelos são aplicados e como podemos avaliar cada um deles. Na segunda parte, após a introdução de cada método que utilizamos, foi abordado um tema novo, o qual seu objetivo era interligar os dois temas anteriores, juntando a metodologia clássica com *machine learning*, assim apresentando florestas aleatórias de sobrevivência. Ao estudar este novo método, vimos em sua teoria como funciona o algoritmo e sua iteração para construir uma árvore e ainda sim utilizar métodos clássicos, como *logrank*, para auxiliar na construção da árvore, além de que ao fim de cada nó, é calculado uma função de risco acumulada para cada invidio com aquelas características, assim podendo chegar uma função de sobrevivência, tal qual como Cox. Então, utilizando um conjunto de dados reais foi feita a comparação entre a nova técnica apresentada e o modelo clássico de Cox, e embora para este conjunto de dados a floresta aleatória de sobrevivência tenha obtido resultados melhores, não é uma conclusão de que de fato é um modelo que sempre será melhor, o objetivo do estudo, tipo dos dados, tipo de resposta que se espera pode influenciar em qual modelo utilizar, então em suma, para cada vertente que o estudo deseja seguir, pode mudar o melhor modelo a se utilizar. Para concluir, fazendo uma ultima ressalva ao modelo de florestas aleatórias de sobrevivência, é um grande avanço em metodologias estatísticas, mostrando que técnicas de *machine learning* estão evoluindo

para ferramentas robustas que por sua vez não tem tantas suposições e lidam bem com a situação atual, onde temos dados muitos complexos (muitas observações e variáveis), além de ter predições precisas e métricas de fáceis avaliação; além do mais, se mostrando como uma metodologia valiosa e que está em constante evolução.

Capítulo 7

Apêndice

Este apêndice detalha os conceitos citados no trabalho que não foram abordados detalhadamente.

Apêndice A: Estrutura de uma Árvore de Decisão

Referenciando o capitulo 3.1.

- **Nó Raiz:** O nó inicial da árvore. Ele contém todo o conjunto de dados de treinamento e representa a primeira divisão (o primeiro teste) que será aplicada.
- **Nó Interno (ou Nó de Decisão):** Qualquer nó da árvore que não seja um nó folha. Cada nó interno representa um teste em uma das variáveis preditoras e se divide em nós filhos com base no resultado desse teste.
- **Nó Folha (ou Nó Terminal):** Os nós finais da árvore, que não se dividem mais. Cada nó folha representa uma região do espaço de covariáveis e contém a predição final do modelo para as observações que chegam até ele (por exemplo, a classe majoritária para classificação).
- **Ramo:** A conexão entre dois nós. Cada ramo representa o resultado de um teste ("idade $<$ 30" ou "idade $>=$ 30") e define o caminho que uma observação segue através da árvore.
- **Profundidade:** O número de ramos no caminho mais longo do nó raiz até o nó folha mais distante. A profundidade é uma medida da complexidade da árvore; árvores mais profundas são mais complexas e mais propensas a sobreajuste.

Apêndice B: Construção e Avaliação de Modelos

- **Dados de Treino:** O subconjunto dos dados originais que é utilizado para construir e ajustar ("treinar") o modelo. O algoritmo aprende os padrões e as relações presentes nestes dados.

- **Dados de Teste:** Um subconjunto dos dados que não foi utilizado durante o treinamento. Ele serve para avaliar o desempenho final do modelo em dados novos e não vistos, fornecendo uma estimativa de sua capacidade de generalização.

- **Conjunto de Validação:** Um subconjunto de dados, separado do treino e do teste, usado especificamente para o ajuste de hiperparâmetros do modelo (como a profundidade máxima da árvore ou o parâmetro de complexidade para a poda).

- **Validação Cruzada (*Cross Validation*):** Uma técnica robusta para avaliar a performance de um modelo. Os dados de treino são divididos em k partes (ou "folds"). O modelo é treinado k vezes, cada vez utilizando k 1 folds para treino e o fold restante para validação. A performance final é a média dos resultados obtidos em cada iteração.

