Métodos de Árvores de Decisão em Análise de Sobrevivência

Uma aplicação a dados de câncer

Vinicius Santos Oliveira

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Ciências

Programa: Estatística
Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima

São Paulo
Dezembro de 2023

Métodos de Árvores de Decisão em Análise de Sobrevivência

Uma aplicação a dados de câncer

Vinicius Santos Oliveira

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 8 de Dezembro de 2023.

Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão julgadora:

Prof. Dr. Antonio de Carlos Pedroso de Lima – IME-USP
Prof. Dr. Rinaldo Artes – Insper
Prof. Dr. Victor Silva Ritter – Stanford

O conteúdo deste trabalho é publicado sob a licença CC BY 4.0
(Creative Commons Attribution 4.0 International License)

Em memória de José Franco e Valdeci que sempre me motivaram imensuravelmente. Estejam em paz.

Agradecimentos

Eu não me aposento enquanto o futuro não for presente.
— Black Alien

Agradeço a minha esposa Rafaela por ter me apoiado, me suportado e me incentivado a todo custo. Não permitiu que eu desistisse e foi minha luz e calmaria durante todo esse processo.

Agradeço aos meus pais Zilma e Antonio por terem me incentivado aos estudos e por acreditarem em mim. Sempre serão meu pilares e exemplos.

Agradeço ao meu irmão Felipe por ter sido sempre meu grande amigo e estar disponível a todo momento. É meu ponto de motivação em muitos momentos da minha vida.

Agradeço a minha madrinha e grande amiga profª Marta Yukie Baba por me motivar a entrar no programa de pós-graduação. As suas aulas e seus conselhos foram determinantes para eu entrar no mestrado.

Agradeço aos meus amigos Gabriel Camarotto, Lucas Daneu e Matheus Jordão por terem sido ombros em meus momentos difíceis e festejarem comigo nos momentos de alegrias. Realmente, nem mesmo a força do tempo é capaz de destruir uma amizade.

Agradeço ao meu grande amigo e ex-gestor Gabriel Antunes que permitiu a realização da pós-graduação durante o meu trabalho. Me aconselhou durante as nossas interações e me apoiou para que eu pudesse dar esse passo na minha carreira.

Agradeço ao meu orientador Prof. Dr. Antonio Carlos Pedroso de Lima por ter sido paciente, por toda ajuda durante a construção desse trabalho e por todo conselho dado.

Resumo

Vinicius Santos Oliveira. **Métodos de Árvores de Decisão em Análise de Sobrevivência: *Uma aplicação a dados de câncer***. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

A análise de sobrevivência é um conjunto de técnicas estatísticas amplamente utilizadas para analisar tempos até a ocorrência de um ou mais eventos. Dentre dos possíveis métodos de modelagem preditiva para dados de sobrevivência, as árvores de decisão têm destaque devido à sua capacidade de modelar relações complexas entre as covariáveis e a ocorrência do evento de interesse. Neste trabalho, são estudadas técnicas de árvore de decisão para dados censurados, revisando suas metodologias, avaliando suas vantagens e desvantangens e apresentando extensões com uso de *ensembles*. Por fim, as diferentes técnicas são aplicadas ao conjunto de dados do ICESP e comparadas com a abordagem usual baseada no modelo de riscos proporcionais de Cox usando métricas de avaliação de performance e técnicas de validação cruzada.

Palavras-chave: Análise de Sobrevivência. Árvore de Decisão. Árvore de Sobrevivência. Ensembles. Predição.

Abstract

Vinicius Santos Oliveira. **Decision Tree Methods in Survival Analysis: *An application to cancer dataset***. Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

Survival analysis is a set of statistical techniques widely used to analyze the time to the occurrence of one ore more events. Among the possible predictive modeling methods for survival data, decision trees stand out for their ability to model complex relationships between covariates and the occurrence of the event of interest. In this work, decision tree techniques for censored data are studied, their methods are reviewed, evaluating advantages and disadvantages, and their extensions using *ensembles* are presented. Finally, the different techniques are applied to the ICESP dataset and compared with the Cox proportional hazards model using predictive performance metrics and cross validation techniques.

Keywords: Survival Analysis. Decision Tree. Survival Tree. Ensembles. Prediction.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Motivação

O uso de métodos de Análise de Sobrevivência para a predição do tempo até a ocorrência de um determinado evento tem crescido consideravelmente nos últimos anos. Este crescimento pode ser atribuído à combinação do avanço da técnica em si e do desenvolvimento dos recursos computacionais. Por exemplo, Löschmann e Smorodina (2020) estudam os conceitos de análise de sobrevivência em assuntos relacionados a aprendizado profundo, usando redes neurais e suas diferentes arquiteturas em diversos conjuntos de dados. Bollepalli et al (2023) testam diferentes algoritmos de aprendizado de máquina para predizer o tempo até a morte de indivíduos com problemas cardiácos.

Uma característica dos métodos de análise de dados de sobrevivência é a possibilidade de se considerar a informação contida em observações incompletas, denominadas censuras. Para essas observações, o tempo de ocorrência não é observado, sabendo-se apenas que ele é superior (ou inferior) ao instante efetivamente observado. Censuras podem ocorrer por diversas razões, como perda de seguimento, retirada voluntária dos participantes do estudo ou quando o evento de interesse não acontece até o final do estudo. Quando não há presença da censuras, poderíamos utilizar diferentes técnicas de classificação ou regressão para solucionar problemas de predição. Entretanto, se essas observações incompletas estão presentes, técnicas que não as consideram se tornam ineficientes, pois teríamos que excluir por completo as informações censuradas.

Talvez a abordagem mais conhecida para lidar com dados de sobrevivência seja aquela envolvendo o modelo de riscos proporcionais de Cox (Cox, 1972) que é um modelo de regressão bastante flexível. Isto deve-se ao fato de que a única suposição para o seu uso é que as taxas de risco sejam proporcionais, facilitando as interpretações. Apesar de ser constantemente utilizado nas áreas médicas, biológicas e industriais, outros segmentos também fazem uso desse modelo. Por exemplo, Wong (2011) aplica o modelo de Cox para predizer o risco de um cliente abandonar um produto de telecomunicações em uma empresa canadense através de informações demográficas e comportamentais. Em outro exemplo, Yala (2016) apresenta a utilização do modelo de Cox para estimar o tempo para um cliente deixar de pagar um empréstimo e discute como a robustez deste modelo pode

apoiar estratégias mais eficientes.

Apesar do modelo de Cox ser muito utilizado, usualmente esta técnica assume que o logarítmo da função da taxa de risco é composta por uma equação linear das covariáveis. Entretanto, pode-se descobrir que existe uma relação não linear e tornando-se necessário o uso de outras técnicas como por exemplo splines, para lidar com o problema. Essas decisões podem tornar o modelo mais complexo devido ao aumento do número de parâmetros a serem estimados. Com isso, quando o nosso foco está em predizer o tempo até a ocorrência de algum evento, torna-se atraente o uso de técnicas de aprendizado de máquina.

O aprendizado de máquina é um subcampo da inteligência artificial que tem como objetivo principal identificar padrões que minimizam alguma medida de incerteza. A sua vantagem é que existem diversas técnicas que podem *aprender* relações não-lineares entre as covariáveis e a variável resposta.

Várias técnicas podem ser utilizadas, dentre as quais temos os os métodos baseados em árvores de decisão que podem ser usados tanto para predições do tipo classificação quanto de regressão. A ideia é transformar as covariáveis em regras do tipo *if–else* de tal maneira que uma medida de risco seja otimizada fornecendo os nós finais que são responsáveis por realizar as predições.

Em um contexto mais geral, as árvores de decisão foram introduzidas por Morgan e Sonquist (1963), que apresentaram um primeiro algoritmo para realizar predição em problemas de regressão. Posteriormente, Breiman et at (1984) fez uma grande contribuição em sua monografia com o desenvolvimento do CART (*Classification and Regression Tree*) e expandindo, assim, o uso de árvores também para problemas de classificação.

Já em Análise de Sobrevivência, a contribuição inicia-se com os trabalhos de Gordon e Olshen (1985), no qual sugerem como critério de divisão (i.é, regras do tipo *if-else* baseadas nas covariáveis) a distância de Wassertein e citam como outros possíveis critérios a estatística log-rank e uma estatística baseada na verossimilhança. Segal (1988) apresenta o desenvolvimento de árvores utilizando estatísticas de comparação e cita as diferentes extensões da estatística *log-rank* como uma medida para divisão dos dados de sobrevivência.

Os métodos usando árvores de decisão podem ser uma importante ferramenta em problemas de predição devido à possibilidade de generalizar diferentes relações e de fornecer melhores métricas de avaliação, como o índice de C (Harrell et al, 1982). Além disso, destacam-se pela sua flexibilidade, uma vez que podem ser utilizados sob uma abordagem não-paramétrica na qual aos dados não se atribui uma distribuição de probabilidade.

A desvantagem do método é a necessidade de se categorizar as covariáveis numéricas, uma vez que elas são transformadas em regras *if–else*, e, portanto, pequenas mudanças podem causar alta variabilidade na predição. Uma forma de resolver esse problema é usando os métodos de *ensemble*, que baseiam-se em algoritmos de aprendizado de máquina que combinam os resultados de diversos modelos na busca por um "modelo melhor". Em um contexto geral, Breiman (1996, 2001) propôs uma abordagem baseada em *Bagging* e *Random Forest* enquanto que Friedman (1999) propôs o *Gradient Boosting Machine*, sendo os métodos de mais famosos e utilizados em problemas de classificação e regressão. No campo da Análise de Sobrevivência, Ishwaran et al (2008) introduziu o método de *Random*

Forest e o nomeou *Random Survival Forest*; Ridgeway (1999) apresentou soluções genéricas para que assim possam ser consideradas diferentes verossimilhanças nos algoritmos de *boosting* e, em particular, a verossimilhança parcial usada no modelo de Cox.

Este trabalho tem como motivação estudar diferentes algoritmos que usam árvores em dados censurados, comparando-os com o modelo de Cox, apresentando as suas teorias, entendendo suas vantagens e desvantagens e aplicando-os a um conjunto de dados fornecido pelo Instituo do Câncer do Estado de São Paulo (ICESP). Esses dados foram coletados durante o período de março de 2010 a agosto de 2011. Os pesquisadores realizaram um estudo de coorte prospectivo nas UTIs de dois hospitais públicos brasileiros especializados no tratamento do câncer, o ICESP e o Hospital do Câncer de Barretos. Após aplicar critérios de inclusão e exclusão bem estabelecidos, foram acompanhados 793 pacientes com difentes tipos de câncer.

O câncer é um termo que abrange diversas doenças que têm em comum o crescimento desordenado de células, gerando possíveis doenças malignas. Essas células se dividem rapidamente e se agrupam formando os tumores. Os tumores invadem os tecidos e podem invadir órgãos vizinhos, podendo estar distantes da origem do tumor, o que origina as metástases.

Esse conjunto de dados apresenta uma variedade de variáveis, mas neste trabalho serão utilizadas as variáveis descritas no Apêncice A.

1.2 Objetivo

Os objetivos desta dissertação são:

- Revisar os métodos de árvores de decisão em análise de sobrevivência, apresentando suas propriedades, vantagens e desvantagens.
- Estudar os métodos de árvores de sobrevivência, avaliando tais métodos em um conjunto de dados real.
- Desenvolver o uso do software Python para gerar as análises necessárias, utilizando a biblioteca *scikit-survival* desenvolvida por Pölsterl (2020).

1.3 Organização do trabalho

A organização do trabalho é feita da seguinte forma. O Capítulo 2 revisa os conceitos básicos da Análise de Sobrevivência, nos quais são apresentadas as principais estatísticas e o modelo de riscos proporcionais de Cox. O Capítulo 3 apresenta os conceitos básicos de aprendizado de máquina, no qual são definidos viés e variância, validação cruzada e como se avalia modelos de análise de sobrevivência usando métricas de performance preditiva. O Capítulo 4 estuda as árvores de decisão em regressão, discutindo como esse método se ajusta aos dados e discutindo suas extensões utilizando os métodos de agrupamento de modelos (*ensemble*): Floresta Aleatória (*Random Forest*) e *Gradient Boosting Machine*. O Capítulo 5 apresenta as árvores de decisão em análise de sobrevivência, discutindo a teoria e o ajuste dessas técnicas aos dados. Além disso, estuda-se como os métodos de

ensemble funcionam nesses casos. O Capítulo 6 analisa os ajustes dos métodos de árvores em um conjunto de dados real comparando-os com o modelo de Cox. Por fim, o Capítulo 7 encerra o trabalho discutindo algumas conclusões e próximos passos.

Capítulo 2

Uma Breve Revisão de Análise de Sobrevivência

A análise de sobrevivência é um ramo da estatística que consiste em analisar de um ou mais eventos de interesse, genericamente denominado o tempo decorrido até ocorrência do evento que está definido como **tempo de falha**. Amplamente utilizada em estudos médicos, essa técnica permite o estudo de tempos de sobrevida, em que muitas vezes são encontradas informações de forma incompletas nos dados, denominadas como **censuras** (Colosimo e Giolo, 2006).

Nesse contexto, a presença de censuras justifica o uso das técnicas de análise de sobrevivência, pois não seria possível aplicar métodos estatísticos tradicionais para modelar esse tipo de evento.

Em estudos de sobrevivência, é importante definir variável aleatória não negativa contínua T como o tempo até a falha. A distribuição dessa variável pode ser caracterizada pelas seguintes funções:

- Função de sobrevivência $\mathbf{S(t)}$.
- Função de taxa de risco $\boldsymbol{\lambda(t)}$.
- Função de taxa de risco acumulado $\boldsymbol{\Lambda(t)}$.

Uma maneira de estudar essas funções é utilizando estimadores não-paramétricos. A vantagem de usá-los é que não precisamos assumir uma distribuição de probabilidade sob os dados. Dentre os estimadores, os mais famosos são o estimador **Kaplan-Meier** para função $S(t)$ e o estimador **Nelson-Aalen** para função $\lambda(t)$.

Além disso, pode-se querer estimá-las levando-se em consideração o uso de covariáveis. O modelo de riscos proporcionais de Cox é uma abordagem muito popular, não assume nenhuma distribuição de probabilidade para os dados; sua única suposição é que as taxas de risco sejam proporcionais.

Nesse capítulo, revisaremos alguns conceitos básicos, porém úteis e importantes para o restante desta dissertação.

2.1 Função de Sobrevivência

Seja T uma variável aleatória contínua não negativa que representa o tempo até a ocorrência de um evento, a qual possui uma distribuição de probabilidade com função densidade de probabilidade $f(t)$ e função de distribuição acumulada $F(t)$.

A função de sobrevivência $S(t)$ é definida como a probabilidade de ocorrência do evento determinado após o tempo t:

$$S(t) = P(T > t), \quad t > 0. \tag{2.1}$$

Em outras palavras, é a probabilidade de uma ocorrência não acontecer até um certo instante t.

Essa função contém algumas propriedades:

- É a probabilidade complementar da $F(t)$, isto é, $S(t) = 1 - F(t)$.
- É uma função monótona descrescente e contínua.
- À medida que t se aproxima de 0, a $S(t)$ se aproxima de 1.
- À medida que t cresce, a $S(t)$ se aproxima de 0.

2.2 Função de Taxa de Risco

A função de taxa de risco é de grande importância no estudo de sobrevivência, pois fornece a intensidade com que os eventos ocorrem ao longo do tempo. Esta função fornece a taxa de falha condicional, ou seja, nos informa a taxa de falha em um intervalo de tempo, dado que o evento não foi observado até o início desse intervalo.

Considere um intervalo de tempo $[t, t + \Delta t)$. A função taxa de risco está associada com probabilidade de que as observações falhem neste intervalo, dado que sobreviveram antes de um tempo determinado t, dividido pelo comprimento do intervalo Δt, quando $\Delta t \to 0$. Matematicamente, a função de taxa de risco expressa-se por:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2}$$

Segue de imediato que:

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{2.3}$$

e que:

$$\lambda(t) = -\frac{d}{dt}\log(S(t)). \tag{2.4}$$

Isto é, a função de taxa de risco $\lambda(t)$ é a razão da função de densidade de probabilidade $f(t)$ e da função de sobrevivência $S(t)$, como também a derivada da função $-\log(S(t))$. Essa função também pode ser implementada como a taxa de falha instantânea em função do tempo t dada a sobrevivência até o tempo t.

2.3 Função de Taxa de Risco Acumalada

A função de taxa de risco acumulada fornece a taxa de falha acumulada até um certo tempo t e é definida como:

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

Além disso, a $\Lambda(t)$ pode ser escrita como:

$$\Lambda(t) = -\log(S(t)). \tag{2.6}$$

Essa função não tem uma interpretação direta, mas pode ser bastante útil para avaliar a função de taxa de risco $\lambda(t)$. Isso ocorre na estimação não-paramétrica, em que $\Lambda(t)$ proporciona um estimador com melhores propriedades, comparativamente a $\lambda(t)$.

2.4 Estimadores não-paramétricos

Uma abordagem para estudar dados censurados é utilizar os estimadores não-paramétricos. A seguir, são discutidos os estimadores e testes utilizados neste trabalho.

2.4.1 Kaplan-Meier

O estimador de *Kaplan-Meier*, introduzido por Kaplan e Meyer (1958) e também conhecido como estimador produto-limite, é uma adaptação da função de sobrevivência empírica dada pela expressão abaixo.

$$\hat{S}(t) = \frac{\text{número de unidades com o tempo de acompanhamento} > t}{n}, \quad t \geq 0, \tag{2.7}$$

sendo que n é o número de observações na amostra.

Quando há a presença de censura, a expressão (2.7) precisa ser generalizada para acomodar as observações incompletas.

Sendo assim, para que se possa aplicar uma generalização em $\hat{S}(t)$, assume-se que:

- r, $r \leq n$, corresponde ao número de tempos da ocorrência do evento de interesse.
- $n - r$ representa o número de tempos censurados.
- $t_{(i)}$ são os tempos ordenados da ocorrência do evento para o indíviduo i, $i = 1, ..., r$.
- n_i é o número de indivíduos em "risco", isto é, que não apresentam o evento de interesse ou foram censurados até o interesse imediatamente anterior a $t_{(i)}$.
- d_i é o número de eventos observados no instante $t_{(i)}$.

Com isso, o estimador de Kaplan-Meier pode ser definido como:

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.8}$$

no qual $\hat{S}(t) = 1$ para $0 \leq t \leq t_{(1)}$.

Tem-se algumas observações sobre esse estimador:

- Quando não existem censuras, a função de sobrevivência empírica e o estimador de Kaplan-Meier coincidem.
- Se o maior tempo observado t^* for não censurado, então $\hat{S}(t) = 0$ para $t \geq t_{(r)}$. Porém, se ele corresponde a uma censura, então $\hat{S}(t) > 0$ para todo $t > t^*$.
- É um estimador consistente e, sob certas condições, pode ser considerado como um estimador de máxima verossimilhança de $S(t)$ (Kaplan e Meier, 1958).

2.4.2 Nelson-Aalen

Um estimador para $\Lambda(t)$ foi inicialmente desenvolvido por Nelson (1972) e retomado por Aalen (1978) que provou suas propriedades assintóticas. Este estimador pode ser expresso da seguinte forma:

$$\hat{\Lambda}(t) = \sum_{j:t_j < t} \left(\frac{d_j}{n_j}\right), \tag{2.9}$$

em que d_j e n_j são definidos como no estimador de Kaplan-Meier na Seção 2.4.1.

Através da expressão em (2.9), podemos encontrar um estimador para função de sobrevivência. Usando a equação (2.6), tem-se que

$$S(t) = \exp(-\Lambda(t)), \tag{2.10}$$

que motiva a definição do estimador

$$\hat{S}(t) = \exp(-\hat{\Lambda}(t)) = \prod_{j:t_j<t} \exp\left(-\frac{d_j}{n_j}\right). \tag{2.11}$$

O estimador $\hat{S}(t)$ é chamado de estimador de Breslow da função de sobrevivência. Pode-se mostrar que o estimador de Kaplan-Meier é uma aproximação de 1ª ordem de (2.11), sendo que as estimativas de Breslow sempre serão maiores ou iguais as estimativas do Kaplan-Meier.

O gráfico das estimativas de Nelson-Aalen podem ser úteis para verificar a adequação de modelos paramétricos e para gerar predições em Análise de Sobrevivência.

2.5 Comparação estatística de duas populações

O teste *log-rank*, proposto por Mantel (1966) e renomeado por Peto e Peto (1972), é utilizado para comparar distribuições em análise de sobrevivência. A estatística de teste avalia se as estimativas das funções da taxa de risco dos grupos em cada tempo de evento observado são iguais. Esse teste é construído calculando o número de eventos observados e esperados - sob a hipotése de igualdade das funções - em um dos grupos em cada tempo de evento e, em seguida, adicionando-os para obter um resumo geral em todos os pontos de tempo em que há o evento.

Suponha dois grupos de pacientes, grupos 1 e 2. Sejam $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ os tempos distintos de eventos observados para os grupos combinados. Pode-se definir as seguintes quantidades:

- $N_{i,j}$: a quantidade de pacientes em risco do grupo i no instante imediatamente anterior a $t_{(j)}$.
- $O_{i,j}$: a quantidade de pacientes do grupo i que apresentaram o evento no instante $t_{(j)}$ para o grupo i.

com $i = 1, 2$ e $j = 1, ..., k$.

A hipótese nula pode ser interpretada como se os dois grupos têm a mesma função de taxa de risco e a hipótese alternativa considerando que essas funções são diferentes:

$$\begin{aligned} H_0 &: \lambda_1(t) = \lambda_2(t), \text{ para todo } t \geq 0; \\ H_1 &: \lambda_1(t) \neq \lambda_2(t), \text{ algum } t \geq 0. \end{aligned} \tag{2.12}$$

Com isso, sob H_0, para cada grupo i, condicionando-se em N_j, $N_{(i,j)}$ e O_j, a variável aleatória $O_{i,j}$ segue uma distribuição hipergeométrica com parâmetros N_j, $N_{i,j}$ e O_j, no qual $N_j = \sum_i N_{i,j}$. Daí, tem-se:

$$E[O_{i,j}] = O_j \frac{N_{i,j}}{N_j} \text{ e } V[O_{i,j}] = E[O_{i,j}] \left(\frac{N_j - O_j}{N_j} \right) \left(\frac{N_j - N_{i,j}}{N_j - 1} \right). \tag{2.13}$$

Pelo teorema central do limite, pode-se encontrar a distribuição assintótica de Z_i dada por:

$$Z_i = \frac{\sum_{j=1}^{k} O_{i,j} - E[O_{i,j}]}{\sqrt{\sum_{j=1}^{k} V[O_{i,j}]}} \xrightarrow{D} N(0,1). \tag{2.14}$$

Para a realização do teste, pode ser considerado apenas um dos grupos. Assim, elevando Z_2 ao quadrado, encontra-se a estatística do teste *log-rank* T,

$$Q = \frac{(\sum_{j=1}^{k} O_{2,j} - E[O_{2,j}])^2}{\sum_{j=1}^{k} V[O_{2,j}]} \xrightarrow{D} \chi_1^2. \tag{2.15}$$

Tarone e Ware (1977) propuseram uma classe de estatísticas para comparação de populações em análise de sobrevivência. Trata-se de uma generalização da estatística Q dada na expressão (2.15). Isto é, a ideia é semelhante ao teste *log-rank*, mas com a adição de um peso w_j, com $j = 1, ..., k$, que pode gerar diferentes estatísticas de teste.