- **Árvore Totalmente Crescida:** Uma árvore de decisão que foi expandida até sua profundidade máxima, ou seja, até que cada nó folha seja o mais "puro" possível (idealmente, contendo observações de uma única classe) ou até que não exista mais divisões possíveis. Tais árvores quase sempre sofrem de sobreajuste.

- **Poda (*Pruning*):** O processo de reduzir o tamanho de uma árvore de decisão, removendo seções (ramos e folhas) que fornecem pouco poder preditivo. É a principal técnica para combater o sobreajuste em árvores, resultando em um modelo mais simples e com melhor capacidade de generalização.

- **Hiperparâmetros:** Configurações externas ao modelo, definidas pelo usuário *antes* do processo de treinamento. Eles não são aprendidos a partir dos dados, mas sim controlam como o algoritmo de aprendizado funciona. Exemplos incluem o número de árvores em uma floresta aleatória ou o parâmetro de complexidade para a poda de uma árvore. A escolha de bons hiperparâmetros é crucial para o desempenho do modelo e geralmente é feita usando um conjunto de validação.

- **Métodos de *Ensemble*:** Abordagens que combinam as predições de múltiplos modelos (chamados de "modelos base") para obter uma predição final mais acurada e estável do que a de qualquer modelo individual. A floresta aleatória é um exemplo clássico.

Apêndice C: Pacotes utilizados no R

O desenvolvimento desta análise foi suportado por diversos pacotes do R, cada um com uma finalidade específica. A seguir, descreve-se brevemente a função de cada pacote principal utilizado.

survival : O pacote fundamental para análise de sobrevivência. Foi utilizado para criar objetos de sobrevivência, ajustar o modelo de Cox e testar suas suposições.

randomForestSRC : A biblioteca central para a modelagem de *machine learning*. Ela fornece a função para treinar os modelos floresta aleatória de sobrevivência (RSF).

dplyr : A principal ferramenta para a manipulação e preparação dos dados. Foi essencial para limpar, filtrar, selecionar e criar novas variáveis de forma eficiente.

ggplot2 : O pacote padrão-ouro para a visualização de dados no R. Utilizado para a criação de gráficos de alta qualidade para a análise exploratória e apresentação de resultados.

pec : Pacote crucial para a avaliação de performance dos modelos. Foi utilizado para calcular e comparar o Brier Score (IBS) e as curvas de erro de previsão entre os modelos Cox e RSF.

readr : Parte do Tidyverse, foi utilizado para a leitura rápida e eficiente dos dados do arquivo `.csv` para o R (`read_csv()`).

tidyr : Um pacote auxiliar para "arrumar" os dados, facilitando a organização e reestruturação dos *data frames* para a análise.

Referências Bibliográ cas

Akram, S.; Ann, Q. U. Newton-Raphson method. *International Journal of Scienti c & Engineering Research*, v. 6(7), p. 1748–1752, 2015.

Blastchar. Telco customer churn. Kaggle Dataset, 2018. URL https://www.kaggle.com/datasets/blastchar/telco-customer-churn. Acessado em: 17 de novembro de 2025.

Breiman, L. Random forests. *Machine learning*, v. 45(1), p. 5–32, 2001.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, v. 78(1), p. 1–3, 1950.

Colosimo, E. A.; Giolo, S. R. *Análise de sobrevivência aplicada*. Edgard Blücher, 2006.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 34(2), p. 187–202, 1972.

Fisher, R. A. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 22(5), p. 700–725, 1925.

Graf, E.; Schmoor, C.; Sauerbrei, W.; Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, v. 18(17-18), p. 2529–2545, 1999.

Grambsch, P. M.; Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, v. 81(3), p. 515–526, 1994.

Greenwood, M. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, v. 33, p. 1–26, 1926.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; Rosati, R. A. Evaluating the yield of medical tests. *JAMA*, v. 247(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(3), p. 841–860, 2008.

Izbicki, R.; dos Santos, T. M. *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, 2020.

Kalbfleisch, J. D.; Prentice, R. L. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.

Kaplan, E. L.; Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53(282), p. 457–481, 1958.

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, v. 50(3), p. 163–170, 1966.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika*, v. 69(1), p. 239–241, 1982.