Com isso, pode-se calcular a estatística QW da seguinte forma:

$$QW = \frac{(\sum_{j=1}^{k} w_j [O_{2,j} - E[O_{2,j}]])^2}{\sum_{j=1}^{k} w_j^2 V[O_{2,j}]}. \tag{2.16}$$

Para definição de w_j, existem diferentes propostas:

1. $w_j = 1$ resulta na estatística *log-rank* (Peto e Peto, 1972) expressa em (2.15).
2. $w_j = N_j$ resulta na estatística de Gehan (1965).
3. $w_j = N_j^{\frac{1}{2}}$ resulta na estatística apresentada por Tarone e Ware (1977).
4. $w_j = S_j^*$ resulta na estatística proposta por Prentice (1978) em que $S_j = \prod_{j=1}^{i} N_j/(N_j + 1)$ é motivada pela estimativa de Kaplan Meier no i-ésimo tempo até o evento não censurado.

Valores elevados de Q e QW fornecem evidências de que a quantidade observada de indivíduos no grupo 2 que experimentam falhas em diferentes intervalos de tempo difere do número esperado, considerando a hipótese de que ambos os grupos 1 e 2 são equivalentes no que diz respeito à sobrevivência.

As estatísticas de teste Q e QW são de suma importância para a construção das árvores de decisão em análise de sobrevivência, uma vez que é utilizada para fazer a divisão dos dados.

2.6 Modelo de Riscos Proporcionais de Cox

2.6.1 Contexto

O modelo de regressão de Cox é baseado em uma relação simples entre a função de risco e um conjunto de variáveis independentes, facilitando a interpretação dos parâmetros associados a elas.

Suponha uma variável aleatória T que representa o tempo até a ocorrência de um evento. Considere um vetor de covariáveis $\mathbf{x}' = (x_1, x_2, ..., x_p)$ avaliadas na origem do acompanhamento. Sejam $\boldsymbol{\beta}' = (\beta_1, ..., \beta_p)$, os coeficientes de regressão desconhecidos que representam os efeitos das covariáveis na sobrevivência e $\lambda_0(t)$ uma função de taxa de risco arbitrária e que não envolve o vetor de covariáveis. Então, o modelo de riscos proporcionais de Cox é especificado através da relação:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_j\right), \quad t \geq 0. \tag{2.17}$$

Note que o efeito das covariáveis é considerado de forma paramétrica porém, tem origem não paramétrica de tal forma que (2.16) define um modelo de regressão semi-paramétrico.

A razão das funções de taxa de risco para duas observações com covariáveis $\mathbf{x_1}$ e $\mathbf{x_2}$, é dada pela seguinte forma:

$$\frac{\lambda(t|\mathbf{x_1})}{\lambda(t|\mathbf{x_2})} = \exp\left(\boldsymbol{\beta}'(\mathbf{x_1} - \mathbf{x_2})\right), \tag{2.18}$$

ou seja, não depende do tempo t.

Pelo fato da razão ser constante no tempo, este modelo é denominado como modelo de riscos proporcionais. Por exemplo, considere um estudo em que estamos querendo estimar a sobrevivência a uma determinada doença. Se um indivíduo num determinado tempo t_1 tem um risco de morte igual a três vezes o risco de um segundo indivíduo, então esta razão de riscos será a mesma para todos outros tempos t. Com isso, a única suposição para o seu uso é que as taxas de risco sejam proporcionais.

Para estimar os parâmetros $\boldsymbol{\beta}$, o método de máxima verossimilhança proposto por Cox e Hinkley (1974) é bastante conhecido e frequentemente utilizado. Contudo, a presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança, torna esse método complicado. Cox (1975) propôs uma solução que consiste em condicionar a verossimilhança para eliminar essa função arbitrária e a denominou como o método da verossimilhança parcial.

2.6.2 Estimação dos Parâmetros

Considere que em uma amostra de n observações existam $r \leq n$ falhas distintas nos tempos $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Tem-se que a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i dada a probabilidade de estar em risco em t_i, pode ser escrita de forma heurística como

$$\frac{\lambda_i(t)}{\sum_{j \in R(t_i)} \lambda_j(t)} = \frac{\lambda_0(t)\exp\left(\mathbf{x_i'}\boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \lambda_0(t)\exp\left(\mathbf{x_j'}\boldsymbol{\beta}\right)} = \frac{\exp\left(\mathbf{x_i'}\boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathbf{x_j'}\boldsymbol{\beta}\right)}, \tag{2.19}$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i. Observe que em (2.19), o componente $\lambda_0(t)$ desaparece.

A função de verossimilhança parcial pode ser encontrada tomando-se o produto dos termos em (2.19) associados aos tempos distintos de falha, isto é,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\mathbf{x_i'}\boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathbf{x_j'}\boldsymbol{\beta}\right)} = \prod_{i=1}^{n} \left(\frac{\exp\left(\mathbf{x_i'}\boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathbf{x_j'}\boldsymbol{\beta}\right)} \right)^{\delta_i}, \tag{2.20}$$

em que δ_i é o indicador de falha. Os valores de $\boldsymbol{\beta}$ que maximizam (2.19) são encontrados resolvendo-se o sistema de equações definido pelo vetor escore $U(\boldsymbol{\beta})$, no qual $U(\boldsymbol{\beta})$ é obtido através das primeiras derivadas da log função de verossimilhança parcial $\log(L(\boldsymbol{\beta}))$,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \in R(t_i)} x_j \exp\left(\mathbf{x_j'}\hat{\boldsymbol{\beta}}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathbf{x_j'}\hat{\boldsymbol{\beta}}\right)} \right] = 0. \tag{2.21}$$

A solução dessa equação pode ser obtida iterativamente utilizando-se o método de Newton-Raphson.

2.6.3 Predição

A predição em análise de sobrevivência difere das técnicas de modelagem em regressão e classificação. Ao invés de estimar a variável T, são estimadas funções atreladas a T: a

função de sobrevivência $S(t)$, a função de taxa de risco $\lambda(t)$ e a função de taxa de risco acumulada $\Lambda(t)$ (apresentadas nas Seções 2.1, 2.2 e 2.3).

No modelo de Cox, pode-se encontrar estimativas das taxas de risco $\lambda(t|x_i)$ da seguinte forma:

$$\hat{\lambda}(t|\mathbf{x_i}) = \hat{\lambda}_0(t)\exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right), \tag{2.22}$$

em que $\hat{\boldsymbol{\beta}}$ é o estimador de verossimilhança parcial. Com relação ao estimador $\hat{\lambda}_0(t)$, Breslow (1972) propôs uma estimativa simples para a função de taxa de risco acumulada $\Lambda_0(t)$ que é expressa por

$$\hat{\Lambda}_0(t_i) = \sum_{j:t_j \leq t} \frac{d_j}{\sum_{i:t_i \in R(t_j)} \exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right)}, \tag{2.23}$$

em que d_j é o número de falhas em t_j e $R(t_j)$ é o conjunto de índices das observações sob risco no tempo t_j. Esse estimador é uma função do tipo escada e o $\hat{\lambda}_0(t)$ representa os saltos da função que ocorrem nos instantes da falha.

Sob a suposição de riscos proporcionais, e a relação entre a função de sobrevivência e a função de taxa de risco acumulada, é possível mostrar que

$$S(t|\mathbf{x}) = [S_0(t)]^{\exp(\mathbf{x_i'}\boldsymbol{\beta})}. \tag{2.24}$$

Em consequência, a estimativa da função de sobrevivência $S_0(t)$ pode ser encontrada através de (2.18) por,

$$\hat{S}_0(t) = \exp\left(-\hat{\Lambda}_0(t)\right). \tag{2.25}$$

Consequentemente, é possível definir o estimador,

$$\hat{\Lambda}(t|\mathbf{x_i}) = -\log(\hat{S}(t|\mathbf{x_i})). \tag{2.26}$$

As quantidades $\hat{\lambda}(t|\mathbf{x_i})$, $\hat{S}(t|\mathbf{x_i})$ e $\hat{\Lambda}(t|\mathbf{x_i})$ são fundamentais quando o objetivo é predizer. Também com essas, pode-se comparar o modelo de Cox com outros modelos preditivos usando métricas de avaliação de performance e identificar quais são as observações mais propensas a sobreviver.

Capítulo 3

Conceitos de Aprendizado de Máquina

O aprendizado de máquina é uma aplicação de inteligência artificial que inclui diversos algoritmos que analisam dados, aprendem com esses e, em seguida, pode ser utilizado para auxiliar a tomar decisões. Um exemplo de algoritmo de aprendizado de máquina é um sistema de recomendação em um serviço de *streaming* de vídeo, como o *YouTube*.

Para que o algoritmo do *YouTube* indique uma decisão sobre quais novos vídeos recomendar para um usuário, algoritmos de aprendizado de máquina associam suas preferências a outros usuários que tenham um gosto semelhante. Essa técnica, que geralmente é simplesmente apresentada como inteligência artificial, é usada em muitos serviços que oferecem recomendações automatizadas.

Existem dois tipos de aprendizados de máquina: o não-supervisionado e o supervisionado. No caso do aprendizado não supervisionado, os algoritmos são aplicados para analisar e agrupar conjuntos de dados sem considerar uma variável resposta. Tais algoritmos podem ser divididos em três categorias diferentes: clusterização, associação e redução de dimensionalidade. Já o aprendizado supervisionado tem como objetivo encontrar uma função $f(\mathbf{x})$, na qual $\mathbf{x}$ é um conjunto de covariáveis, que minimiza a incerteza associada a uma variável resposta Y. Isto é, a ideia é identificar uma função $f(\mathbf{x})$ que consiga predizer a variável resposta Y.

Neste capítulo, o foco é mostrar os conceitos básicos em aprendizado supervisionado, que é o utilizado em problemas oriundos da análise de sobrevivência. São apresentados os conceitos de viés e variância, as técnicas de validação cruzada e métricas de avaliação de performance.

3.1 Viés e Variância

Considere uma variável aleatória Y representando a variável resposta. Também considere um vetor de covariáveis fixas $\mathbf{x}' = (x_1, ..., x_p)$. Assume-se que existe uma função $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$Y = f(\mathbf{x}) + \epsilon, \tag{3.1}$$

isto é, o comportamento da variável Y é explicado em parte por uma componente sistemática que depende de x por meio da função $f(\cdot)$ e, aquilo que não é captado por essa parte está contida na componente aleatória ϵ, que é o termo de erro, independente de $\mathbf{x}$ e com média zero e variância σ^2.

Em geral a função $f(\mathbf{x})$ depende de um ou mais parâmetros desconhecidos, que precisam ser estimados. Uma vez obtidas essas estimativas, pode-se contruir um procedimento de predição da variável Y subtituindo-se os parâmetros desconhecidos na função original. Denote essa função de predição por $\hat{f}(\mathbf{x})$. Pode-se então considerar o erro quadrádico médio dado por

$$E[(Y - \hat{f}(\mathbf{x}))^2]. \tag{3.2}$$

Realizando os devidos cálculos, tem-se que

$$\begin{aligned} E[(Y - \hat{f}(\mathbf{x}))^2] &= E[(Y - f(\mathbf{x}) + f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2] \\ &= E[(Y - f(\mathbf{x}))^2] + 2E[Y - f(\mathbf{x})]E[\hat{f}(\mathbf{x}) - f(\mathbf{x})] + E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2]. \end{aligned} \tag{3.3}$$

Como $E(Y) = f(\mathbf{x})$ e $Var(Y) = Var(\epsilon) = \sigma^2$,

$$\begin{aligned} E[(Y - \hat{f}(\mathbf{x}))^2] &= E[(Y - f(\mathbf{x}))^2] + E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] \\ &= Var[\epsilon] + E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] \\ &= \sigma^2 + E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2]. \end{aligned} \tag{3.4}$$

Considerando-se o segundo termo na última expressão em (3.4), tem-se que

$$\begin{aligned} E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] &= E[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})] + E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2] \\ &= E[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2] - 2E[\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})]]E[E[\hat{f}(\mathbf{x})] - f(\mathbf{x})] + \\ &+ E[(E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2]. \end{aligned} \tag{3.5}$$

Como $E[\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})]] = 0$ e $E[(E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2] = (E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2$,

$$E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] = E[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2] + (E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2. \tag{3.6}$$

Logo, a expressão (3.2) pode ser escrita por

$$E[(Y - \hat{f}(\mathbf{x}))^2] = \sigma^2 + E[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2] + (E[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2, \tag{3.7}$$

em que:

- $E[\hat{f}(\mathbf{x})] - f(\mathbf{x})$ é o **viés** decorrente de se predizer Y por $\hat{f}(\mathbf{x})$, que pode ser interpretada como a diferença entre o valor real e o predito.
- $E[(\hat{f}(\mathbf{x}) - E[\hat{f}(\mathbf{x})])^2]$ é a **variância**, representando o quão homogêneas são as predições em relação ao valor esperado de $\hat{f}(\mathbf{x})$.
- σ^2 é o **erro irredutível**, isto é, um erro que não pode ser reduzido por $f(\mathbf{x})$. Geralmente, esse termo contém o ruído no conjunto de dados que não é captado pela componente sistemática do modelo.

Os termos viés e variância são importantes para o desenvolvimento e escolha de modelos preditivos. À medida que é construído um modelo mais complexo - por exemplo, um modelo com diversos parâmetros e/ou alta profundidade numa árvore de decisão - tende-se a diminuir o viés e aumentar a variância. Na figura a seguir, pode-se ver isso. Tal comportamento é denominado *trade-off*.

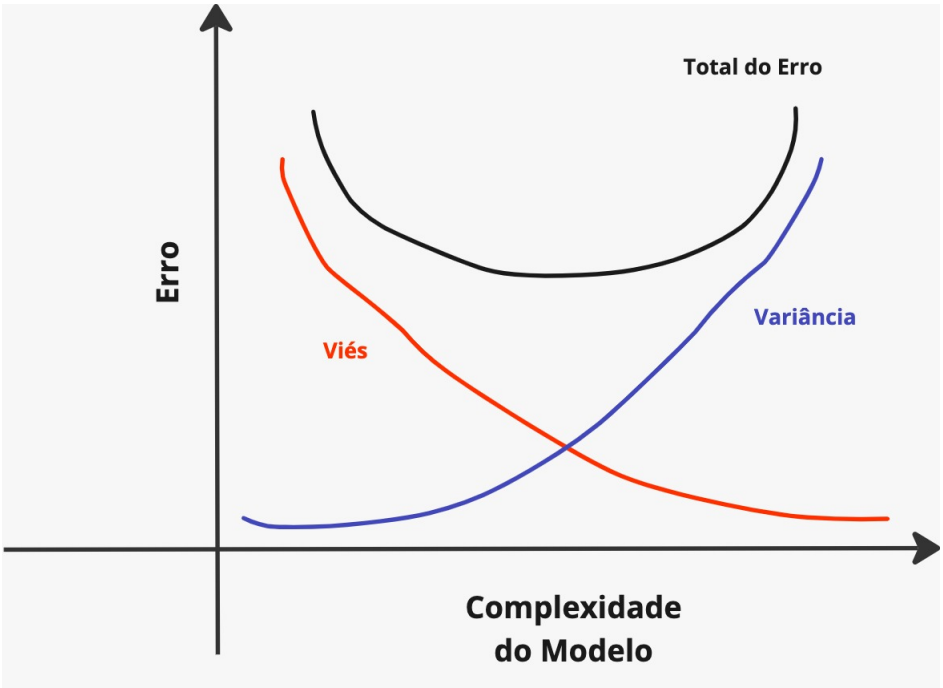

Figura 3.1: *Trade-off entre viés e variância.*

Com o conceito de *trade-off*, tem-se dois problemas que podem ocorrer ao se construir um modelo:

- **Sobreajuste** (*Overfitting*), que ocorre quando um modelo tem baixo viés mas alta variância.
- **Subajuste** (*Underfitting*), que se apresenta quando um modelo tem alto viés mas baixa variância.

Para que se possa estudar o impacto associado a esses conceitos, é necessário realizar uma partição no conjunto de dados que leva à **validação cruzada** possiblitando medir através de **métricas de avaliação de performance**, o quanto o modelo está sendo preciso no processo de predição.

3.2 Validação Cruzada

Em aprendizado de máquina, o termo generalização se refere à capacidade de um algoritmo aprender determinados padrões a partir de um conjunto de dados. O objetivo é medir a capacidade do modelo em realizar predições em novos conjuntos de dados, sem que haja perda de performance preditiva.

Para avaliar a capacidade de generalização de um modelo de aprendizado de máquina e encontrar um equilíbrio entre sobreajuste e subajuste, a validação cruzada é uma ferramenta fundamental. Trata-se de uma técnica para avaliar um modelo de aprendizado de máquina e testar seu desempenho em diferentes conjuntos de dados. Essa técnica ajuda a comparar e selecionar um modelo para o problema específico de modelagem preditivo, pois permite quantificar o viés e variância.

Existem várias técnicas diferentes que podem ser usadas para realizar validação. Entretanto, todas elas têm um algoritmo semelhante:

1. Dividir o conjunto de dados em duas partes: uma para **treinamento** e outra para **teste**.
2. Ajustar o modelo no conjunto de treinamento. Normalmente, esse procedimento é chamado de *treinamento do modelo*.
3. Calcular medidas de avaliação de performance no conjunto de teste denominada *etapa de validação*.
4. Repetir as etapas 1 e 3 algumas vezes. O número de repetições depende do método que está sendo utilizado.

Existem muitas técnicas de validação cruzada. Algumas delas são mais comumente utilizadas. Neste trabalho, serão abordados os métodos: *Hold-out* e *k-Fold*.

3.2.1 Hold-out

Hold-out é a técnica mais simples e popular. Consiste-em:

1. Dividir aleatoriamente o conjunto de dados em duas partes: o conjunto de treinamento e o conjunto de teste. Normalmente, 70% do conjunto de dados vai para o conjunto de treinamento e 30% para o conjunto de teste.
2. Ajustar (i.é, treina-se) o modelo no conjunto de treinamento.
3. Calcular as métricas de avaliação de performance nos conjuntos de treino e de teste.

Figura 3.2: *Ilustração do algoritmo do Hold-out.*

Por mais que seja simples, o *hold-out* tem uma grande desvantagem. Se o conjunto de dados não corresponder a realizações de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), pode-se encontrar uma alta variância ao se avaliar o desempenho nos conjuntos de treino e no teste. Além disso, o fato de se testar o modelo apenas uma vez pode ser um ponto negativo devido ao tamanho do conjunto de dados (James et al, 2013).

3.2.2 Validação cruzada em k etapas (k-fold)

O método de validação cruzada em k etapas, também conhecido como *k-fold*, procura minimizar as desvantagens do método *hold-out* auxiliando a lidar com a questão de "testar"uma única vez.

O algoritmo se baseia nos seguintes passos:

1. Definir um número k de etapas (*folds*). Geralmente, k assume o valor de 5 ou 10, mas pode-se escolher qualquer número tal que $k < n$, sendo n o tamanho do conjunto de dados.
2. Em cada etapa i, $i = 1, \dots, k$, considerar uma partição D_i do conjunto de dados originais, composta por dois subconjuntos: T_i, um subconjunto de treinamento e V_i um subconjunto de validação.
3. Com o modelo ajustado em (3), calcular uma ou mais métricas de avaliação de performance considerando-se o subconjunto V_i.
4. Os passos (3) e (4) são repetidos para todas as k partições $D_1, \dots, D_k$.
5. Com o modelo ajustado em (3), calcular uma ou mais métricas de avaliação de performance considerando-se os subconjuntos T_i e V_i.

6. Calcular algumas medidas resumo (por exemplo, média e variância) para as métricas de avaliação obtidas nas etapas $1, 2, \ldots, k$.

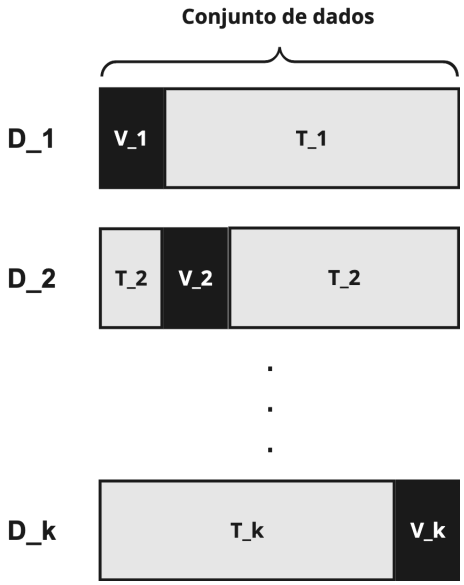

Figura 3.3: *Ilustração do algoritmo para validação cruzada em k etapas.*

A vantagem é que o *k-fold* fornece um resultado mais estável e confiável, pois o treinamento e o teste são realizados em diferentes partes do conjunto de dados. Podemos tornar a etapa de teste ainda mais robusta se aumentarmos o número k de etapas para testar o modelo em subconjuntos de dados diferentes.

O método *k-fold* tem uma desvantagem. Aumentar k resulta no treinamento de mais modelos e este processo pode ser caro e demorado computacionalmente.

3.3 Métricas de Avaliação de Performance

As métricas de avaliação são usadas para medir a qualidade de um modelo preditivo. Assim, avaliar modelos ou algoritmos de aprendizado de máquina é essencial para qualquer tipo de problema relacionado com predição.

Existem diferentes tipos de métricas de avaliação para testar um modelo. No caso de análise de sobrevivência, tem-se como métricas mais populares o **Índice C** e o **Escore de Brier**.

3.3.1 Índice C

O índice C, também conhecido como índice de concordância e introduzido em Harrell et al. (1982), é uma métrica de performance utilizada para avaliar modelos preditivos em análise de sobrevivência.

A intuição por trás do índice C de Harrell é a seguinte. Para o paciente i, um modelo de sobrevivência atribui uma estimativa de risco em seu tempo de falha T_i dada por η_i.

Se o modelo for adequado, os pacientes que tiveram tempos mais curtos até o evento devem ter estimativas $\hat{\eta}_i$ mais altas. Considerando apenas dois pacientes, se $\hat{\eta}_i > \hat{\eta}_j$, então $T_i < T_j$.

Para calcular a métrica, considere que para cada par de pacientes i e j (com $i \neq j$), obtém-se as estimativas $\hat{\eta}_i$ e $\hat{\eta}_j$ bem como os correspondentes tempos de falha T_i e T_j. Assim, o índice C pode ser encontrado pelos seguintes passos:

- Se T_i e T_j não forem censurados, pode-se observar quando ambos os pacientes tiveram a ocorrência do evento. Dizemos que o par (i, j) é um par concordante se $\hat{\eta}_i > \hat{\eta}_j$ e $T_i < T_j$ e é um par discordante se $\hat{\eta}_i > \hat{\eta}_j$ e $T_i > T_j$.
- Se T_i e T_j forem censurados, não é possível identificar quando ocorreu o evento para nenhum dos pacientes e, portanto, esse par não é considerado no cálculo do índice.
- Se um dos tempos de falha T_i e T_j for censurado, observa-se apenas uma ocorrência de evento. Suponha que foi observado o evento para o indivíduo i e que o tempo T_j foi censurado. Com isso,
  - Se $T_j < T_i$, não é possível saber qual paciente apresentou o evento primeiro. Então, o par é desconsiderado no cálculo do índice.
  - Se $T_j > T_i$, então o indivíduo i apresentou o evento primeiro. Portanto, (i, j) é um par concordante se $\hat{\eta}_i > \hat{\eta}_j$, e é um par discordante se $\hat{\eta}_i < \hat{\eta}_j$.

Logo, o índice C é definido como,

$$C = \frac{\sum_{i \neq j} I(T_i > T_j) I(\hat{\eta}_i < \hat{\eta}_j) \delta_j}{\sum_{i \neq j} I(T_i > T_j) \delta_j}, \tag{3.8}$$

em que $I(\cdot)$ é a função indicadora e δ_j é o indicador de falha para a j-ésima observação. Note que a estimativa η pode ser obtida considerando-se a função de sobrevivência, a função de taxa de risco ou a função de taxa de risco acumulada.

Algumas interpretações para essa medida:

- Quando c encontra-se próximo a 0,5, conclui-se que as predições não são melhores do que o arremessar de uma moeda honesta para determinar qual observação é mais propensa a apresentar o evento.
- Quando c está próximo de 1, conclui-se que as predições possuem uma boa capacidade discriminativa, sendo capaz em classificar corretamente a maioria das observações em relação ao tempo de sobrevivência. Em outras palavras, o modelo possui uma habilidade em ordenar as observações de acordo com o tempo até o evento de interesse.
- Quando c encontra-se próximo a 0, conclui-se que as predições são piores do que uma predição aleatória.

3.3.2 Escore de Brier

A métrica escore de Brier, proposta por Brier (1950), mede a distância entre a função de sobrevivência predita e a proporção de falhas observadas em um determinado ponto t. É uma maneira de avaliar se as curvas de sobrevivência que o modelo produz estão próximas em relação ao que foi observado.

Dado um conjunto de dados de sobrevivência com tamanho n, definido por $(T_i, \delta_i, \mathbf{x_i})$, com $i = 1, ..., n$. A função de sobrevivência predita é dada por $\hat{S}(t, \mathbf{x_i})$ com $t > 0$. Então, supondo ausência de censura à direita, o escore de Brier pode ser definido como,

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} (I(T_i > t) - \hat{S}(t, \mathbf{x_i}))^2, \tag{3.9}$$

no qual $I(\cdot)$ é uma função indicadora.

No caso de presença da censuras à direita, é necessário ajustar essa medida usando o **método da ponderação pelo inverso da probabilidade de censura**. Este método calcula para cada observação no tempo t um peso que é representado pelo inverso da probabilidade de censura. Supondo que $G(t) = P[C > t]$ representa a função de sobrevivência da variável da censura C, utiliza-se o método de Kaplan-Meier aplicado às observações censuradas para obter estimativas $\hat{G}(\cdot)$ e define-se o escore de Brier por

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\hat{S}(t, \mathbf{x_i})^2 I(T_i \leq t, \delta_i = 1)}{\hat{G}(T_i)} + \frac{(1 - \hat{S}(t, \mathbf{x_i}))^2 I(T_i > t)}{\hat{G}(t)} \right]. \tag{3.10}$$

Um escore de Brier próximo de 0 indica uma correspondência mais precisa entre as estimativas das funções de sobrevivência em t e as proporções de falhas experimentadas em um ponto t, indicando um desempenho superior do modelo em termos de calibração e precisão. Em contrapartida, um maior escore de Brier sugere uma disparidade entre as predições e os eventos reais, indicando uma menor confiabilidade do modelo em fornecer estimativas.

A vantagem do uso do escore de Brier é a possibilidade de avaliar o quanto a função de sobrevivência $\hat{S}(t, \mathbf{x})$ dada pelo modelo está calibrada isto é, o quão $\hat{S}(t, \mathbf{x})$ difere as taxas reais de eventos observados no tempo t.

Capítulo 4

Árvores de Decisão em Regressão

As árvores de decisão representam a uma técnica de aprendizado de máquina supervisionado. Elas são bastante úteis em problemas de regressão, quando o objetivo é predizer uma variável contínua com base em diversas covariáveis. Essa técnica é baseada no conceito de *árvores binárias*. Uma árvore binária é definida como uma árvore na qual todos os nós contêm exatamente dois nós filhos.

Em uma árvore de decisão, as folhas (ou nós terminais) representam as predições, os nós são definidos com base nas covariáveis e os ramos são regras aplicadas a essas covariáveis. A ideia geral é utilizar um algoritmo que busque selecionar a melhor covariável para realizar a divisão dos dados em cada nó da árvore, com base em alguma *medida de impureza* ou *erro na predição*. Existem vários algoritmos para construir uma árvore de decisão, mas o mais utilizado é o *CART* (Breiman et al, 1984).

O algoritmo CART consegue lidar com dados numéricos e categóricos e executa o crescimento máximo da árvore por meio de um particionamento binário recursivo. Ou seja, quando é ajustada uma árvore, não há um critério de parada, podendo ocorrer que ao final haja apenas uma observação em cada folha. Para evitar que isso aconteça, têm-se as técnicas de poda (*pruning*).

Poda refere-se um método de regularização que tem como objetivo de remover alguns ramos e nós da árvore. Existem duas abordagens principais:

- Pré-poda: essa abordagem envolve definir um critério para impossibilitar o crescimento da árvore antes que ela atinja sua profundidade máxima ou que garanta um determinado número mínimo de elementos da amostra em uma folha.
- Pós-poda: nessa abordagem, a árvore é construída sem restrições. Em seguida, alguns nós da árvore são removidos usando uma *função de custo*.

Para ambas abordagens de poda, são necessários definir os hiperparâmetros. Os hiperparâmetros são parâmetros externos ao modelo de aprendizado de máquina que não são aprendidos durante o seu ajuste, mas que afetam a construção do modelo e, consequentemente, o desempenho preditivo. Eles são determinados antes do treinamento do modelo e

são essenciais para otimizar a performance preditiva.

As principais vantagens do uso de árvores de decisão são:

- Não é necessário assumir uma distribuição de probabilidade para a variável resposta.
- A técnica assume uma relação não linear entre as covariáveis e a variável resposta.
- Identifica algoritmicamente interações entre as covariáveis.
- É facilmente interpretável.
- Requer pouca preparação dos dados.
- É capaz de lidar com dados numéricos e categóricos.
- É capaz de lidar com problemas envolvendo múltiplas variáveis respostas.

Como desvantagens, tem-se que:

- As árvores de decisão tendem a não levar em consideração a variância das covariáveis, uma vez que elas são categorizadas. Com isso, uma simples mudança no valor da covariável, pode causar um grande impacto na predição, podendo levar a um modelo com uma baixo poder preditivo.
- As predições geradas são aproximações constantes por partes, logo, as árvores não são boas em realizar extrapolações.
- Os algoritmos de aprendizado de árvore de decisão são baseados em algoritmos heurísticos, sendo que as decisões localmente ótimas são tomadas em cada nó. Tais algoritmos não podem garantir o retorno de uma árvore de decisão globalmente ótima.

Existem algumas técnicas que conseguem lidar com essas desvantagens, chamadas de agrupamento (*ensemble*). Os métodos de agrupamento combinam diversos modelos em busca de uma melhor predição. Existem diversos algoritmos de agrupamento, porém os mais utilizados em árvores de decisão são a Floresta Aleatória (*Random Forest*) e o *Gradient Boosting Machine*.

Neste capítulo, é introduzida a técnica de árvores de decisão em regressão, explicando como são construídas e suas extensões Floresta Aleatória e *Gradient Boosting Machine*.

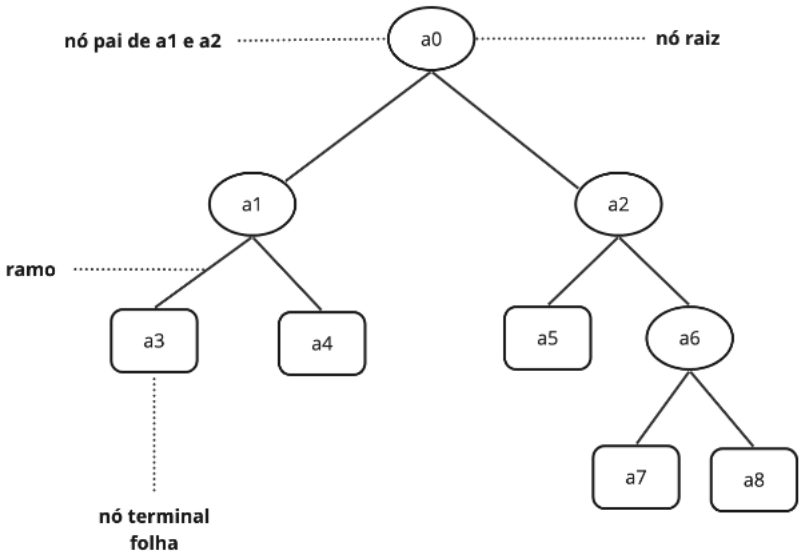

Figura 4.1: *Ilustração de uma árvore de decisão.*

4.1 Construindo árvores em regressão

Assume-se um vetor de covariáveis $\mathbf{x} = (x_1, ..., x_p)$, com $x_j \in \mathbb{R}$ com $j = 1, ..., p$ e um vetor de resposta $\mathbf{y} \in \mathbb{R}^n$. Segundo Segal (1988), para se construir uma árvore de decisão são necessários de quatro componentes:

1. Um conjunto de questões binárias da forma "$\mathbf{x} \in S$?"em que $S \subset \chi$, com χ sendo espaço das covariáveis e S um sub-espaço das covariáveis. A resposta a essa pergunta induz a uma partição do espaço das covariáveis. Ou seja, os casos para os quais a resposta é sim são associados à região S e aqueles que não são associados é definido pelo complemento de S. Essa indução realiza a divisão da amostra em subamostras chamadas de nós.

2. Uma qualidade de divisão satisfatória $\phi(s, t)$ que pode ser avaliada para qualquer divisão s de qualquer nó t. Essa qualidade é definida por meio de uma função de custo, como por exemplo a soma dos quadrados de resíduos, e é utilizada para avaliar as possíveis divisões. Com isso, pode-se encontrar a melhor divisão s para a criação dos nós da árvore.

3. Critérios de parada para determinar um tamanho apropriado da árvore.

4. Resumos estatísticos para as folhas da árvore treinada que podem ser simples como a média ou mediana ou mais complexos como uma curva de sobrevivência de Kaplan-Meier, dependendo do contexto.

4.1.1 Candidatos à divisão binária

A quantidade de divisões possíveis que podem ser encontradas em S, é reduzida a um número computacionalmente viável de tal forma que:

1. Cada divisão depende do valor de apenas uma covariável. Isto é, para um nó t, é esperado que se tenha apenas uma divisão definida por uma única covariável.
2. Para as covariáveis x_j, são consideradas apenas divisões resultantes de questões da forma "$x_j \leq c$?", com c sendo uma constante.

Pode aparentar que as reduções citadas acima são inúteis. Em (1), o número de divisões é restrito apenas à forma univariada, sendo que se poderia criar divisões multivariadas e em (2) restringe-se apenas em encontrar uma constante c, sendo que é possível identificar um intervalo $[c_1, c_2]$ tal que $c_1 \leq x_j \leq c_2$. No entanto, a ideia é diminuir o número incontável e infinito de partições que se pode definir.

Além disso, é fixado apenas um número finito de valores na amostra em questão e, assim, examina-se apenas os valores para c que resultam em um caso de troca de "lados". Ou seja, é necessário apenas avaliar o conjunto de valores distintos em x_j. Com isso, existem no máximo $n - 1$ divisões dadas por $x_j \leq c$ com c encontrado entre valores distintos observados para x_j.

Portanto, uma árvore é desenvolvida da seguinte maneira. Para cada nó:

- Encontra-se todas as divisões possíveis com base em cada covariável x_j.
- Seleciona-se e cria-se um nó direito e um nó esquerdo (i.é, nós filhos) com a divisão encontrada por $\phi(s, t)$.

As etapas 1 e 2 (citadas na Seção 4.1) são reaplicadas a cada um dos nós filhos e assim por diante até que seja atingido algum critério de parada.

4.1.2 Critério de qualidade da divisão

A maneira como é escolhida a divisão, baseia-se em uma métrica de qualidade, que depende de s na qual é a divisão realizada na covariável x_j, com $j = 1, ..., p$, bem como do nó t a ser dividido. Essa medida de qualidade é calculada usando uma **função de impureza** $i(t)$. A função de impureza $i(t)$ é uma métrica utilizada para medir a heterogeneidade dos dados em um nó. Quando construímos uma árvore de decisão, o objetivo é dividir os dados em subconjuntos homogêneos para tomar decisões mais precisas e fazer predições. A função de impureza determina quão bem uma determinada covariável x_j divide os dados.

Para problemas de regressão, existem diversas funções de impureza que podem ser aplicadas. A mais utilizada é a soma dos quadrados residual,

$$SS(t) = \sum_{k=1}^{m} (y_k - \bar{y}(t))^2, \tag{4.1}$$

em que

$$\bar{y}(t) = \frac{1}{m_t} \sum_{k=1}^{m_t} y_k, \tag{4.2}$$

sendo m_t o número de observações contidas no nó t.

A função $SS(t)$ fornece uma comparação baseada na decomposição subaditiva "entre/-dentro", e refere à homogeneidade ou medida de perda aplicada ao nó pai. Assim, se for considerado que $i(t) = SS(t)$, obtém-se o critério de divisão por mínimos quadrados.

Intuitivamente, quando dividimos as amostras queremos que a região correspondente a cada nó folha seja “pura”, ou seja, que a maioria das observações dessa região contenham a menor variância sob a variável resposta y. Com isso, a função de impureza $SS(t)$ tem como objetivo mensurar esse comportamento, isto é, calcular se um determinado nó está sendo o mais puro possível.

Logo, suponha-se t_d e t_e representando nós filhos direito e esquerdo definidos por uma divisão s no nó t. Então, o critério de divisão é dado por:

$$\phi(s,t) = i(t) - i(t_d) - i(t_e) = SS(t) - SS(t_d) - SS(t_e). \tag{4.3}$$

Com isso, pode-se encontrar a melhor divisão s^*, expressa por:

$$\phi(s^*,t) = \max_{s\in\Omega} \phi(s,t), \tag{4.4}$$

em que Ω é o espaço de todas possíveis divisões s.

Note que quando é escolhida a medida $SS(t)$ para o critério de divisão, a árvore de regressão é construída com nós cada vez mais homogêneos. A subaditividade é equivalente à não negatividade de ϕ, isto é, $SS(t) \geq SS(t_e) + SS(t_d)$ e, assim, essa desigualdade garante a homogeneidade crescente (estritamente não decrescente), garantindo que à medida que é aumentado o número de nós, tem-se nós mais homegêneos.

4.1.3 Critério de parada

Da mesma forma que é definido um critério para particionar um nó, precisamos de um critério para determinar quando fixar um nó, determiná-lo como uma folha e parar de particionar. Para isso, existem duas estratégias: pré-poda e pós-poda.

Pré-Poda

O objetivo do pré-pruning é evitar o sobreajuste, que ocorre quando a árvore de decisão se ajusta muito bem aos dados de treinamento, mas tem um desempenho ruim em dados

novos. Ao restringir o crescimento da árvore, a pré-poda ajuda a melhorar a generalização do modelo.

Sendo assim, existem alguns parâmetros que podem ser definidos antes do ajuste de uma árvore:

- A profundidade máxima, que restringe o número de vezes que o algoritmo divide os dados parando assim que atingir o valor especificado.
- O número de observações para divisão, que fixa um tamanho mínimo de observações em um nó para permitir a divisão.
- O número mínimo de observações nas restrições da folha, que é o número de observações necessárias para considerar um nó como uma folha, resultando em observações mínimas para realizar as predições.
- O número total de folhas, que é o número máximo de nós terminais que a árvore deve conter.
- A quantidade de impureza aceitável, que é a medida de impureza mínima para realizar a divisão dos dados.

Esses parâmetros são denominados como hiperparâmetros e são essenciais para o ajuste de uma árvore de decisão.

Em geral, a profundidade máxima, o número de amostras para divisão e o número de amostras mínimas nas folhas são os hiperparâmetros mais utilizados nessa estratégia, pois impossibilitam o crescimento da árvore.

Uma desvantagem do uso da pré-poda é que se esses hiperparâmetros forem escolhidos arbitrariamente, pode resultar em um modelo com a presença de subajuste. Para resolver essa questão, é de suma importância usar as técnicas de validação cruzada para que se possa medir esses *trade-offs* e testar diferentes conjuntos de hiperparâmetros.

Pós-poda

De forma análoga a pré-poda, o objetivo da pós-poda é evitar o sobreajuste, porém a estratégia é diferente. Ao invés de considerar diversos hiperparâmetros para definir um critério de parada, Breiman et al. (1984) propôs uma abordagem para a realização de uma poda (*pruning*) em uma árvore máxima até se chegar a uma árvore reduzida.

Com isso, Breiman et al (1984) apresentou a solução chamada custo de complexidade à poda (*cost complexity pruning*) denotado por $C_\alpha(A)$. Seja $i_j(A)$, a medida da impureza da j-ésima folha da árvore A cujo número de observações alocadas é n_j, com $j = 1, ..., J$. O custo de complexidade é definido como:

$$C_\alpha(A) = \sum_{j=1}^{J} n_j i_j(A) + \alpha J. \tag{4.5}$$

$C_\alpha(A)$ considera a medida da impureza e o tamanho da árvore. O parâmetro $\alpha \geq 0$ é denotado como hiperparâmetro e gerencia o *trade-off* entre o erro da predição e o tamanho da árvore. Grandes valores de α resultam em árvores pequenas, $\alpha = 0$ constitui a árvore com sua profundidade máxima.

Note que para encontrar o valor de α é necessário o uso das técnicas de validação cruzada.

4.2 Agrupamento de modelos (Ensemble)

O *ensemble* é uma técnica poderosa e versátil de aprendizado de máquina que busca combinar as predições de vários modelos individuais para melhorar a precisão e a robustez dos resultados.

A ideia subjacente dessa técnica é que diferentes modelos podem capturar diferentes aspectos e nuances dos dados, e ao combiná-los, é possível obter um modelo com menor variância. Cada modelo individual pode ser treinado de forma independente, usando diferentes algoritmos, conjuntos de dados ou parâmetros.

Existem várias técnicas populares de *ensemble*:

- *Bagging* (Breiman, 1996): os modelos individuais são treinados em diferentes subconjuntos dos dados originais por meio da técnicas de reamostragem *Bootstrap* (Efron, 1979).
- *Boosting* (Friedman, 1996): os modelos são treinados sequencialmente, dando maior ênfase às amostras que levam aos maiores erros nos modelos anteriores.
- *Stacking* (Breiman, 1996): envolve a combinação de várias predições de diferentes modelos em um meta-modelo, que aprende a ponderar as contribuições de cada modelo individual.

Como as árvores de decisão apresentam algumas deficiências, métodos como *bagging* e *boosting* podem ajudar a minimizá-las e, assim, produzir um modelo melhor. Por exemplo, esses métodos conseguem manter a variância das covariáveis, uma vez que produz diversas árvores de decisão; as predições deixam de ser uma aproximação constante por partes para serem mais suaves e contínuas. Com uma grande dimensionalidade devido ao número de covariáveis, é possível diminuir a complexidade e o tamanho da árvore, uma vez que cada árvore pode conter um subconjunto de covariáveis, entre outras vantagens.

Na próxima seção, são apresentados e discutidos os métodos de agrupamento de modelos em árvores de regressão: Floresta Aleatória e *Gradient Boosting Machine*.

4.2.1 Bagging

Como motivação, imagine que existem duas funções de predição para Y, denotadas $\hat{f}_1(\mathbf{x})$ e $\hat{f}_2(\mathbf{x})$. Considere um estimador combinando essas duas quantidades da seguinte maneira:

$$\hat{f}(\mathbf{x}) = \frac{\hat{f}_1(\mathbf{x}) + \hat{f}_2(\mathbf{x})}{2}. \tag{4.6}$$

Usando a expressão (3.7), pode-se encontrar a esperança do erro quadrático de $f(\mathbf{x}')$:

$$E[(Y - \hat{f}(\mathbf{x}))^2] = \sigma^2 + \frac{1}{4}(Var[\hat{f}_1(\mathbf{x}) + \hat{f}_2(\mathbf{x})]) + \left(\frac{E[\hat{f}_1(\mathbf{x})] + E[\hat{f}_2(\mathbf{x})]}{2} - f(\mathbf{x}) \right)^2. \tag{4.7}$$

Assumindo que $\hat{f}_1(\mathbf{x})$ e $\hat{f}_2(\mathbf{x})$:

- são não correlacionados, isto é, $Cov[\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x})] = 0$.
- são não enviesados, isto é, $E[\hat{f}_i(\mathbf{x})] = f(\mathbf{x})$ com $i = 1, 2$.
- apresentam a mesma variância, isto é, $Var[\hat{f}_1(\mathbf{x})] = Var[\hat{f}_2(\mathbf{x})]$

Pode-se chegar a seguinte desigualdade:

$$E[(Y - \hat{f}(\mathbf{x}))^2] = Var[\epsilon] + \frac{Var[\hat{f}_1(\mathbf{x})]}{2} \leq E[(Y - \hat{f}_i(\mathbf{x}))^2], \tag{4.8}$$

sob as suposições anteriores, é melhor utilizar o estimador combinado $\hat{f}(\mathbf{x})$ do que usar $\hat{f}_i(\mathbf{x})$, com $i = 1, 2$, separadamente. A conclusão permanece válida quando se combina as B funções.

Os métodos de *bagging* têm o objetivo de melhorar as predições dadas por árvores. A ideia consiste em criar B árvores distintas e combinar seus resultados para melhorar o poder preditivo em relação a uma única árvore. Para criar as B árvores, o *bagging* utiliza B amostras bootstrap da amostra original.

Bootstrap (Efron, 1979) é um procedimento de reamostragem que usa dados de uma amostra para gerar uma distribuição amostral empírica de alguma estimativa, tomando repetidamente sub-amostras aleatórias da amostra observada de mesmo tamanho da amostra original, com reposição.

Para cada sub-amostra boostrap, cria-se uma árvore utilizando as técnicas descritas na Seção 4.1. Contudo, como é assumido na expressão (4.8) que os estimadores são não viesados, não podamos as árvores. Ou seja, não é aplicada a técnica pós-poda. Ainda assim, pode ser interessante considerar alguns hiperparâmetros para cada árvore i, $i = 1, ..., B$, utilizando-se a técnica pré-poda.

Seja $\hat{f}_i(\mathbf{x})$ uma função preditora obtida segundo a i-ésima árvore. A função de predição dada pelo *bagging* é dada por:

$$\hat{f}_b(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} \hat{f}_i(\mathbf{x}). \tag{4.9}$$

Note que o número de árvores B é escolhido usando-se as técnicas de validação cruzada e, à medida que $B \to \infty$, maior pode ser o *overfitting*.

4.2.2 Florestas Aleatórias

Na expressão (4.8), além da condição de que as funções de predição são não viesadas, assumiu-se também que são não correlacionadas. Em geral, mesmo utilizando B amostras bootstrap para construir cada uma das árvores, elas tendem a dar predições parecidas e, portanto, as funções $\hat{f}_i(\mathbf{x})$ apresentam alta correlação.

Floresta aleatória apresenta uma solução para diminuir a correlação das árvores ajustadas nas B amostras booststrap. Ao invés de escolher qual das p covariáveis são utilizadas em cada um dos nós da árvore, só é permitido que seja escolhida uma dentre as $m < p$ covariáveis. Essas m covariáveis são escolhidas aleatoriamente dentre as observadas e, a cada nó desenvolvido, um novo subconjunto de covariáveis é sorteado. Aleatorizando as amostras e as covariáveis, é esperado que as árvores construídas apresentam predições menos correlacionadas.

Tendo um conjunto de dados $D = (\mathbf{x}_i, y_i)$ com $i = 1, ..., n$ e $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, as etapas para a construção de uma floresta aleatória são dadas por:

1. Define-se o número de árvores B, o número de covariáveis k e os hiperparâmetros especificados na Seção 4.1.4 para cada árvore i com $i = 1, ..., B$.
2. Para cada $i = 1, ...B$, realizar os seguintes passos:
   (a) Seleciona-se uma amostra bootstrap D_i de tamanho n com reposição.
   (b) Constrói-se uma árvore A_i utilizando a amostra D_i da seguinte maneira:
      i. Seleciona-se, aleatoriamente em cada nó k covariáveis, em que $k \leq p$.
      ii. Utiliza-se somente as k covariáveis para encontrar a melhor partição.
      iii. Declara-se um nó terminal ao atingir os critérios especificados através dos hiperparâmetros.

Após o ajuste do modelo, a função de predição pode ser definida por:

$$\hat{f}(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} A_i(\mathbf{x}). \tag{4.10}$$

Os valores B e k são os hiperparâmetros do algoritmo da floresta aleatória e são encontrados usando as técnicas de validação cruzada.

4.2.3 Gradient Boosting Machine

Como motivação, assume-se uma função $f(x) = (x-1)^2$ no qual $x \in \mathbf{R}$ e $f : \mathbf{R} \to \mathbf{R}$. Como esta função é diferenciável, para encontrar o mínimo global de $f(x)$, basta que:

$$f'(x) = -2x + 1 = 0 \Rightarrow x_0 = 1, \tag{4.11}$$

em que x_0 é o valor mínimo.

Algebricamente, é fácil calcular o mínimo de $f(x)$. No entanto, existem funções qmais complexas e há a necessidade de se considerar métodos iterativos para identificar o mínimo. Um método que se tornou bastante popular é o método do gradiente descendente, que é um método numérico usado em problemas de otimização. Para encontrar um ponto mínimo de uma função usa-se um esquema iterativo, em que em cada passo se toma a direção negativa do gradiente. Na imagem abaixo, tem-se a ideia de como ocorre essa "descida". Note que a cada passo é identificado um novo valor de x mais próximo do ponto de mínimo.

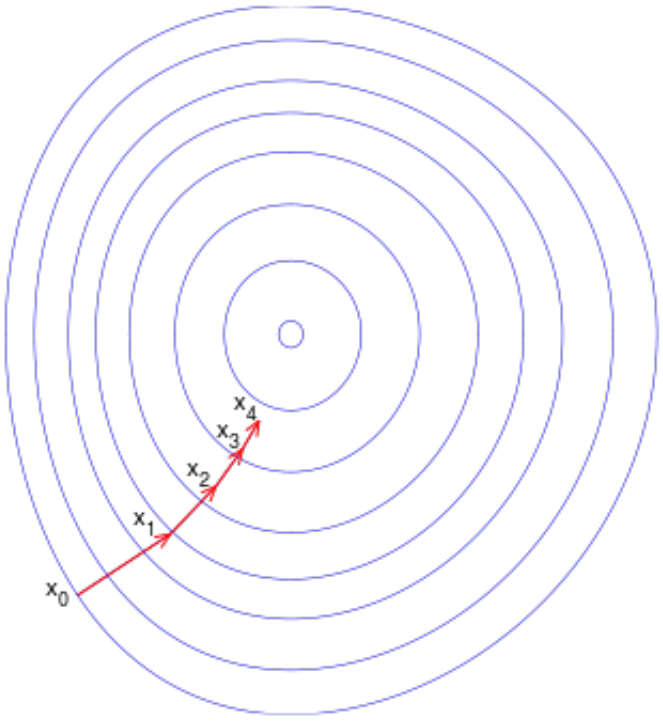

Figura 4.2: *Ilustração das curvas de níveis e cada iteração do método do gradiente. Fonte: Wikipédia*

O método do gradiente descedente é desenvolvido da seguinte maneira:

1. Define-se um número de iterações m e uma taxa de aprendizagem γ.
2. Inicializa-se um vetor $\mathbf{x}_0$ com o objetivo de alcançar o ponto mínimo de $f(\mathbf{x})$.

3. Calcula-se $f(\mathbf{x}_0)$.
4. Encontra-se $\mathbf{x}_1$, sendo que $\mathbf{x}_1 = x_0 - \gamma \mathbf{d}_1$.
5. Realiza-se os passos 3 e 4 iterativamente até se atingir a convergência ou um critério de parada, em geral o número de iterações m especificado em (1).

Note que $\mathbf{d}_i$, com $i = 1, ..., m$ e m é o vetor de derivada parciais de $f(\mathbf{x}_i)$, também chamado de vetor gradiente.

O valor γ é a taxa de aprendizagem (*learning rate*) que representa o peso considerado para atualização dos valores. A escolha do valor γ pode compensar entre a taxa de convergência e a "ultrapassagem". Embora a direção da descida seja determinada a partir do gradiente $\mathbf{d}$, a taxa de aprendizado determina o tamanho do passo nessa direção. Uma alta taxa de aprendizado γ fará com que os valores encontrados ultrapassem os mínimos, mas uma baixa taxa de aprendizado pode aumentar o tempo de convergência ou encontrar uma solução local o que é indesejável.

O *Gradient Boosting* se espelha nesse método. Através de uma função de perda e uma taxa de aprendizagem γ, a ideia é melhorar sequencialmente o treinamento do modelo. Da mesma forma que o gradiente descendente corrige os valores para se aproximar do ponto de mínimo, o *Gradient Boosting* ajusta um modelo para cada etapa da iteração com o objetivo de corrigir as observações com os maiores erros em interações anteriores. Isto é, o objetivo é desenvolver diversos modelos "fracos" para que a cada etapa as predições sejam atualizadas, obtendo-se assim um modelo melhor.

O *boosting* pode ser pensado como tendo os seguintes componentes:

- **Aprendiz base** é um modelo "simples" como a regressão linear ou a árvore de decisão. O *boosting* apresenta uma solução que melhora iterativamente qualquer modelo preditivo. Entretanto, as árvores de decisão são escolhidas devido a sua robustez e não-linearidade.
- **Aprendiz fraco** é um modelo cuja medida de erro é ligeiramente melhor que uma média. Quando o aprendiz base é uma árvore de decisão, o aprendiz fraco é uma árvore com relativamente poucas divisões, denominada árvore rasa.
- **Treinamento sequencial** no qual cada etapa é otimizada usando os pseudo-resíduos (a serem definidos adiante) de árvores previamente treinadas (i.é, ajustadas) para melhorar o desempenho preditivo. Ao se fazer isso, permite-se que cada etapa do treinamento foque nos erros da árvore anterior.

Suponha um conjunto de dados $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, com n o número de observações e árvores de decisão A_k, $k = 1, ..., m$ e m sendo o número de árvores. Também, seja γ a taxa de aprenzidado escolhida arbitrariamente, $0 < \gamma \leq 1$, e uma função de perda diferenciável $L(y, F(\mathbf{x}))$, Friedman (2001) propôs o algoritmo do *Gradient Boosting Machine* da seguinte maneira:

1. Encontra-se o valor do $F_0(\mathbf{x})$, dado por

$$F_0(\mathbf{x}) = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} L(y, \alpha).$$

2. Para $k = 1, ..., m$:

   (a) Calcula-se os pseudo-resíduos r_{ik} (i.é, os gradientes) para cada amostra, dados por:

   $$r_{ik} = -\left[\frac{\partial L(y_i, F(\mathbf{x}))}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{k-1}(\mathbf{x})},$$

   com $i = 1, ..., n$.

   (b) Ajusta-se uma árvore de decisão A_k usando como variável resposta r_{ik} e o vetor de covariáveis $\mathbf{x}_i$. Com isso, a árvore A_k cria j folhas com $j = 1, ..., J_k$.

   (c) Para cada folha j, $j = 1, ..., J_k$, calcula-se o valor da predição α_{jk} dado por:

   $$\alpha_{jk} = \underset{\alpha}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in R_j} L(y_i, F_{k-1}(\mathbf{x_i}) + \alpha),$$

   sendo R_j as partições sob o vetor de covariáveis $\mathbf{x}$ que resultam na folha j.

   (d) Atualiza-se a predição para cada observação i de acordo com a taxa de aprendizado γ:

   $$F_k(\mathbf{x}_i) = F_{k-1}(\mathbf{x}_i) + \gamma A_k(\mathbf{x}_i),$$

   no qual $A_k(\mathbf{x}_i) = \sum_{j=1}^{J_k} \alpha_{jk} I(\mathbf{x_i} \in R_j)$.

3. Calculam-se as predições usando $F_m(\mathbf{x})$.

Note que o número de árvores m e a taxa de aprendizado γ podem ser escolhidos usando as técnicas de validação cruzada.

Se a função de perda escolhida for idêntica ao método de mínimos quadrados,

$$L(y, \alpha) = \frac{\sum_{i=1}^{n}(y_i - \alpha)^2}{2}, \tag{4.12}$$

pode-se encontrar os valores de $F_0(x)$ e α_{jk}, isto é,

$$\begin{aligned}
&F_0(\mathbf{x}_i) = \underset{\alpha}{\operatorname{argmin}}\, L(y, \alpha) \\
&\implies \frac{\partial L(y, \alpha)}{\partial \alpha} = -\sum_{i=1}^{n}(y_i - \alpha) = 0 \implies \alpha = \bar{y}. \\
&\alpha_{jk} = \underset{\alpha}{\operatorname{argmin}}\, L(y, \alpha + F_{k-1}(\mathbf{x})) \\
&\implies \frac{\partial L(y, \alpha + F_{k-1}(\mathbf{x_i}))}{\partial \alpha} = -\sum_{i\,:\,\mathbf{x_i} \in R_j} (y_i - F_{k-1}(\mathbf{x}_i) - \alpha) = 0 \\
&\implies \alpha = \frac{\sum_{i\,:\,\mathbf{x_i} \in R_j} y_i - F_{k-1}(\mathbf{x}_i)}{|\mathbf{x}_i \in R_j|}.
\end{aligned} \tag{4.13}$$

Logo, o algoritmo do *Gradient Boosting* pode ser definido pelos seguintes passos:

1. Define-se o número de árvores m e a taxa de aprendizado γ.
2. Calcula-se $\bar{y}$ e assumir que $F_0(\mathbf{x}_i) = \bar{y}$.
3. Calcula-se os pseudo-resíduos r_{i0} para cada amostra, dado por:

$$r_{i0} = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x}_i)=F_0(\mathbf{x}_i)} = y_i - F_0(\mathbf{x}_i) = y_i - \bar{y}. \tag{4.14}$$

4. Ajusta-se uma árvore de regressão A_1 usando r_{i0} como a variável resposta e $\mathbf{x}_i$ como o vetor de covariáveis.
5. Para cada folha j da árvore A_1, calcula-se a predição da seguinte forma:

$$\alpha_{j1} = \frac{\sum_{i:\mathbf{x}_i \in R_j} y_i - F_0(\mathbf{x}_i)}{|\mathbf{x}_i \in R_j|}, \tag{4.15}$$

   sendo R_j, $j = 1, ..., J_0$, os estratos definidos sob o vetor de covariáveis x que resulta na folha j.
6. Atualiza-se a predição para cada observação i usando a seguinte equação:

$$F_1(\mathbf{x}_i) = F_0(\mathbf{x}_i) + \gamma \alpha_{j0} I(\mathbf{x}_i \in R_j). \tag{4.16}$$

7. Repetem-se as etapas 1 à 6 até se encontrar a função de predição $F_m(\mathbf{x}_i)$.

O *Gradient Boosting Machine* pode apresentar resultados mais satisfatórios que a floresta aleatória e uma única árvore de decisão, pois como é ajustado para o corrigir os erros dos modelos de forma sequencial, esse algoritmo é capaz de capturar padrões mais complexos nos dados. No entanto, o método apresenta algumas desvantagens ao ser utilizado:

- Como o objetivo é melhorar o modelo minimizando os erros sequencialmente, isso pode superenfatizar *outliers* e causar sobreajuste.
- Computacionalmente é caro treiná-lo, isto é, é bastante exaustivo treinar o algoritmo em termo de tempo e memória computacional.

Capítulo 5

Árvores de Sobrevivência

As técnicas de árvores de decisão em análise de sobrevivência podem ser utilizadas para predizer a duração do tempo até um evento ocorrer. Essas abordagens combinam os princípios das árvores de decisão, que são modelos de aprendizado de máquina interpretáveis e intuitivos, com os conceitos da análise de sobrevivência, que lidam com conjuntos de dados de sobrevivência $(T_i, \delta_i, \mathbf{x}_i)$ com $i = 1, ..., n$, em que $\mathbf{x}_i$ é o vetor de covariáveis, $T_i = min(T_i^*, C_i)$ em que o T_i^* é o tempo até o evento e C_i é o tempo até a censura, e δ_i é o indicador de evento.

Segundo Segal (1988), a divisão usando a estatística de Welch (Welch, 1947), isto é, a estatística T para comparação de duas médias populacionais, e a divisão usando os métodos de mínimos quadrados apresentam resultados comparáveis quando é mantida a estrutura dos dados. Quando há transformação na variável resposta, pode-se obter resultados diferentes. Entretanto, a opção pelo uso de uma estatística de teste como critério de divisão pode trazer algumas vantagens e benefícios adicionais:

1. **Invariância sob transformação monótona da resposta Y**. As árvores ajustadas pelo métodos de mínimos quadrados possuem essa invariância apenas em relação às transformações monótonas das covariáveis. Ou seja, a divisão ótima é a mesma, independentemente de usarmos X_1 ou $X_1^* = g(X_1)$ para alguma função monótona g. No entanto, é somente por meio do uso de estatísticas de teste para duas amostras que as melhores divisões são preservadas sob transformações monótonas.

2. **Falta de sensibilidade para detectar *outliers* na variável resposta**. O uso de mínimos quadrados está sujeito à sensibilidade a observações extremas de Y. Isso, no contexto da árvore de regressão, não necessariamente é uma desvantagem uma vez que tais *outliers* podem ser isolados em seus próprios nós terminais. Ainda assim, a influência na topologia da árvore pode ser distorcida e a interpretação das divisões que levam ao isolamento do *outlier*, podendo ocasionar em um modelo sobreajustado.

3. **Extensão para variável resposta com censura**. A motivação principal para a mudança do critério de divisão é possibilitar o uso de técnicas de árvore para dados de sobrevivência. Em vez de usar estatísticas na comparação de duas amostras considerando apenas os dados não censurados como critérios de qualidade da divisão, são utilizadas estatísticas modificadas que levam em consideração a censura.

A construção de uma árvore de decisão em análise de sobrevivência envolve a criação de regras de divisão com base em estatísticas, como a utilizada no teste de *log-rank* apresentada na Seção 2.5. Essas regras de divisão são aplicadas iterativamente para dividir os dados em grupos cada vez mais homogêneos.

Uma vez construída a árvore, ela pode ser usada para predizer a duração da sobrevivência para novos indivíduos com base em suas características. A estrutura da árvore permite uma interpretação direta das covariáveis mais importantes e de suas interações, facilitando a compreensão dos fatores que influenciam a sobrevivência.

Além disso, existem variações e extensões das árvores de decisão, como a Floresta de Sobrevivência Aleatória (*Random Survival Forest*) e o *Gradient Boosting Survival Tree*, que oferecem melhorias e recursos adicionais. Essas técnicas têm se mostrado eficazes na modelagem em dados de sobrevivência e fornecem insumos valiosos para a compreensão dos fatores que afetam a sobrevida em diferentes contextos.

Neste capítulo, são apresentadas as árvores de sobrevivência, explicando como são desenvolvidas e como são úteis para problemas de predição; são também discutidos os conceitos de Floresta de Sobrevivência Aleatória e *Gradient Boosting Survival Tree*.

5.1 Divisão dos dados

Ao invés de dividir os dados de forma a otimizar a homegeneidade **dentro** do nó, pode-se obter divisões que resultam em separação **entre** os nós. Assim, qualquer estatística de comparação de duas amostras fornece um critério de divisão.

Para a construção de uma árvore de sobrevivência, o algoritimo de divisão dos dados é definido pelas seguintes etapas:

1. Para cada covariável x_k, $k = 1, ..., p$:
   (a) Constrói-se um vetor $\mathbf{c}_k$ que contém os valores distintos observados para a covariável x_k.
   (b) Para cada valor c_{ik}, $i = 1, ..., l$, em que $l = |\mathbf{c}_k|$:
      i. Cria-se uma divisão $I(x_k > c_{ik})$, em que $x_k > c_{ik}$ é considerada como população 1 e $x_k \leq c_{ik}$ é considerada como população 2.
      ii. Calcula-se os valores descritos na Tabela 5.1.
      iii. Dado o peso w_j pré-especificado, calcula-se a estatística QW expressa na equação (2.16).
   (c) Seleciona-se uma constante c_k^* que representa o valor c_{ik} associado ao maior valor da estatística QW.
2. Seleciona-se a covariável x^* e a constante c^* que representam o maior valor da estatística QW entre todas as covariáveis x_k e constantes c_k^* da etapa 1.
3. Constrói-se o nó usando a função indicadora $I(x^* > c^*)$.
4. Repete-se o processo em cada nó iterativamente, até atingir as folhas (i.é, nós finais).

Apesar das diferentes formas de construir a estatística QW, há muita similaridade nas árvores construídas por essas diferentes medidas conforme resultados obtidos por simulações e aplicações em diferentes conjuntos de dados (Segal, 1988). Por isso, a estatística mais utilizada é a *log-rank* expressa na equação (2.15).

Note que o algoritmo acima apresenta uma solução para particionar os dados em nós, porém é necessário um critério de parada para que o treinamento da árvore de sobrevivência possa atingir as folhas e realizar predições.

5.2 Critério de parada

De forma análoga à árvore de regressão, as árvores de sobrevivência apresentam alguns métodos para a definição de um critério de parada:

- Antes da fase de treinamento, são especificados alguns hiperparâmetros para controlar o crescimento da árvore. O método mais utilizado é a pré-poda e os critérios são os mesmos especificados na Seção 4.1.4.
- Uma outra maneira de realizar o critério de parada é através de uma generalização do custo de complexidade apresentado por Gordon e Olshen (1985). A ideia é treinar uma árvore de tamanho máximo e, através de uma função de custo, reduzí-la. De forma um pouco mais específica, seja A uma árvore ajustada em seu tamanho máximo e $G(h)$ o valor da estatística TW para o nó h. Então o custo de complexidade pode ser definido como

$$G_\alpha(h) = \sum_{h \notin F} G(h) - \alpha|A - F|, \tag{5.1}$$

  em que F é o conjunto de folhas (i.é, nós terminais) de A, $|F|$ é a quantidade de folhas e $|A|$ a quantidade de nós. Observe que o somatório na expressão (5.2) é apenas dos nós internos. O sinal negativo reflete o fato de que G_α deve ser maximizada diferentemente de C_α expresso em (4.5) que deve ser minimizada. Para escolha de α, é necessário o uso de técnicas de validação cruzada.
- Durante o treinamento, pode-se "podar" alguns nós usando uma técnica recomendada por Segal (1988). Para cada nó h em uma árvore A, incluindo o nó raiz, calcula-se o máximo da estatística $|QW|$ considerando todas as divisões que são antecedentes ao nó h. Em seguida, os valores para todos os nós são ordenados em ordem decrescente, e um ponto de corte é determinado. Se um nó interno corresponde ao menor valor do ponto de corte, todos os seus descendentes são podados. Este método é conhecido como abordagem *bottom-up*. Essa abordagem apresenta menor custo computacional em comparação com a definição de um ponto de corte antes do ajuste da árvore, pois evita a necessidade de um reajuste da árvore a cada escolha de um ponto de corte.

No método pré-poda, escolhe-se um conjunto de hiperparâmetros arbitrariamente que apresente o melhor desempenho preditivo, avaliado através de técnicas de validação cruzada. Nota-se que outros métodos de controle podem ser utilizados.

5.3 Predição

Após o ajuste da árvore, as folhas resultantes são utilizadas para construir as quantidades apresentadas no Capítulo 2. Com isso, pode-se aplicar algumas estatísticas ou funções discutidas anteriormente:

- A estimativa da função de taxa de risco acumulada $\hat{\Lambda}_f(t|\mathbf{x})$, baseada no estimador de Nelson-Aalen, é calculada a partir de todas as amostras dos dados de treinamento que estão na mesma folha f.
- A estimativa da função de sobrevivência $\hat{S}_f(t|\mathbf{x})$, obtida através do estimador Kaplan-Meier, é calculada a partir de todas amostras no conjunto de treinamento que estão na mesma folha f.
- O número total de eventos, que pode ser estimado pela soma da função de taxa de risco acumulada no nó terminal

$$\sum_{j=1}^{n_f} \hat{\Lambda}_f(t_j|\mathbf{x}), \tag{5.2}$$

  em que n_f o número de tempo distintos na folha f.

5.4 Floresta de Sobrevivência Aleatória

A floresta de sobrevivência aleatória (do inglês, *Random Survival Forest*) é uma especialização da técnica apresentada na Seção 4.2.2. e proposta por Ishwaran et al. (2008) para lidar com problemas em análise de sobrevivência. O procedimento é descrito a seguir.

Considerando um conjunto de dados $D = (T_i, \delta_i, \mathbf{x}_i)$ com $i = 1, ..., n$ e $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$:

1. Define-se o número de árvores B, o número de covariáveis k e os hiperparâmetros especificados na Seção 4.1.4.
2. Para cada árvore $i = 1, ..., B$, realizam-se os seguintes passos:
   (a) Seleciona-se uma amostra bootstrap D_i de tamanho n, com reposição.
   (b) Constrói-se uma árvore de sobrevivência A_i utilizando a amostra D_i da seguinte maneira:
      i. Seleciona-se aleatoriamente, em cada nó, k covariáveis em que $k \leq p$.
      ii. Utiliza-se somente as k covariáveis para identificar a melhor divisão dos dados.
      iii. Definem-se as folhas quando se atingir os critérios de parada.

Após o treinamento do algoritmo, podem ser utilizadas diferentes formas de realizar uma predição:

- A estimativa da função de taxa de risco acumulada em t. Para cada árvore i, a função de taxa de risco acumulada $\Lambda_i(t|\mathbf{x})$ pode ser calculada usando o estimador de *Nelson-Aalen*. Assim, para se obter o valor no ponto t dessa função, aplica-se a média sob todas as árvores i,

$$\hat{\Lambda}(t|\mathbf{x}) = \sum_{i=1}^{B} \frac{\hat{\Lambda}_i(t|\mathbf{x})}{B}. \tag{5.3}$$

- A estimativa da função de sobrevivência em t. O cálculo ocorre de forma análoga a função de taxa de risco acumulada, porém é utilizado o estimador *Kaplan-Meier* e, posteriormente, aplica-se a média sob todas as árvores i,

$$\hat{S}(t|\mathbf{x}) = \sum_{i=1}^{B} \frac{\hat{S}_i(t|\mathbf{x})}{B}. \tag{5.4}$$

- A média do número total de eventos, que pode ser estimada pela soma da função de taxa de risco acumulada calculada pelo estimador de Nelson-Aalen,

$$\sum_{i=1}^{B} \frac{\sum_{j=1}^{n_i} \hat{\Lambda}_i(t_j|\mathbf{x})}{B}. \tag{5.5}$$

  sendo t_j os tempos de falhas distintos e observados no conjunto de treinamento, n_i a quantidade de tempos distintos na sub-amostra i e B o número de árvores/sub-amostras.

5.5 Gradient Boosting Survival Tree

Da mesma maneira que há uma extensão da floresta aleatória para dados de sobrevivência, existe uma extensão para o *Gradient Boosting Machine*. Ridgeway (1999) propõe diferentes formas de usar o algoritmo de *Gradient Boosting* em outros tipos de variável resposta. Dentre elas, o uso da função de log-verossimilhança parcial de Cox (Cox, 1975) como função de perda.

Assim, considerando o conjunto de dados $(T_i, \delta_i, \mathbf{x}_i)$ com $i = 1, ...n$ e as expressões apresentadas na Seção (2.6), pode-se encontrar uma função $f(\cdot)$ em que $f : \mathbf{R}^p \to \mathbf{R}$ maximiza a seguinte quantidade:

$$L(T, \delta, f(\mathbf{x})) = \sum_{i=1}^{n} \delta_i \left[f(\mathbf{x}_i) - \log \left(\sum_{j \in R(t_i)} \exp(f(\mathbf{x}_j)) \right) \right] \tag{5.6}$$

com $R(t_i)$ é o conjunto dos índices das observações não censuradas e para as quais o evento ainda não ocorreu até o instante t.

Considerando como função de perda $\Phi(t, \delta, f(\mathbf{x})) = -L(t, \delta, f(\mathbf{x}))$, tem-se algoritmo desenvolvido por Ridgeway (1999) que consiste dos seguintes passos:

1. Define-se o número de árvores m.
2. Assume-se que o logaritmo da função de risco inicial, representada aqui por $F_0(\mathbf{x}_i)$ é igual a zero.
3. Para $k = 1, ..., m$, realizam-se as seguintes etapas:
   (a) Calculam-se pseudo-resíduos r_{ik} dados por:

$$r_{ik} = -\left[\frac{\partial \Phi(T_i, \delta_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right]_{f(\mathbf{x}_i)=F_{k-1}(\mathbf{x}_i)} = \delta_i - \sum_{k \in R(t_i)} \delta_k \frac{\exp(F_{k-1}(\mathbf{x}_i))}{\sum_{j \in R(T_i)} \exp(F_{k-1}(\mathbf{x}_j))} \quad (5.7)$$

   em que $F_k(x_i)$ é o log da função de risco no passo k.

   (b) Ajusta-se uma árvore de regressão descrita no Capítulo 4 usando como variável resposta r_{ik} e vetor de covariáveis $\mathbf{x}_i$. Assim, pode se obter a predição $\hat{r}_{ik}$.

   (c) Ajusta-se um modelo de riscos proporcionais de Cox (visto na Seção 2.6) usando como variável resposta (T_i, δ_i) e como covariável $\hat{r}_{ik}$. Com isso, pode-se obter o coeficiente da regressão γ_k da covariável $\hat{r}_{ik}$, denominado como taxa de aprendizagem.

   (d) Atualiza-se a predição para cada observação i de acordo com a taxa de aprendizado γ_k:

$$F_k(\mathbf{x}_i) = F_{k-1}(\mathbf{x}_i) + \gamma_k \hat{r}_{ik} \quad (5.8)$$

4. Calculam-se as predições usando $F_m(\mathbf{x})$.

Algumas observações sobre o algoritmo:

- O aprendiz base (descrito na Sub-seção 4.2.3) é uma árvore de regressão. Isto ocorre pelo fato de que os pseudo-resíduos podem ser tanto positivos como negativos.
- Como o objetivo do algoritmo é minimizar os "erros" sequencialmente, é realizada uma transformação na função $L(t, \delta, f(\mathbf{x}))$, ou seja, é construída uma função $\Phi : \mathbf{R} \to \mathbf{R}$ em que $\Phi(t, \delta, f(\mathbf{x})) = -L(t, \delta, f(\mathbf{x}))$. Ao fazer isso, se $L(t, \delta, f(\mathbf{x}))$ apresenta um ponto de máximo, então $\Phi(t, \delta, f(\mathbf{x}))$ apresenta um ponto de minímo.
- Ao invés de fixar a taxa de aprendizagem, o valor de γ_k é ajustado a cada interação do algoritmo usando o modelo de Cox conforme explicado no item 3.(c).

Para realizar as predições usando o *Gradient Boosting Survival Tree*, tem-se:

- A log função de taxa de risco é dada por $F_m(\mathbf{x})$. Consequentemente, a função de taxa de risco é definida por $\exp(F_m(\mathbf{x}))$.

- A função de taxa de risco acumulada que é expressa por $\hat{\Lambda}(t|\mathbf{x}) = \exp(F_m(\mathbf{x}))\hat{\Lambda}_0(t)$. O valor $\hat{\Lambda}_0(t)$ é calculado através do estimador de Breslow.
- A função de sobrevivência que é medida por $\hat{S}(t|\mathbf{x}) = \hat{S}_0(t)^{\exp(F_m(\mathbf{x}))}$. O valor $\hat{S}_0(t)$ é encontrado através do estimador de Breslow.

Os valores discrepantes terão resíduos muito maiores do que os não discrepantes, portanto, o aumento de gradiente concentrará uma quantidade desproporcional de sua atenção nesses pontos.

Da mesma forma que o *Gradient Boosting Machine*, este método também apresenta uma desvantagem quando lidamos com conjuntos de dados contendo *outliers*. Essas observações atípicas têm o potencial de influenciar de forma desproporcional o processo iterativo de ajuste do modelo. Isso ocorre porque os resíduos associados a esses *outliers* podem ser significativamente maiores do que os resíduos das observações não atípicas. Como resultado, o aumento do gradiente pode concentrar uma quantidade desproporcional de atenção nesses pontos atípicos, podendo impactar negativamente o desempenho preditivo do modelo. Além disso, outra consideração importante é o custo computacional associado ao treinamento do *Gradient Boosting Survival Tree*. Devido ao seu processo sequencial, pode-se demandar mais recursos computacionais e tempo para ajustar os parâmetros e gerar resultados precisos, especialmente em conjuntos de dados com alta dimensionalidade.

Capítulo 6

Aplicação aos dados ICESP

Este capítulo apresenta uma análise exploratória dos dados, os resultados provenientes da aplicação dos métodos de árvores de sobrevivência abordados no Capítulo 5 e o modelo de risco proporcionais de Cox especificado na Seção 2.6. Com o intuito de explicar e aplicar os métodos apresentados, foram escolhidas algumas variáveis dos dados ICESP que se mostraram clinicamente importantes em análises anteriores (não apresentadas aqui) e em discussões com os pesquisadores clínicos. As variáveis consideradas estão descritas na Tabela A.1.

6.1 Análise Exploratória dos Dados

Uma breve análise exploratória dos dados ICESP é apresentada a seguir. No total, foram considerados 793 pacientes. Na tabela abaixo, apresentamos algumas estatísticas das variáveis numéricas. Pode-se observar, a príncipio, que:

- Temos apenas 2 registros sem pontos sobre a qualidade de vida (QV.0).
- A idade média e mediana são aproximadamente de 60 anos.
- Há uma alta variação na quantidade de dias de hospitalização, caracterizando que podemos ter pacientes que ficam mais dias hospitalizados do que em média e caracterizando uma distribuição assimétrica.
- Apesar do IMC médio estar dentro do peso normal, temos que 25% dos pacientes estão acima de 27,45 de ICM indicando pré-obesidade.

Tabela 6.1: *Descrição das variáveis numéricas.*

Variável	Dados Omissos (%)	Min	Média	D.P	Q1	Q2	Q3	Max
Idade	0 (0,0%)	16,23	61,57	14,28	54,00	63,00	71,00	91,20
IMC	0 (0,0%)	11,93	24,26	5,29	20,67	23,53	27,45	44,27
Capac. Func.	0 (0,0%)	0,00	1,15	1,23	0,00	1,00	2,00	4,00
Hosp. pré-UTI	0 (0,0%)	0,00	4,14	16,52	0,00	1,00	3,00	134,00
QV0	2 (0,2%)	-0,59	0,47	0,43	-0,59	0,62	0,80	1,00

(a) *D.P: desvio padrão, Q1: primeiro quartil, Q2: mediana, Q3: terceiro quartil.*

Pode-se notar a distribuição dos pacientes segundo algumas variáveis qualitativas dicotômicas é apresentada na tabela a seguir. A maioria dos pacientes apresentam as seguintes características:

- Não apresenta insuficiência renal, respiratória e cardíaca.
- Não apresenta diagnóstico de diabetes, cirrose, alcoolismo e delírio.
- Realiza cirurgia.

Além disso, são poucos os pacientes que têm dados faltantes para essas variáveis.

Tabela 6.2: *Descrição das variáveis qualitativas dicotômicas.*

Variável	Sim (%)	Não (%)	Dados Faltantes (%)
Insuf. Renal	20 (2,5%)	773 (97,5%)	0 (0,0%)
Insuf. Resp	19 (2,4%)	773 (97,5%)	1 (0,1%)
Cirrose	13 (1,6%)	780 (98,4%)	0 (0,0%)
Alcoolismo	93 (11,7%)	700 (88,3%)	0 (0,0%)
Diabetes	122 (15,4%)	671 (84,6%)	0 (0,0%)
Insuf. Cardiaca	32 (4,0%)	761 (96,0%)	0 (0,0%)
Delírio	124 (15,6%)	669 (84,4%)	0 (0,0%)
Cirurgia	481 (60,7%)	311 (39,2%)	1 (0,1%)
Quimioterapia	302 (38,1%)	489 (61,7%)	2 (0,2%)
Radioterapia	171 (21,6%)	619 (78,1%)	3 (0,3%)

Analisando a Figura 6.1, tem-se as seguintes interpretações em relação aos tipos de cirurgia, quimioterapia e radioterapia:

- Mais de 50% dos pacientes faz cirurgia com intuito curativo ou adjuvante.
- A maioria dos pacientes não faz radioterapia e quimioterapia.
- Aproximadamente 20% dos pacientes faz radioterapia com intuito curativo ou adjuvante.
- Aproximadamente 30% dos pacientes faz quimioterapia com intuito curativo ou adjuvante.
- Uma minoria dos pacientes fazem quimioterapia, radioterapia e/ou cirurgia como opção de melhora ou alívio momentâneo (palitativa).

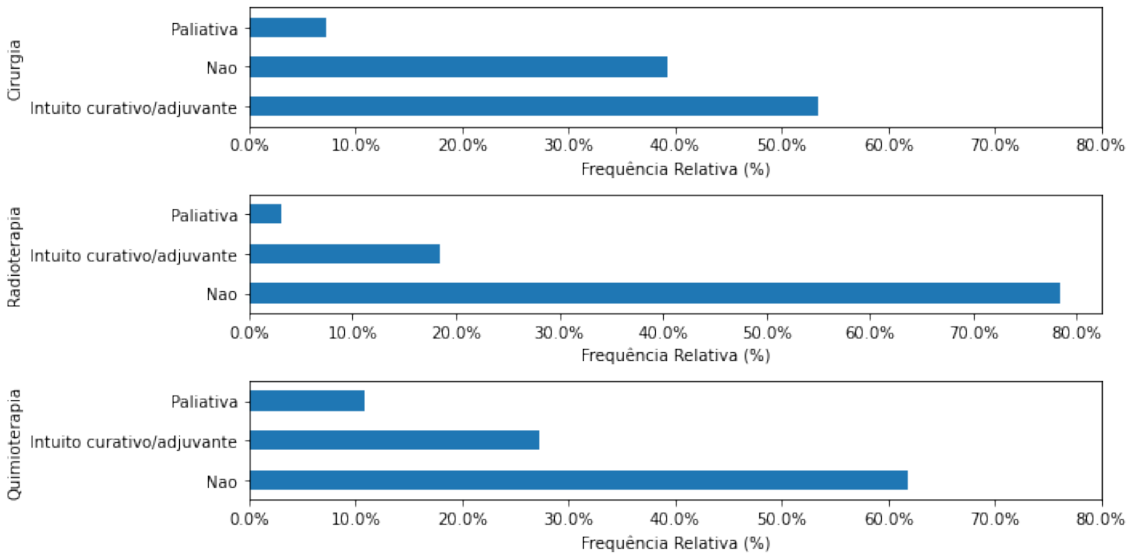

Figura 6.1: *Distribuição dos tipos de cirurgia, quimioterapia e radioterapia.*

Na Figura 6.2, a distribuição dos pacientes segundo sexo é ilustrada no gráfico de barras. Pode-se concluir que aproximadamente 60% dos pacientes são do sexo masculino e 40% dos pacientes são mulheres.

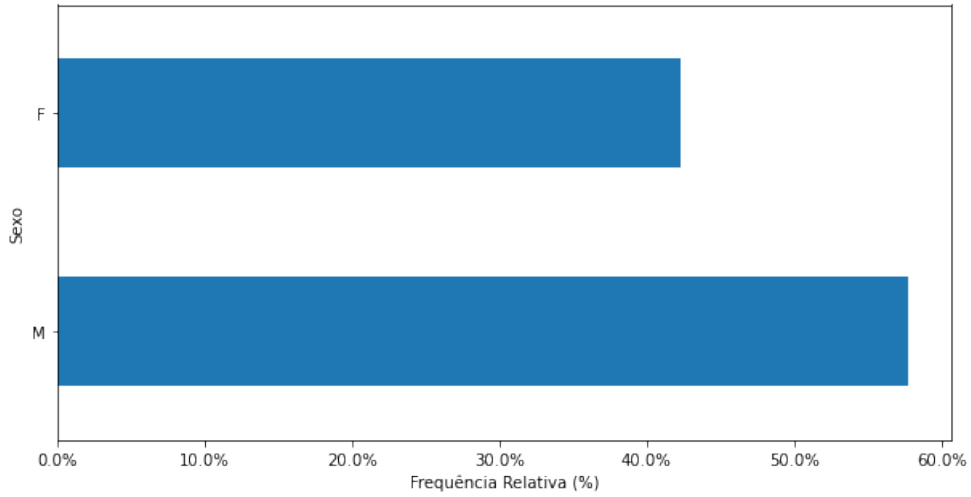

Figura 6.2: *Distribuição do sexo dos pacientes.*

Na Figura 6.3, nota-se que mais de 90% dos pacientes apresentam o status do câncer como ativo, sendo aproximadamente 60% com o diagnóstico recente e 40% em recaída e progressão. Além disso, menos de 10% dos indivíduos apresentam o câncer em remissão.

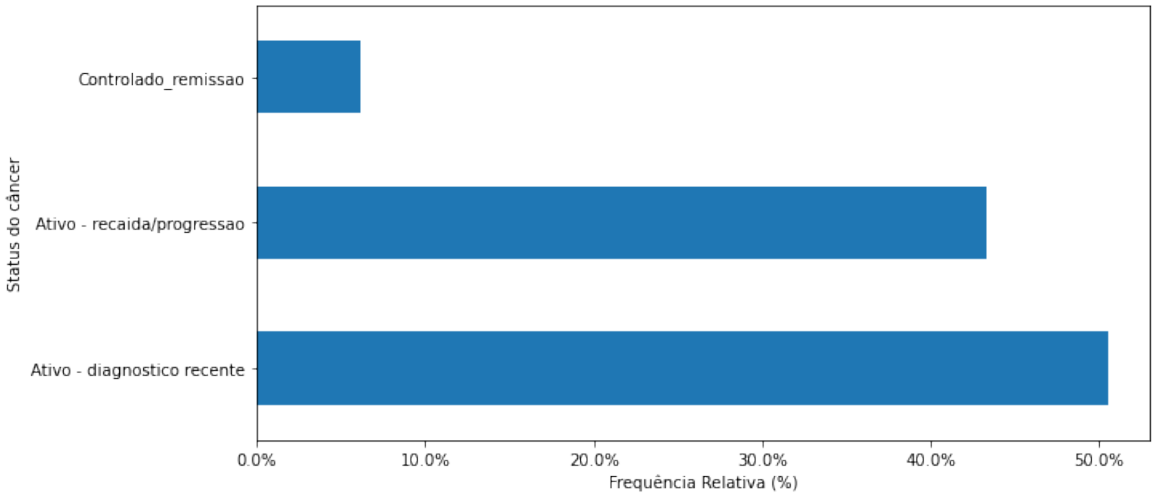

Figura 6.3: *Distribuição do status do câncer dos pacientes.*

Na Figura 6.4, percebe-se que 50% dos pacientes são admitidos na UTI após de cirurgias eletivas, aproximadamente 40% dos pacientes por motivos clínicos e menos 10% admitidos devido a situações emergenciais.

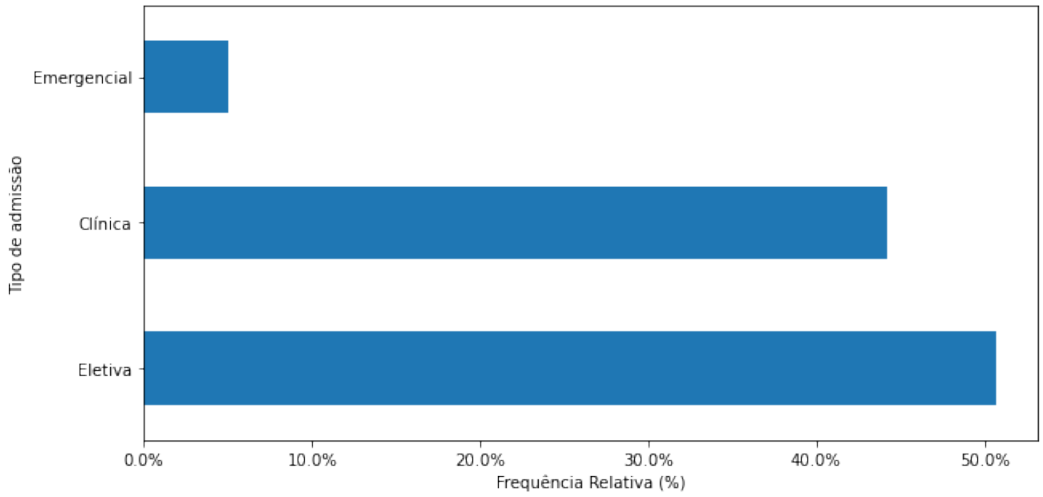

Figura 6.4: *Distribuição do tipo de admissão dos pacientes.*

Na Figura 6.5, é apresentada a distribuição do local de neoplasia dos pacientes. Observe que mais de 30% dos pacientes apresentam neoplasia no trato digestivo, menos de 20% no sistema genital e aproximadamente 10% no trato unitário. A minoria dos pacientes tem neoplasia no cérebro.

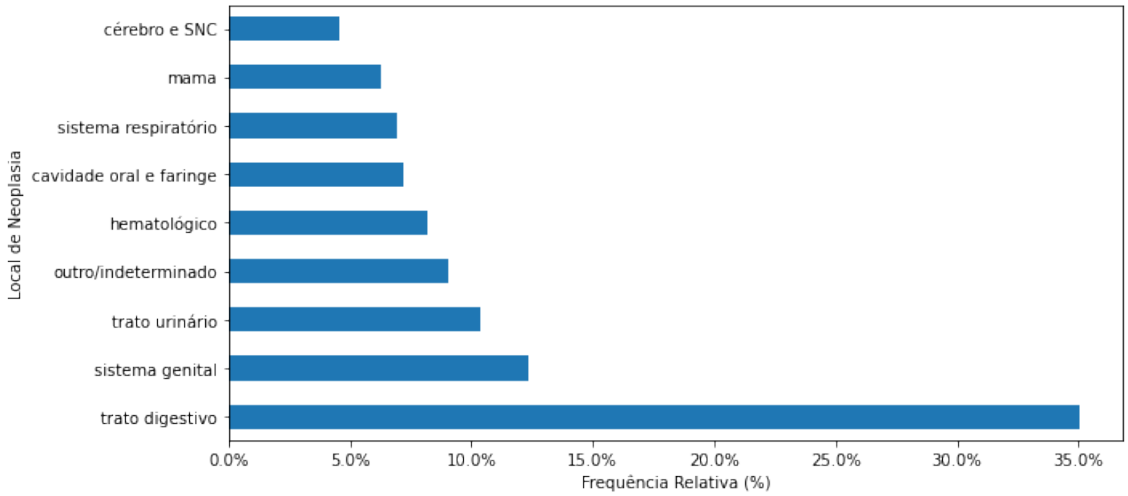

Figura 6.5: *Distribuição do local de neoplasia dos pacientes.*

Pela Figura 6.6 pode-se avaliar como a extensão do câncer está distribuída entre os pacientes. Em geral, a maioria dos pacientes tem câncer localmente avançado ou limitado, representando mais de 60%. Ainda, tem-se que mais de 30% apresenta metastase a distancia. Menos de 5% dos pacientes tem leucemias.

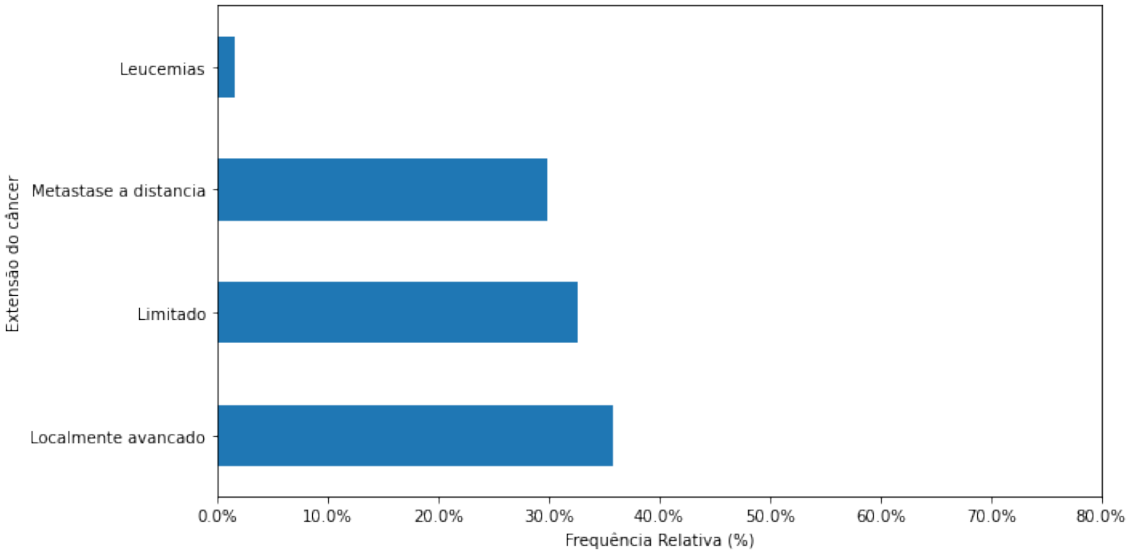

Figura 6.6: *Distribuição da extensão do câncer.*

A experiência de sobrevivência dos pacientes na amostra é ilustrada através do gráfico a seguir, que apresenta estimativas Kaplan-Meier da probabilidade de sobrevivência ao longo do tempo. Pela figura 6.7 pode-se ver que a sobrevivência diminui rapidamente nos primeiros dias. O tempo mediano é aproximadamente de 200 dias com erro padrão de aproximadamente 8 dias.

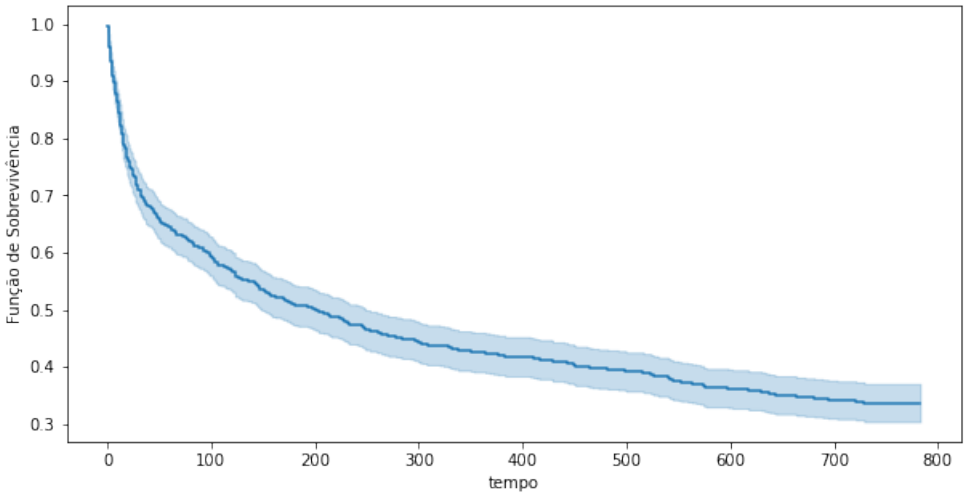

Figura 6.7: *Probabilidade de sobrevivência em diferentes tempos t.*

Um análise mais detalhada desses dados, incluindo informações sobre a qualidade de vida dos pacientes, é apresentada em Normillo et al (2016).

6.2 Construindo Árvores de Sobrevivência

Ao empregar árvores de sobrevivência nos dados da ICESP, é possível identificar fatores prognósticos relevantes, como idade, estágio da doença e tipos específicos de tratamento, que influenciam diretamente na sobrevivência dos pacientes.

Antes de realizar o ajuste dos dados em uma árvore de decisão, é possível fazer a escolha arbitrária dos hiperparâmetros dados pelos valores de profundidade e pelo número mínimo de pacientes nas folhas. Nas análises a seguir, é possível avaliar como esses parâmetros influenciam os ajustes da árvore e como as covariáveis são utilizadas para prever o tempo até o óbito de pacientes com câncer.

Na Figura 6.8, está representada uma árvore de sobrevivência ajustada aos dados ICESP, com uma profundidade igual a 2. Cada nó representa uma condição, onde o ramo da direita indica que essa condição é verdadeira, enquanto o ramo da esquerda indica que a condição é falsa. É notável que apenas duas variáveis foram consideradas na construção da árvore: o escore que procura avaliar a mortalidade hospitalar na admissão na UTI (SAPS3_Pontos) e o índice de massa corporal (imc) do paciente. Além disso, observa-se que a árvore resultante possui apenas quatro folhas, identificadas pelos nós 2, 3, 5 e 6.

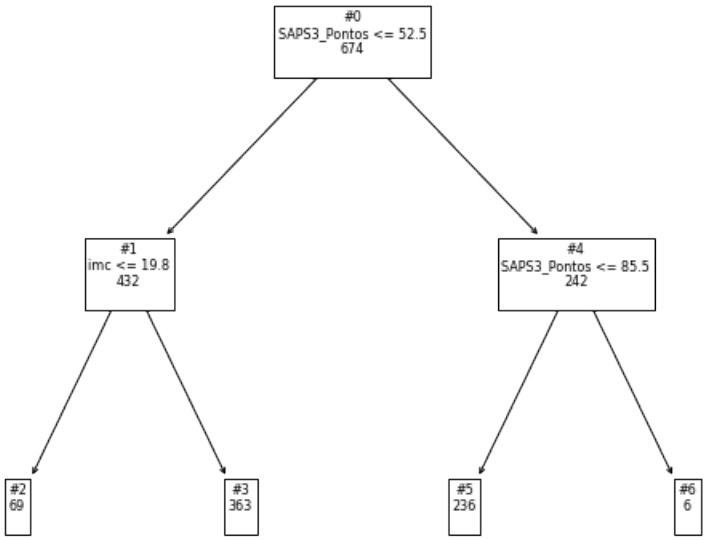

Figura 6.8: *Árvore de sobrevivência ajustada aos dados ICESP com profundidade 2.*

Pode-se analisar as diferenças na experiência de sobrevivência entre essas quatro folhas utilizando as estimativas de Kaplan-Meier. A partir da Figura 6.9, pode-se obter as seguintes interpretações:

- a folha 3 engloba pacientes com escores SAPS3 abaixo ou igual a 52,5 e IMC acima de 19,8. Este grupo contém 363 pacientes e demonstra a maior probabilidade de sobrevivência, com um tempo mediano de 729 dias.
- a folha 5 representa pacientes com escores SAPS3 entre 52,5 e 85,5. Notavelmente, este é o segundo grupo com maior probabilidade de sobrevivência, evidenciando um tempo mediano de 130 dias e contendo 236 pacientes.
- a folha 2 abrange pacientes com escores SAPS3 abaixo ou igual a 52,5 e IMC abaixo ou igual a 19,8. Nesse contexto, este grupo contém 69 pacientes e apresenta o segundo maior risco, com um tempo mediano de 24 dias.
- a folha 6 inclui pacientes com escores SAPS3 acima de 85,6, representando o grupo com menor quantidade de pacientes (6 pacientes) e com maior risco, exibindo um tempo mediano de 3 dias.
- pode-se observar que quanto maior os valores dos escores SAPS3 sobre a situação de vida, menor é o tempo de sobrevida do paciente.
- o IMC serve como um discriminador entre grupos de pacientes com maior ou menor risco, dados os escores SAPS3 do paciente.

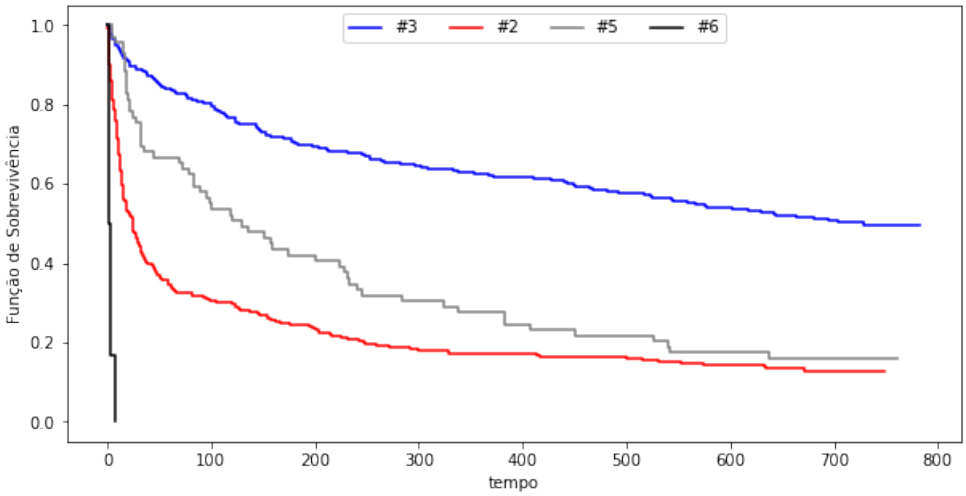

Figura 6.9: *Estimativa de Kaplan-Meier por folhas extraídas da árvore de sobrevivência com profundidade 2.*

Na Figura 6.10, é perceptível que ao ampliar a profundidade para 3, incorporam-se duas novas variáveis: se o tipo de quimioterapia foi paliativa (tipo_quimioterapia_Paliativa) e a capacidade funcional ECoG (capac_func_ecog). Além disso, ocorre um aumento no número de folhas, passando de 4 para 8.

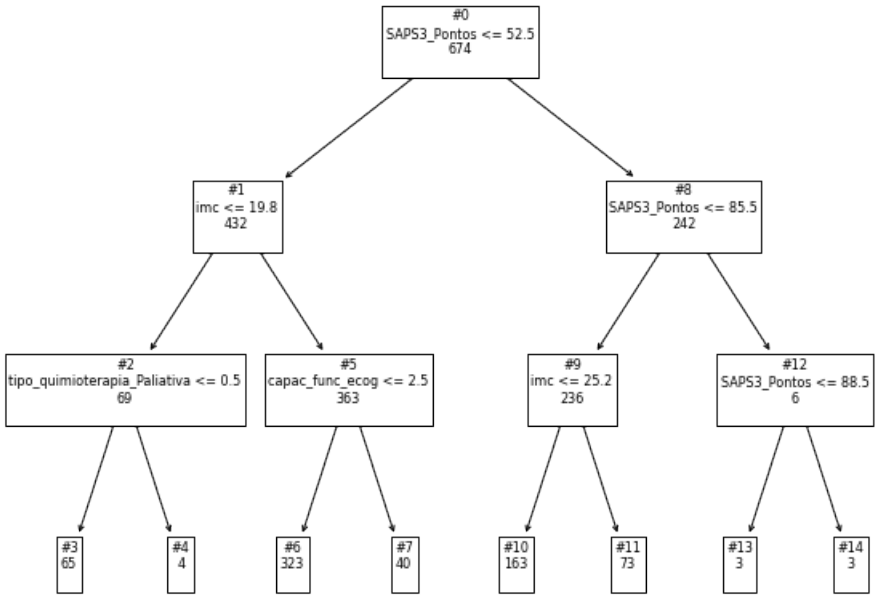

Figura 6.10: *Árvore de sobrevivência ajustada aos dados ICESP com profundidade 3.*

Na Figura 6.11, são apresentadas as estimativas de Kaplan Meier para as duas folhas com os maiores riscos e as duas folhas com os menores riscos. Destaca-se que a folha 6

compartilha as mesmas condições de IMC e escores SAPS3 da folha 3, conforme indicado na árvore da Figura 6.8, mas inclui a condição adicional de a função ECoG ser menor ou igual a 2,5. Além disso, é relevante notar que, na folha 14, ou seja, quando os pacientes têm escores SAPS3 acima de 88,5, são observadas as menores probabilidades de sobrevivência, com um tempo mediano de 3 dias.

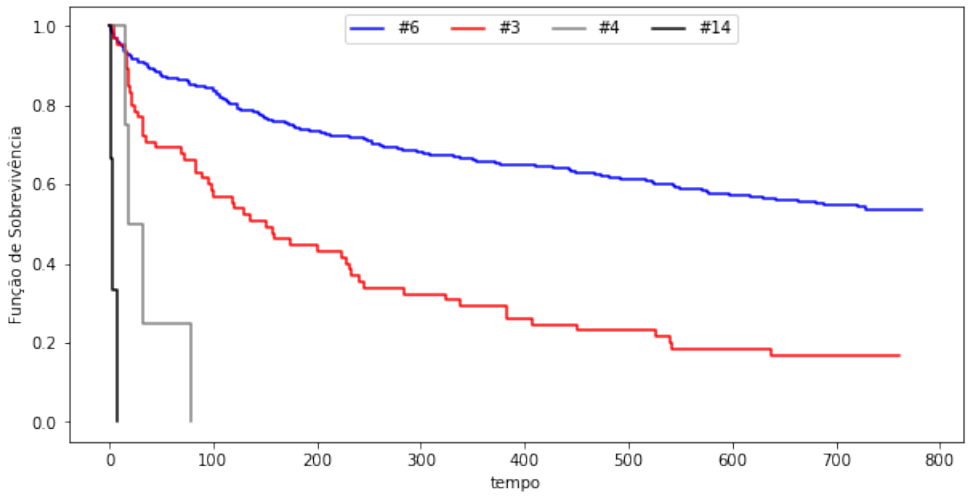

Figura 6.11: *Estimativa de Kaplan-Meier por folhas extraídas da árvore de sobrevivência com profundidade 3.*

A profundidade é um parâmetro de suma importância no processo de desenvolvimento de uma árvore de decisão, contudo, sua sensibilidade aumenta à medida que a profundidade é ampliada. Quanto maior for a profundidade, maior a redução no número de observações nas folhas no conjunto de treinamento. Por exemplo, na Figura 6.10, as folhas 4, 13 e 14 foram construídas com até 4 observações. A presença de folhas contendo um número reduzido de observações está diretamente ligada ao fenômeno de sobreajuste. Em outras palavras, a árvore demonstra um desempenho preditivo eficaz no conjunto de treinamento, porém essa capacidade não se reflete de maneira correspondente no conjunto de teste.

Considere a abordagem de *hold-out* para a separação dos dados do ICESP em conjuntos de treino e teste. Ao aleatorizar 85% para o conjunto de treinamento e 15% para o conjunto de teste, é possível analisar como o comportamento do índice C varia com o aumento da profundidade. Na Figura 6.12, observa-se que, à medida que a profundidade aumenta, o índice C melhora no conjunto de treinamento, mas piora no conjunto de teste.

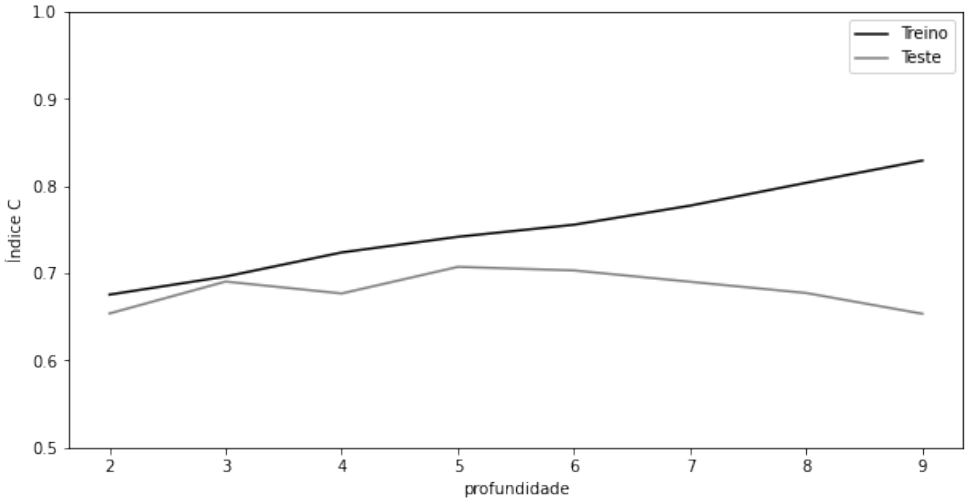

Figura 6.12: *Desempenho do índice C nos conjuntos de treino e teste para diferentes profundidades.*

Uma maneira de lidar com o problema da profundidade consiste em estabelecer um número mínimo de indivíduos nas folhas. Isso significa que é possível regular a complexidade do modelo, influenciando diretamente sua capacidade de generalização e prevenindo problemas como sobreajuste. Essa consideração proporciona uma maneira de ajustar a profundidade da árvore de decisão de forma mais controlada e adaptada à natureza dos dados.

Ao considerar um número mínimo de 50 observações, a figura seguinte possibilita uma explicação detalhada do processo de construção da árvore sob essa condição. É importante destacar que limitar o número de observações nas folhas possibilita a obtenção de resultados preditivos mais robustos e com maior confiabilidade.

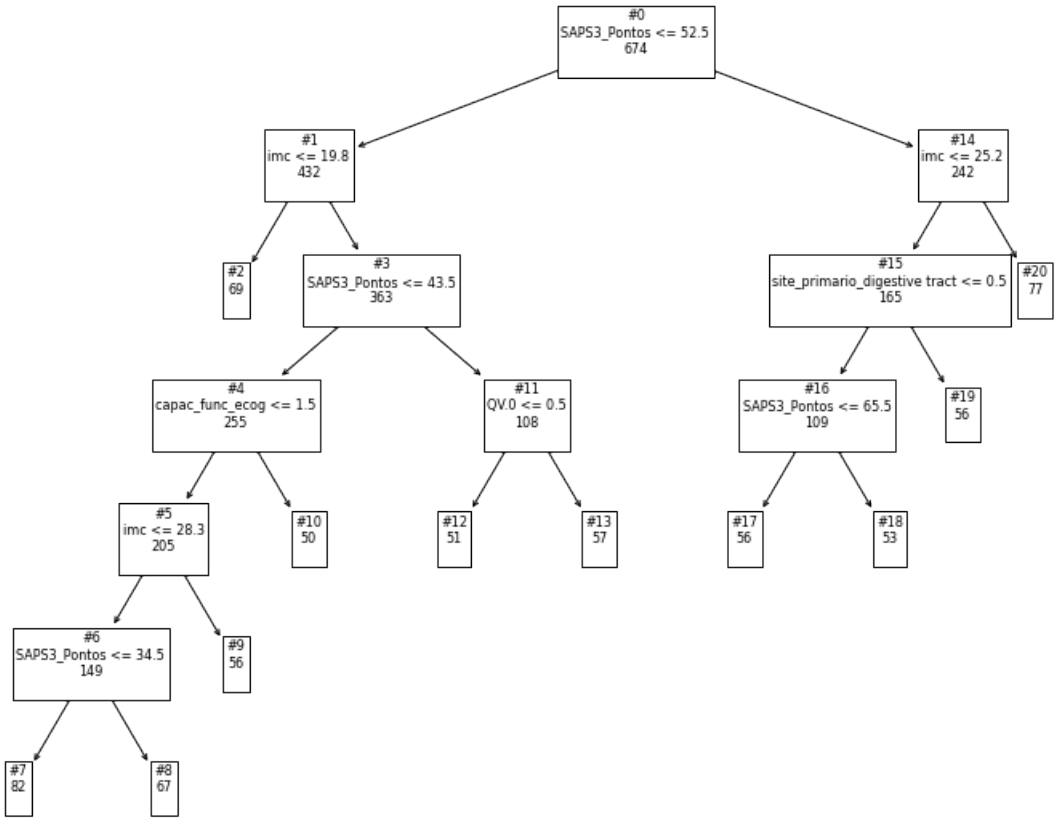

Figura 6.13: *Árvore de sobrevivência ajustada aos dados ICESP com número mínimo de amostras nas folhas igual a 50.*

De forma análoga à Figura 6.12, é possível ajustar diferentes números mínimos de observações nas folhas para avaliar os efeitos na performance preditiva nos conjuntos de treino e teste. Observa-se que, à medida que o número mínimo de observações diminui, a discrepância no índice C entre os conjuntos de treino e teste aumenta. No entanto, ao aumentar o número mínimo de observações, não parece haver alteração no índice C, cujos valores nos conjuntos de treino e teste se aproximam.

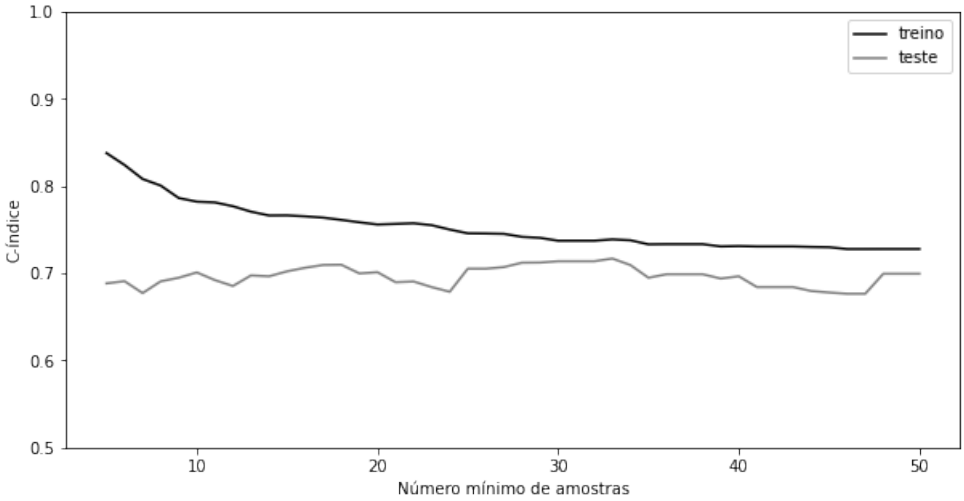

Figura 6.14: *Desempenho do índice C nos conjuntos de treino e teste para diferentes números mínimos de observações nas folhas.*

Ajustar uma árvore de sobrevivência com a configuração apropriada de número mínimo de observações nas folhas e profundidade é crucial para otimizar o desempenho do modelo. O número mínimo de observações nas folhas influencia diretamente a granularidade da divisão, determinando quão específicas ou generalizadas as condições de decisão serão. Ao definir um número mínimo adequado, evitamos subdivisões excessivamente específicas que podem levar a um sobreajuste, garantindo, assim, uma melhor capacidade de generalização do modelo para novos dados.

Quanto à profundidade da árvore, é uma medida crítica para controlar a complexidade do modelo. Uma árvore muito profunda pode resultar em uma representação excessivamente complexa dos dados de treinamento, o que, novamente, pode levar a um sobreajuste e prejudicar a capacidade do modelo de generalizar novos dados. Ao limitar a profundidade, buscamos um equilíbrio entre a capacidade de capturar padrões nos dados de treinamento e a capacidade de generalização para situações não vistas anteriormente.

Portanto, ajustar o número mínimo de observações nas folhas e a profundidade da árvore é uma estratégia essencial para desenvolver um modelo de árvore de sobrevivência capaz de realizar previsões precisas e robustas em diversos conjuntos de dados.

Na Figura 6.15, é apresentado um mapa de calor que reflete as seguintes informações: quanto mais intensa a cor, maior é a discrepância do índice C entre os conjuntos de treino e teste, resultando em um modelo sobreajustado; quanto mais clara a cor, menor a diferença, reduzindo o efeito de sobreajuste. Os valores em cada quadrado indicam o índice C no conjunto de teste. Concluímos que a configuração com profundidade igual a 6 e número mínimo de observações igual a 30 proporciona um equilíbrio ótimo entre viés e variância. Essa combinação resulta em um aumento significativo no desempenho do índice C no conjunto de teste, enquanto minimiza a disparidade entre os conjuntos de treino e teste. Em resumo, essa combinação de parâmetros destaca-se por conseguir equilibrar de forma eficaz a precisão do modelo com sua habilidade de generalização.

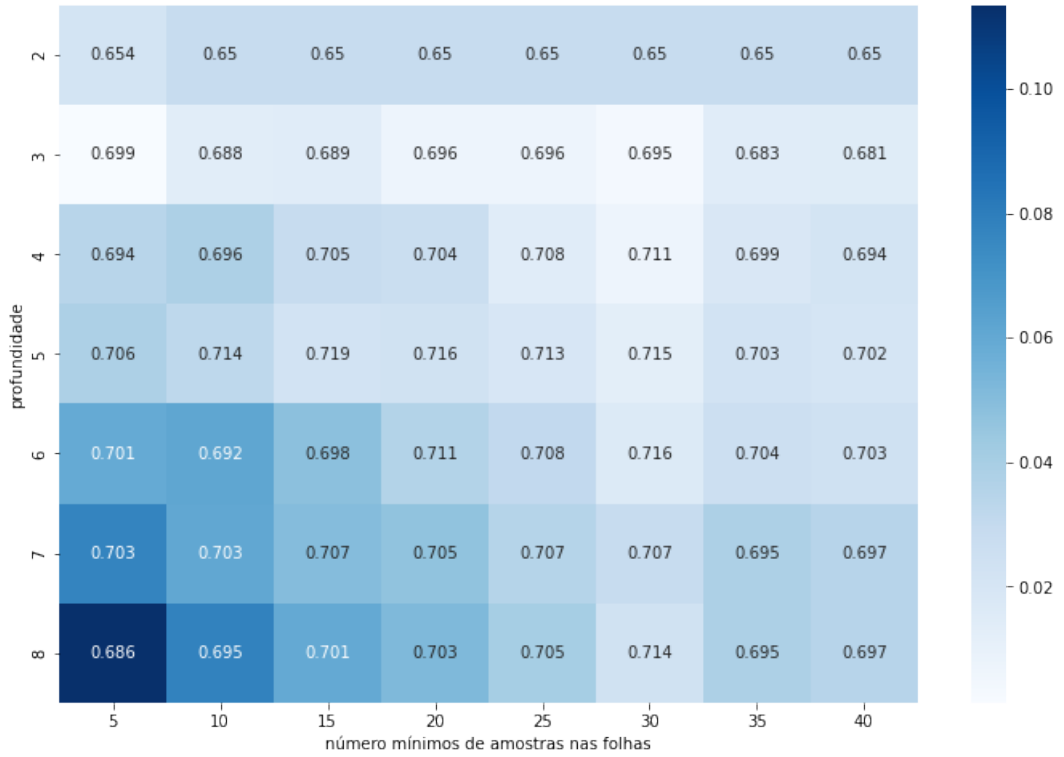

Figura 6.15: *Desempenho do índice C nos conjuntos de treino e teste para diferentes números mínimos de observações nas folhas e profundidades.*

6.3 Resultados dos Modelos

Para efetuar o treinamento dos modelos apresentados, são empregadas as técnicas de validação cruzada descritas no Capítulo 3. A ideia consiste em escolher os melhores hiperparâmetros para as técnicas de árvores utilizando a técnica de validação cruzada em k etapas (k-*fold*). Após obter os resultados mais promissores, é possível realizar a comparação com o modelo de Cox em um conjunto de teste estabelecido pela técnica *Hold-out*.

O procedimento consiste nas seguintes etapas:

- Utilizando a técnica *Hold-Out*, 85% dos dados do ICESP são aleatoriamente separados para compor um conjunto de treinamento, enquanto os 15% restantes formam um conjunto de teste.
- No conjunto de treinamento, é aplicada a técnica k-*fold* com $k = 10$ para identificar os melhores hiperparâmetros da árvore de sobrevivência, da floresta de sobrevivência aleatória e *Gradient Boosting Survival Tree*. Em seguida, são ajustados novamente os modelos no conjunto de treinamento utilizando os hiperparâmetros que apresentaram melhores resultados.
- Utilizando o conjunto de teste, são comparadas as técnicas de árvores com o modelo de Cox.

Os dados faltantes apresentados na Seção 6.1 foram excluídos dessas análises.

6.3.1 Árvore de Sobrevivência

Para ajustar a árvore de sobrevivência nos dados ICESP, são estabelecidos os seguintes hiperparâmetros:

- Profundidade máxima P_i em que é escolhido um valor mínimo de 2 e um máximo de 10, resultando $|P| = 9$.
- Número mínimo de observações nas folhas N_i em que é definido um valor mínimo de 5 e um máximo de 50, obtendo $|N| = 46$.

Utilizando a técnica *k-fold* com $k = 10$, são ajustadas 4140 árvores de sobrevivência, pois cada ajuste da árvore com profundidade P_i e número mínimo de amostra nas folhas N_i é repetida 10 vezes. Para determinar os valores de P^* e N^* que proporcionam o melhor desempenho preditivo, é utilizada a métrica índice C, conforme apresentado na Sub-seção 3.3.1. A escolha do índice C é motivada pela sua invariância a t, diferente do *Brier Score* apresentado na Sub-seção 3.3.2.

Na Tabela 6.3, destacam-se as 5 melhores média do índice C, juntamente com seus correspondentes valores P e N. Observa-se que a melhor árvore apresenta profundidade máxima $P = 6$ e número mínimo de observações $N = 30$, corroborando com os mesmos valores identificados na Figura 6.15. Além disso, nota-se um empate quando N é fixo em 30 e P varia de 8 à 10. Essa situação ocorre devido a possibilidade N excluir algumas folhas e, consequentemente, reduzir o tamanho da árvore (ou seja, sua profundidade).

Tabela 6.3: *Top 5 melhores médias do índice C usando a árvore de sobrevivência.*

P	N	Média C	Desvio Padrão C
6	30	0,6973	0,048
10	30	0,6964	0,046
8	30	0,6964	0,046
9	30	0,6964	0,046
7	30	0,6961	0,045

Após identificar os melhores valores de P^* e N^*, o modelo é reajustado no conjunto de treinamento, permitindo a observação das regras desenvolvidas. Na figura a seguir, apresenta-se a árvore de sobrevivência ajustada com seus nós e folhas. Nota-se que em cada nó há uma condição e o valor absoluto do teste log-rank, enquanto em cada folha é fornecido o número de observações. As covariáveis selecionadas foram:

- SAPS3_Pontos é a variável de maior influência na construção da árvore.
- imc, que condiz com índice de massa corporal do paciente.
- site_primario, que representa o local da neoplasia.
- capac_func_ecog, que significa a capacidade de função ECoG.
- idade, que é a idade em anos do paciente.
- QV.0, que representa o escore sobre a qualidade de vida dos pacientes.

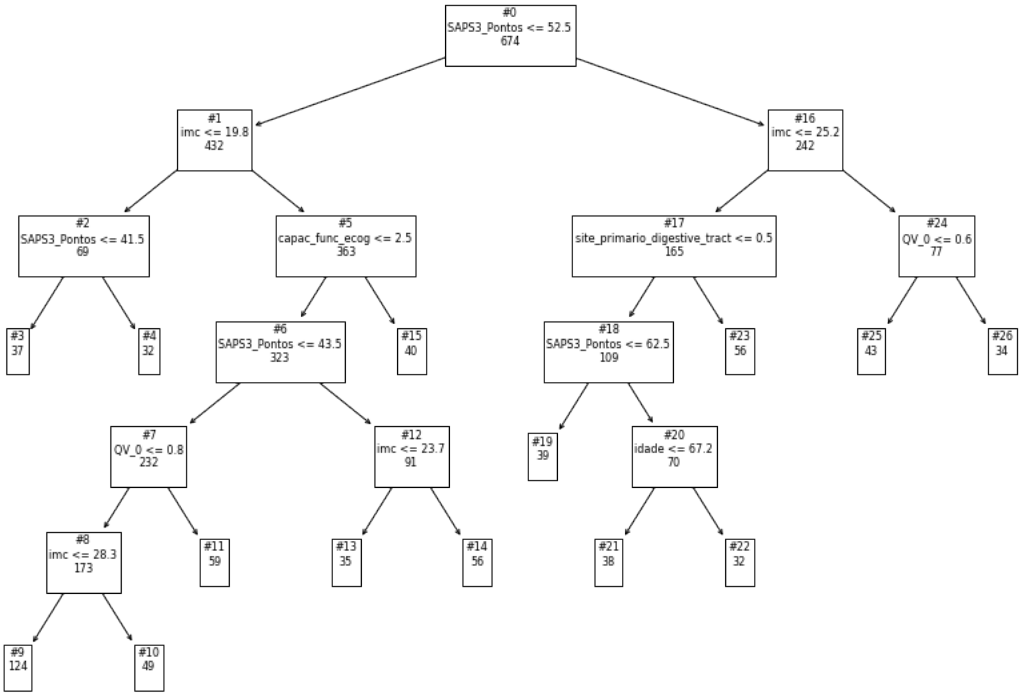

Figura 6.16: *Árvore de sobrevivência ajustada em todo conjunto de treinamento com* $P = 6$ *e* $N = 30$.

Observa-se que os pacientes foram distribuídos em 14 grupos distintos, correspondendo ao número de folhas na árvore ajustada. Assim, torna-se viável avaliar o risco associado a cada grupo utilizando a função de sobrevivência ou a taxa acumulada de risco. A utilização do estimador de Kaplan-Meier no conjunto de treinamento permite a visualização das duas folhas com as maiores probabilidades de sobrevivência e das duas folhas com as menores probabilidades, conforme ilustrada na figura abaixo. Algumas interpretações estão descritas a seguir:

- A folha 11 identifica o grupo de pacientes com o menor risco de óbito, indicando que esses indivíduos apresentam probabilidades de sobrevivência superiores a 0,8. Esses pacientes se caracterizam por escore SAPS3 abaixo de 43,5, IMC acima de 19,8, capacidade funcional de ECoG abaixo de 2,5 e escore de qualidade de vida acima de 0,8.
- A folha 9 identifica o grupo de pacientes com segundo menor risco de óbito. Esses indivíduos evidenciam uma redução mais acentuada nas probabilidades de sobrevivência em comparação com a folha 11, embora o tempo mediano seja de aproximadamente 600 dias. São pacientes com escore SAPS3 abaixo de 43,5, capacidade funcional ECoG abaixo de 2,5, IMC entre 19,8 a 28,3 e escore de qualidade de vida abaixo de 0,8.
- A folha 22 caracteriza o grupo de pacientes com segundo maior risco de óbito. Estes pacientes são identificados por escore SAPS3 superior a 52,5, IMC abaixo de 25,2, ausência de neoplasia no trato digestivo e idade acima de 67,2 anos. Importante notar que, no máximo, esses pacientes sobrevivem até aproximadamente 500 dias.
- A folha 14 apresenta o grupo de pacientes com o maior risco de óbito. Esses pacientes

caracterizam-se por escore SAPS3 superior a 43,5, índice de massa corporal acima de 23,7 e capacidade funcional de ECoG abaixo de 2,5. O tempo mediano é de aproximadamente 3 anos.

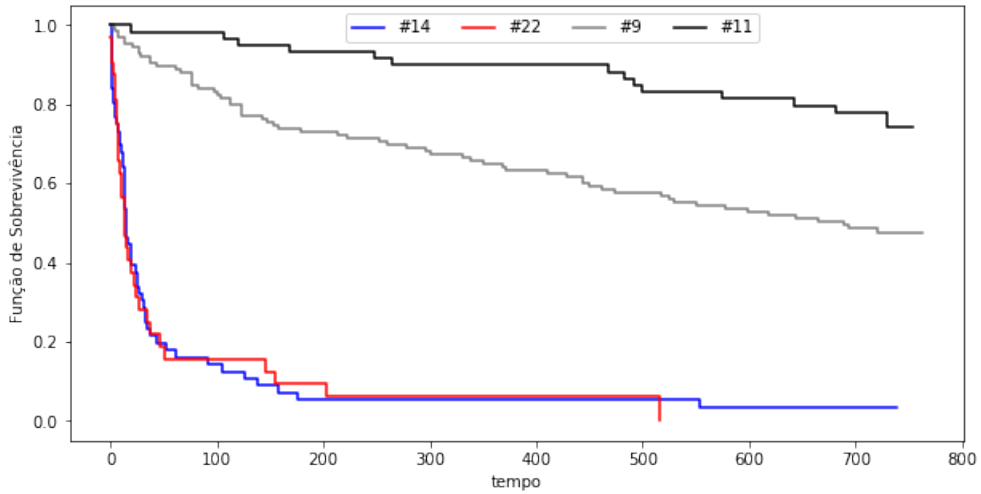

Figura 6.17: *Função de sobrevivência para diferentes folhas da árvore.*

6.3.2 Floresta de Sobrevivência Aleatória

Para o ajuste da floresta de sobrevivência aleatória, são estabelecidos os seguintes hiperparâmetros:

- Profundidade máxima P_i em que é escolhido um valor mínimo de 2 e um máximo de 10.
- Número mínimo de observações nas folhas N_i, definido um valor mínimo de 5 e um máximo de 50.
- Número de árvores B_i com um valor mínimo de 50 e um máximo de 200.
- Número máximo de covariáveis M_i em que é dado um valor mínimo de 5 e um máximo de 20.

Assim como descrito na Subseção 6.3.1, é aplicada a técnica *k-fold* com $k = 10$, avaliando cada etapa com o índice C. Os resultados sumarizados para os cinco melhores conjuntos de hiperparâmetros são apresentados na tabela a seguir. É importante ressaltar que os hiperparâmetros mais eficazes são $P^* = 7$, $N^* = 5$, $B^* = 50$ e $M^* = 5$, proporcionado um índice C de $0,7374$.

Tabela 6.4: *Top 5 melhores médias do índice C usando o algoritmo RSF.*

P	N	M	B	Média C	Desvio Padrão C
7	5	5	50	0,7374	0,029
7	10	5	150	0,7355	0,035
7	15	5	150	0,7352	0,032
7	5	5	150	0,7347	0,030
7	5	10	150	0,7339	0,031

Ajustando a floresta de sobrevivência aleatória em todo o conjunto de treinamento com os melhores hiperparâmetros, observa-se a distribuição da média do número total de eventos, conforme especificado na Seção 5.4. A escolha da média do número total de eventos se deve pelo fato de ser uma estimativa invariante ao tempo t. Na Figura 6.18, nota-se que a distribuição é assimétrica à direita, indicando que à medida que o valor estimado do número total de eventos aumenta, a quantidade de observações diminui.

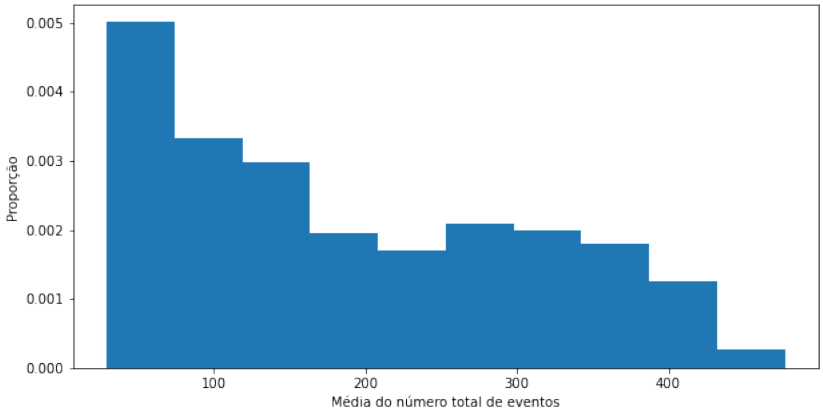

Figura 6.18: *Distribuição da média do número total de eventos no conjunto de treinamento.*

Dessa forma, torna-se viável a formação de grupos de pacientes com base nos quartis da média do número total de eventos, possibilitando a avaliação do comportamento da função de sobrevivência entre esses grupos. O propósito é analisar se existe uma ordenação em relação à sobrevida dos pacientes e identificar decisões distintas para cada grupo. A aplicação dos 1º, 2º e 3º quartis resulta no cálculo da função de sobrevivência para cada grupo, utilizando o estimador de Kaplan-Meier, conforme demonstrado na Figura 6.11.

Destaca-se que, à medida que o intervalo do número total de eventos aumenta, observa-se uma diminuição na sobrevida dos pacientes e um declínio significativo na probabilidade de sobrevivência. Por exemplo, uma decisão estratégica poderia envolver a priorização do atendimento médico para o grupo de pacientes abrangidos no intervalo $[290,53;\infty)$, uma vez que esses representam os indivíduos com maiores riscos de óbitos.

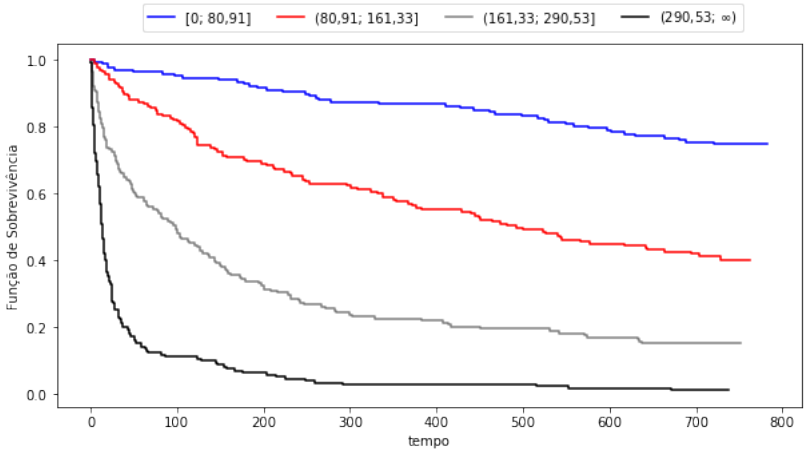

Figura 6.19: *Função de sobrevivência por grupos definidos pelos quartis do número total de eventos.*

6.3.3 Gradient Boosting Survival Tree

No processo de treinamento do *Gradient Boosting Survival Tree*, são selecionados os hiperparâmetros que influenciam significativamente o desempenho do modelo. Os seguintes parâmetros são considerados:

- Profundidade máxima (P_i), cujo valor é restrito a um intervalo entre 2 e 10, garantindo uma abordagem equilibrada entre complexidade e generalização do modelo.
- Número mínimo de observações nas folhas (N_i), variando de um mínimo de 5 a um máximo de 50, proporcionando flexibilidade na sensibilidade do modelo às variações nos dados.
- Número de árvores (B_i), estabelecido com um mínimo de 50 e um máximo de 200, permitindo uma ampla gama de possibilidades para capturar padrões complexos nos dados.

A abordagem de validação cruzada utilizando *k-fold* com $k = 10$ e o índice C é empregada para avaliar o desempenho do modelo em diferentes conjuntos de dados. Os resultados obtidos, apresentados na Tabela 6.5, revelam que os hiperparâmetros ótimos para maximizar a eficácia do modelo são $P^* = 2$, $N^* = 5$ e $B^* = 100$. Essa configuração proporciona um equilíbrio entre a complexidade do modelo e sua capacidade de generalização, resultando em um desempenho robusto na realização de predições.

Tabela 6.5: *Top 5 melhores médias do índice C usando o* Gradient Survival Boosting.

P	N	B	Média C	Desvio Padrão C
2	5	100	0,7313	0,033
7	45	50	0,7309	0,031
2	15	100	0,7284	0,036
2	5	50	0,7273	0,035
2	15	50	0,7267	0,033

É possível realizar predições utilizando o logaritmo da função de taxa de risco que é uma estimativa invariante ao tempo t, conforme detalhado na Seção 5.5. A Figura 6.20 apresenta a distribuição dessa função, revelando que 50% dos pacientes possuem previsões abaixo de -0,07, sugerindo que o modelo prediz que metade dos pacientes está inclinada a um menor risco.

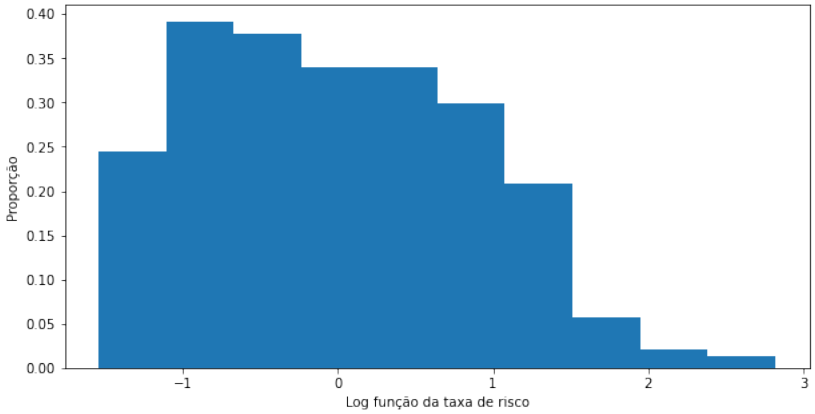

Figura 6.20: *Distribuição do logaritmo da função de taxa de risco dada pelo Gradient Survival Boosting.*

Da mesma forma que é abordada na subseção 6.2.1, pode-se formar grupos de pacientes utilizando o logaritmo da função de taxa de risco e analisar o comportamento da sobrevivência entre esses grupos. Na figura 6.21, vale ressaltar que à medida que o intervalo do logaritmo da função de taxa de risco aumenta, a sobrevida dos pacientes diminui, acompanhada por uma queda mais acentuada na probabilidade de sobrevivência.

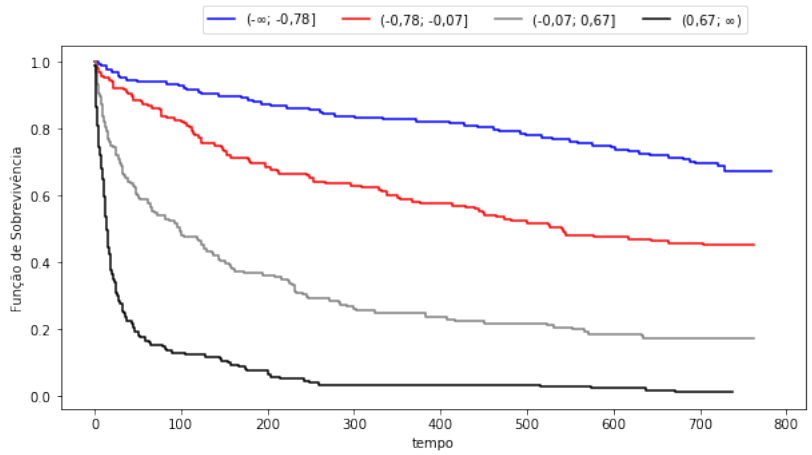

Figura 6.21: *Função de sobrevivência por grupos definidos pelos quartis das predições.*

6.4 Comparação do Desempenho Preditivo

Após a determinação dos melhores hiperparâmetros, os métodos baseados em árvores são recalibrados em todo o conjunto de treinamento. Ao ajustar o modelo de Cox a esse

conjunto, é possível comparar os resultados com os métodos apresentados, utilizando o índice C no conjunto de teste especificado na Seção 6.3. Na Tabela 6.6, é perceptível que a floresta de sobrevivência aleatória alcança o índice C mais elevado, embora com uma diferença de aproximadamente 2 pontos em relação ao modelo de riscos proporcionais de Cox. Ademais, a árvore de sobrevivência registra o menor índice C.

Tabela 6.6: *Comparação dos modelos usando o índice C.*

Modelo	Índice C
Modelo de Cox	0,729
Árvore de Sobrevivência	0,716
Floresta de Sobrevivência Aleatória	**0,756**
Gradient Boosting Survival Tree	0,752

É possível realizar uma análise do *Brier Score* em diferentes instantes de tempo t para cada modelo. Assim como no caso do índice C, a árvore de sobrevivência demonstra um desempenho inferior em comparação com os outros modelos. A floresta de sobrevivência aleatória destaca-se ao apresentar os melhores resultados, embora haja um empate com o modelo de Cox quando o tempo está na faixa de 100 a 250, e um empate em todos os instantes de tempos com o *Gradient Boosting Survival Tree*. Vale mencionar que, conforme estabelecido por Harrell (2001), quando os modelos apresentam valores inferiores a 0,25 no *Brier Score*, considera-se que há uma boa performance preditiva. Nesse sentido, todos os modelos analisados cumprem essa condição, sugerindo que cada um deles pode ser escolhido de forma viável para a predição da sobrevivência dos pacientes.

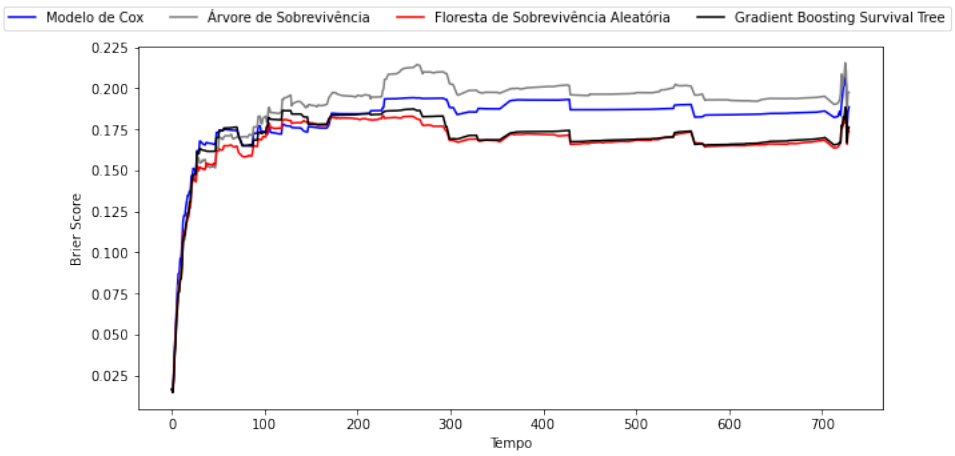

Figura 6.22: *Brier Score para diferentes tempos t e modelos.*

É possível gerar a curva de aprendizado para cada método de árvore. A curva de aprendizado é uma representação visual da relação entre o desempenho do modelo e o tamanho do conjunto de dados. Ao testar diferentes tamanhos amostrais, essa curva permite avaliar como o modelo se beneficia do aumento da quantidade de dados.

Inicialmente, em conjuntos de treinamento pequenos, o modelo pode apresentar sobre-ajuste ou subajuste devido à falta de dados para aprender padrões complexos. Conforme o

tamanho amostral aumenta, a curva de aprendizado geralmente mostra uma melhora no desempenho, indicando que o modelo está capturando melhor a relação entre variáveis e se tornando mais preciso.

No entanto, é importante observar que, em algum ponto, o benefício do aumento do tamanho amostral pode diminuir, indicando que o modelo atingiu seu potencial máximo de aprendizado com os dados disponíveis. Identificar esse ponto ótimo é crucial para evitar a coleta de dados desnecessários e otimizar o processo de treinamento do modelo.

Na Figura 6.23, pode-se avaliar a curva de aprendizado utilizando o índice C como métrica de avaliação para cada método, empregando a técnica *k-fold* com $k = 10$. Nesse contexto, calcula-se a média do índice C nos conjuntos de treinamento e teste.

Observa-se na árvore de sobrevivência que, após atingir 400 observações, há um aumento no índice C no conjunto de teste, acompanhado por uma redução na diferença entre os conjuntos de treinamento e teste. Para o *Gradient Boosting Survival Tree*, à medida que o número de observações aumenta, nota-se uma maior convergência entre os índices C do conjunto de treinamento e teste. No entanto, na floresta de sobrevivência aleatória, o índice C aparenta manter-se constante para diferentes tamanhos amostrais, com uma notável disparidade entre os conjuntos de treinamento e teste.

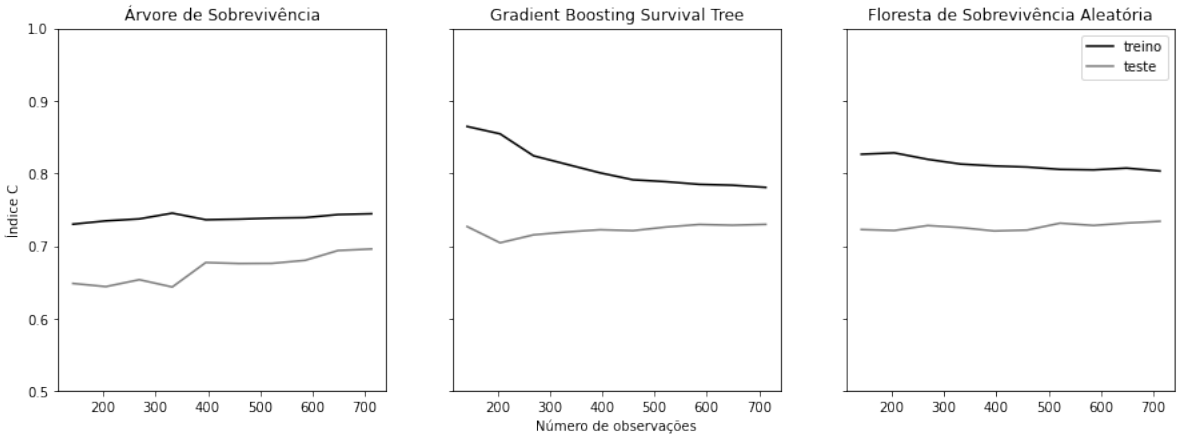

Figura 6.23: *Curva de aprendizado utilizando o índice C para diferentes métodos de árvores.*

6.5 Avaliação do Efeito das Observações Censuradas

Em análise de sobrevivência, a simulação de censuras desempenha um papel fundamental, uma vez que permite avaliar e aprimorar modelos de maneira mais robusta. Com isso, pode-se introduzir diferentes cenários e níveis de censura artificialmente nos conjuntos de dados. Isso possibilita testar a resistência dos diferentes modelos de análise de sobrevivência sob diversas condições, avaliando sua capacidade de lidar com informações truncadas e eventos não observados. Essa abordagem é crucial para validar a robustez e a generalização dos modelos em situações do mundo real, onde dados censurados são comuns.

Nesta seção, observações censuradas são adicionadas artificialmente aos dados do ICESP. Posteriormente, os modelos discutidos na Seção 6.4 são treinados por meio da técnica de *Hold-out*. Em resumo, 85% dos dados do ICESP são aleatoriamente separados para formar um conjunto de treinamento, enquanto os 15% restantes constituem um conjunto de teste. O propósito é ajustar cada método baseado em árvore no conjunto de treinamento e examinar o comportamento do índice C tanto nos conjuntos de treinamento quanto nos de teste, com o intuito de compreender possíveis situações de sobreajuste e subajuste.

Na figura a seguir, são exibidos os resultados da simulação, na qual a linha tracejada representa a proporção real de censura, fixada em $0,34$. É perceptível que, à medida que a proporção de censura aumenta, todos os modelos tendem a exibir uma crescente tendência de sobreajuste. É interessante notar que a floresta de sobrevivência aleatória se revelou particularmente sensível às variações na censura, manifestando maiores disparidades entre os conjuntos de treinamento e teste.

Por outro lado, o *Gradient Boosting Survival Tree* se destacou ao manter uma performance superior no conjunto de teste, mesmo diante do aumento da censura e com um sobreajuste considerável. Esse método mostrou-se mais estável perante as variações na censura, sugerindo uma capacidade mais robusta de generalização.

A diferença de comportamento entre esses modelos frente às variações na censura destaca a importância de selecionar a abordagem mais apropriada, levando em consideração o contexto específico e as condições dos dados. Essa análise reforça a necessidade de avaliar a sensibilidade dos modelos a diferentes cenários de censura ao escolher a melhor estratégia para análises preditivas em sobrevivência.

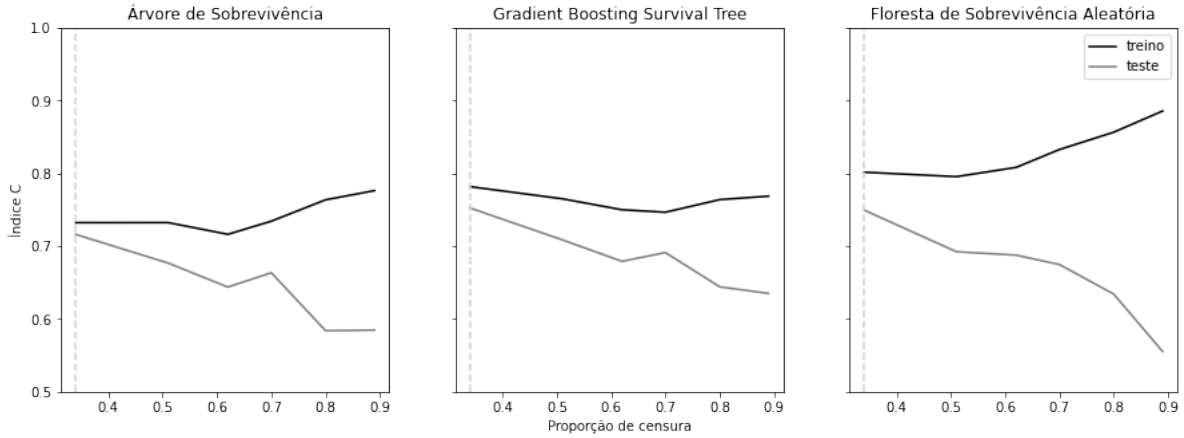

Figura 6.24: *Desempenho do índice C para diferentes modelos variando artificialmente a proporção de censura.*

Capítulo 7

Conclusão

Para se construir predições em Análise de Sobrevivência, o modelo de riscos proporcionais de Cox é uma opção bastante interessante uma vez que não há suposições no modelo além da proporcionalidade dos riscos. Entretanto, é esperado que as covariáveis e o logaritmo da taxa de risco sejam linearmente correlacionadas. Como as covariáveis podem apresentar relações não-lineares, as árvores de decisão se tornam uma possibilidade para construção de modelos preditivos em dados de sobrevivência.

Dessa forma, neste trabalho foram revisados os métodos de árvores de decisão, apresentando como são construídos em problemas de regressão, estudando como são feitas as divisões usando o método de mínimos quadrados e quais são as técnicas para que se possa criar as folhas. Além disso, foram explorados os algoritmos mais famosos de *ensembles*: Floresta Aleatória e *Gradient Boosting Machine*.

Foram estudadas as árvores de sobrevivência, mostrando um novo conceito de divisão dos dados e, assim, apresentando as estatísticas de comparação de duas populações como possíveis medidas. Também foram considerados os algoritmos de *ensemble* denominados como Florestas de Sobrevivência Aleatórias e *Gradient Boosting Survival Tree*.

Por fim, foram aplicadas as diferentes técnicas de árvores de sobrevivência no conjunto de dados ICESP e comparadas com o modelo de Cox. Os dados ICESP fornecem uma rica fonte para essa análise, permitindo a criação de modelos preditivos personalizados. Ao explorar os diferentes métodos de árvores de decisão, os profissionais de saúde podem ganhar insumos sobre quais pacientes têm maior probabilidade da sobrevivência, orientando assim as decisões de tratamento e fornecendo uma abordagem mais personalizada e eficaz.

Usando o índice C, foi observado que a Floresta de Sobrevivência Aleatória apresentou um melhor desempenho preditivo para identificar os pacientes com maiores riscos de óbitos, em comparação ao modelo de Cox e Árvore de Sobrevivência e o *Gradient Boosting Survival Tree*. Entretanto, quando é avaliado o Brier Score, observa-se que há um empate entre o *Gradient Boosting Survival Tree* e a Floresta de Sobrevivência Aleatória.

Como proposta para trabalhos futuros, sugere-se:

- Aplicar outros métodos de otimização de hiperparâmetros.

- Estudar a importância das covariáveis nos métodos de árvores de sobrevivência.
- Propor outras funções de perdas para o algoritmo do *boosting*, além da função de verossimilhança parcial de Cox.
- Pesquisar outros critérios para divisão de dados em árvores de sobrevivência.

Apêndice A

Dicionário de dados ICESP

Tem-se o dicionário do conjunto de dados disponibilizado. No total, apresenta-se 793 observações e 27 variáveis:

- Uma variável sendo o identificador do paciente (idpac).
- 24 variáveis que indicam diversas características dos pacientes.
- A variável tempo que indica a variável resposta.
- A variável delta que é o indicador de censura.

Sobre a variável resposta, um ponto de observação é que neste estudo a censura é a direita, isto é, existe a falta da informação quando o evento não é observado até o término do estudo.

Tabela A.1: *Descrição das colunas do conjunto de dados ICESP*

Variável	Descrição
idpac	identificador do paciente
idade	idade do paciente
sexo	sexo do paciente
imc	IMC do paciente
insuf_renal	Se o paciente tem insuficiência renal.
insuf_resp	Se o paciente tem insuficiência respiratória.
cirrose	Se o paciente tem o diagnóstico de cirrose.
alcoolismo	Se o paciente tem o diagnóstico de alcoolismo.
diabetesSN	Se o paciente tem o diagnóstico de diabetes.
insuf_cardiaca	Se o paciente tem insuficiência cardíaca.
deliriumSN	Se o paciente teve o diagnóstico de delírio.
tipoadm	Tipo de admissão.
site_primario	Um agrupamento da neoplasia.
status_cancer	Status do câncer.
extensao_cancer	Extensão do câncer.
cirurgia	Se o paciente realizou a cirurgia.
tipo_cirurgia	Tipo de cirurgia
quimioterapia	Se o paciente realizou quimioterapia.
tipo_quiometerapia	Tipo de quimioterapia.
radioterapia	Se o paciente realizou radioterapia
tipo_radioterapia	Tipo de Radioterapia.
capac_func_ecog	Capacidade de função ECoG.
hospit_antes_uti	Quantidades de dias de hospitalização antes de ir para a UTI.
QV.0	Escore sobre qualidade de vida extraído de um questionário.
SAPS3_Pontos	Escore da propensão à morte na admissão à UTI extraído de um questionário.
tempo	Tempo em dias até ao óbito ou ao final do estudo.
delta	Se ocorreu o óbito (censura).

Apêndice B

Códigos

Todo o desenvolvimento da dissertação foi realizada no software Python. Os códigos podem ser encontrados nesse link do GitHub: https://github.com/vinisantosol/ime-mestrado/.

Referências

[Bazoukis *et al.* 2023] George Bazoukis, Sri Charan Bollepalli, Cheng Ting Chung *et al.* "Application of artificial intelligence in the diagnosis of sleep apnea". Em: *Journal of Clinical Sleep Medicine* 19.7 (2023), pgs. 1337–1363.

[Bou-Hamad e Larocque 2011] Imad Bou-Hamad e Denis Larocque. "Gradient boosting for survival data". Em: *Computational Statistics & Data Analysis* 55.3 (2011), pgs. 1479–1490.

[Bou-Hamad, Larocque e Ben-Ameur 2011] Imad Bou-Hamad, Denis Larocque e Hatem Ben-Ameur. "A review of survival trees". Em: *Statistics Surveys* 5 (2011), pgs. 44–71.

[Breiman 1996a] Leo Breiman. "Bagging predictors". Em: *Machine Learning* 24 (1996), pgs. 123–140.

[Breiman 1996b] Leo Breiman. "Stacked regressions". Em: *Machine Learning* 24 (1996), pgs. 49–64.

[Breiman 2001] Leo Breiman. "Random forests". Em: *Machine Learning* 45 (2001), pgs. 5–32.

[Breiman *et al.* 1984] Leo Breiman, Jerome Friedman, Charles J. Stone e Richard A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[Breslow 1972] Norman E. Breslow. "Discussion of the paper by d. r. cox". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (1972), pgs. 216–217.

[Breslow 1974] Norman E. Breslow. "Covariance analysis of censored survival data". Em: *Biometrics* 30 (1974), pgs. 89–99.

[Brier 1950] Glenn W. Brier. "Verification of forecasts expressed in terms of probability". Em: *Monthly Weather Review* 78.1 (1950), pgs. 1–3. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

[Colosimo e Giolo 2006] Enrico Antonio Colosimo e Suely Ruiz Giolo. *Análise de Sobrevivência Aplicada*. ABE-Projeto Fisher, Edgard Blücher, 2006.

[Cox 1972] D. R. Cox. "Regression models and life-tables". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* (1972), pgs. 187–220.

[Cox 1975] D. R. Cox. "Partial likelihood". Em: *Biometrika* 62 (1975), pgs. 269–276. DOI: 10.1093/biomet/62.2.269.

[Cox e Hinkley 1974] D. R. Cox e D. V. Hinkley. *Theoretical Statistics*. London: Chapman e Hall, 1974.

[Efron 1977] Bradley Efron. "The efficiency of Cox's likelihood function for censored data". Em: *Journal of the American Statistical Association* 72 (1977), pgs. 557–565.

[Efron 1979] Bradley Efron. "Bootstrap methods: another look at the jackknife". Em: *The Annals of Statistics* 7.1 (1979), pgs. 1–26.

[Friedman 1999] Jerome H. Friedman. *Greedy Function Approximation: a Gradient Boosting Machine*. Rel. técn. Department of Statistics, Stanford University, 1999.

[Friedman 2002] Jerome H. Friedman. "Stochastic gradient boosting". Em: *Computational Statistics & Data Analysis* 38 (2002), pgs. 367–378.

[Gehan 1965] E. A. Gehan. "A generalized wilcoxon test for comparing arbitrarily singly-censored samples". Em: *Biometrika* 52 (1965), pgs. 203–223. DOI: 10.2307/2333825. URL: https://doi.org/10.2307/2333825.

[Gill 1984] R. D. Gill. "Understanding Cox's regression model: a martingale approach". Em: *Journal of the American Statistical Association* 79 (1984), pgs. 441–447.

[Gordon e Olshen 1985] Leo Gordon e Richard A. Olshen. "Tree-structured survival Analysis". Em: *Cancer Treatment Reports* 69 (1985), pgs. 1065–1069.

[Grambsch e Therneau 1994] P. M. Grambsch e T. M. Therneau. "Proportional hazards tests and diagnostics based on weighted residuals". Em: *Biometrika* 81 (1994), pg. 515.

[Harrell 2001] Frank E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag New York, 2001.

[Harrell *et al.* 1982] Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee e Robert A. Rosati. "Evaluating the yield of medical tests". Em: *Journal of the American Medical Association* 247 (1982), pgs. 2543–2546.

[Hothorn *et al.* 2006] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro e Mark J. van der Laan. "Survival ensembles". Em: *Biostatistics* 7.3 (2006), pgs. 355–373.

[Hothorn e Lausen 2003] Torsten Hothorn e Berthold Lausen. "On the exact distribution of maximally selected rank statistics". Em: *Computational Statistics & Data Analysis* 43 (2003), pgs. 121–137.

[Ishwaran e Kogalur 2007] Hemant Ishwaran e Udaya B. Kogalur. "Random survival forests for R". Em: *R News* 7 (2007), pgs. 25–31.

[Ishwaran *et al.* 2008] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone e Michael S. Lauer. "Random survival forests". Em: *Annals of Statistics* 2 (2008), pgs. 841–860.

[Izbicki e Santos 2020] Rafael Izbicki e Tiago Mendonça dos Santos. *Aprendizado de máquina: uma abordagem estatística*. 2020. 268 pp. isbn: 978-65-00-02410-4.

[James *et al.* 2013] Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013. isbn: 978-1461471370.

[Kaplan e Meier 1958] Edward L. Kaplan e Paul Meier. "Nonparametric estimation from incomplete observations". Em: *Journal of the American Statistical Association* 53 (1958), pgs. 457–481.

[Klein e Moeschberger 2003] John P. Klein e Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Verlag, 2003.

[LeBlanc e Crowley 1993] Michael LeBlanc e John Crowley. "Survival trees by goodness of split". Em: *Journal of the American Statistical Association* 88 (1993), pgs. 457–467.

[Lee e Rosati 1982] K. L. Lee e R. A. Rosati. "Evaluating the yield of medical tests". Em: *JAMA* 247 (1982), pg. 2543.

[Löschmann e Smorodina 2020] Lennart Löschmann e Daria Smorodina. *Deep Learning for Survival Analysis*. Humboldt-Universität zu Berlin. Viewed 8 August 2020. Disponível em https://humboldt-wi.github.io/blog/research/information_systems_1920/group2_survivalanalysis/. 2020.

[Mantel 1966] Nathan Mantel. "Evaluations of survival data and two new rank order statistics arising in its consideration". Em: *Cancer Chemotherapy Reports* 50 (1966), pgs. 163–170.

[Morgan e Sonquist 1963] J. Morgan e J. Sonquist. "Problems in the analysis of survey data and a proposal". Em: *Journal of the American Statistical Association* 58 (1963), pgs. 415–434.

[Naftel *et al.* 1985] David Naftel, Eugene H. Blackstone e Mark Turner. "Conservation of events". 1985.

[Nelson 1969] Wayne Nelson. "Hazard plotting for incomplete failure data". Em: *Journal of Quality Technology* 1.1 (1969), pgs. 27–52.

[Normilio-Silva *et al.* 2016] Karina Normilio-Silva *et al.* "Long-term survival, quality of life, and quality-adjusted survival in critically ill patients with cancer". Em: *Crit Care Med* 44 (2016), pgs. 1327–1337.

[Obuda 2016] Felix Yala Obuda. "Analysis of Credit Risk on Bank Loans Using Cox's Proportional Hazards Model". Tese de dout. University of Nairobi, 2016.

[Oehlert 1992] Gary Oehlert. "A note on the delta method". Em: *American Statistician* (1992), pgs. 27–29.

[R. Peto e J. Peto 1972] Richard Peto e Juliet Peto. "Asymptotically efficient rank invariant test procedures (with discussion)". Em: *Journal of the Royal Statistical Society, Series A* 135 (1972), pgs. 185–206.

[Prentice e Gloeckler 1978] Ross L. Prentice e Louis A. Gloeckler. "Regression analysis of grouped survival data with application to breast cancer data". Em: *Biometrics* 34 (1978), pgs. 57–67. doi: 10.2307/2529588. url: https://doi.org/10.2307/2529588.

[Ridgeway 1999] Greg Ridgeway. "The state of boosting". Em: *Computing Science and Statistics* 31 (1999), pgs. 172–181.

[Schoenfeld 1982] David Schoenfeld. "Partial residuals for the proportional hazards regression model". Em: *Biometrika* 69 (1982), pg. 239.

[Segal 1988] Mark R. Segal. "Regression trees for censored data". Em: *Biometrics* 44 (1988), pgs. 35–47.

[Tarone e Ware 1977] Robert E. Tarone e James Ware. "On distribution-free tests for equality of survival distributions". Em: *Biometrika* 64 (1977), pgs. 156–160. doi: 10.1093/biomet/64.1.156. url: https://doi.org/10.1093/biomet/64.1.156.

[Therneau e Atkinson 2019] Terry Therneau e Bradley Atkinson. *An Introduction to Recursive Partitioning Using the RPART Routines*. Mayo Clinic, 2019.

[Therneau e Grambsch 2000] Terry Therneau e Patricia Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.

[Wang e Yao 2020] Menglan Wang e Shan Yao. "Gradient boosting machine for right-censored survival data". Em: *BMC Bioinformatics* 21.1 (2020), pgs. 1–18.

[Welch 1947] B. L. Welch. "The generalization of student's problem when several different population variances are involved". Em: *Biometrika* 34 (1947), pgs. 28–35.

[Wong 2011] K.K. Wong. "Using cox regression to model customer time to churn in the wireless telecommunications industry". Em: *Journal of Target Measurement Analysis in Marketing* 19 (2011), pgs. 37–43. doi: 10.1057/jt.2011.1.