UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATISTICA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

CLEVIA BENTO DE OLIVEIRA

ANÁLISE DE SOBREVIVÊNCIA E RANDOM SURVIVAL FOREST NA PREDIÇÃO DO TEMPO ATÉ A MORTE DE CÂNCER DE MAMA EM MULHERES DO ESTADO DE PERNAMBUCO

CAMPINA GRANDE -PB
2023

CLEVIA BENTO DE OLIVEIRA

ANÁLISE DE SOBREVIVÊNCIA E RANDOM SURVIVAL FOREST NA PREDIÇÃO DO TEMPO ATÉ A MORTE DE CÂNCER DE MAMA EM MULHERES DO ESTADO DE PERNAMBUCO

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

Área de concentração: Estatística

Orientador: Prof.Dr. Tiago Almeida de Oliveira.

CAMPINA GRANDE -PB
2023

O48a Oliveira, Clevia Bento de.

Análise de sobrevivência e *random survival forest* na predição do tempo até a morte de câncer de mama em mulheres do estado de Pernambuco [manuscrito] / Clevia Bento de Oliveira. - 2023.

26 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Coordenação do Curso de Estatística - CCT. "

1. Câncer de mama. 2. Análise de sobrevivência. 3. Aprendizagem de máquina. I. Título

21. ed. CDD 610.28

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

CLEVIA BENTO DE OLIVEIRA

ANÁLISE DE SOBREVIVÊNCIA E RANDOM SURVIVAL FOREST NA PREDIÇÃO DO TEMPO ATÉ A MORTE DE CÂNCER DE MAMA EM MULHERES DO ESTADO DE PERNAMBUCO

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

Área de concentração: Estatística

Aprovado em: 05/07/2023

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira (Orientador)
Universidade Estadual da Paraíba (UEPB)

Cleanderson Romualdo Fidelis

Prof. Me. Cleanderson Romualdo Fidelis
Universidade Estadual da Paraíba (UEPB)

Milena Edite Casé de Oliveira

Me. Milena Edite Casé de Oliveira
Universidade Estadual da Paraíba (UEPB)

Dedico este trabalho a minha família por estarem sempre apoiando minha trajetória.

AGRADECIMENTOS

Agradeço primeiramente a Deus por estar sempre a me guiar com minhas orações. Aos meus pais Rosana Bento e Geraldo Pereira por estarem sempre me apoiando e me dando suporte para conseguir concluir meu curso. A todos professores que tive durante todo curso por contribuírem com meu aprendizado. E a todos os amigos que fiz durante o curso, em especial Maria Karolina Ramos, Joseferson Barreto e Jefferson Santos por tornarem minha trajetória mais leves além de me apoiarem e me ajudarem nas horas difíceis.

“Deem graças ao Senhor, porque Ele é bom.
O seu amor dura para sempre!”
(Salmos 136:1)

RESUMO

O câncer de mama é a primeira causa de morte por câncer na população feminina na maioria das regiões do Brasil. A análise de sobrevivência é uma técnica estatística que visa analisar dados onde a variável de interesse é o tempo até a ocorrência de um evento, que é o mesmo objetivo em muitos estudos de câncer. Os algoritmos de *Machine Learning* se destacam no desenvolvimento de pesquisas em saúde por terem uma melhor capacidade na predição de resultados capturando relações complexas nos dados. Dentre esses algorítimos temos o *Random Survival Forest*, que é um método projetado para a análise de dados de sobrevivência. Com o objetivo de analisar o tempo de sobrevivência em mulheres acometidas por câncer de mama no estado de Pernambuco, através de técnicas de análise de sobrevivência e *Random Survival Forest*, foi possível concluir que o risco de falecimento por câncer de mama é maior em mulheres acima de 81 anos, mulheres mais jovens com idade entre 21 e 40 anos têm maior probabilidade de sobrevivência, a extensão do tumor possui impacto na sobrevivência dos pacientes além de que o modelo de *Random Survival Forest* pode ser uma alternativa interessante quando é preciso analisar um grande conjunto de dados.

Palavras-chaves: câncer de mama; análise de sobrevivência; random survival forest.

ABSTRACT

Breast cancer is the leading cause of cancer death in the female population in most regions of Brazil. Survival analysis is a statistical technique that aims to analyze data where the variable of interest is the time until the occurrence of an event, which is the same objective in many cancer studies. Machine Learning algorithms excel in development of health research because they have a better ability to predict outcomes by capturing complex relationships in the data. Among these algorithms we have the *Random Survival Forest*, which is a method designed for the analysis of survival data. With the aim of analyzing the survival time in women affected by breast cancer in the state of Pernambuco, through survival analysis techniques and *Random Survival Forest*, it was possible to conclude that the risk of death from breast cancer is greater in women over 81 years old, women younger people between the ages of 21 and 40 are more likely to survive, the extent of the tumor has an impact on the survival of patients, in addition to the *Random Survival Forest* model can be an interesting alternative when you need to analyze a large data set

Keywords: breast cancer; survival analysis; random survival forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Caracterizado pela proliferação anormal das células do tecido mamário, o câncer de mama é o tipo de câncer mais incidente em mulheres em todas as regiões do Brasil, após o câncer de pele não melanoma, este também acomete homens, porém é considerado raro, representando apenas 1% do total de casos da doença (INCA, 2022b). Segundo o INCA (2023) o número estimado de casos novos de câncer de mama no Brasil, para o triênio de 2023 a 2025, é de 73.610 casos, correspondendo a um risco estimado de 66,54 casos novos a cada 100 mil mulheres.

O câncer de mama é a primeira causa de morte por câncer na população feminina em todas as regiões do Brasil, exceto na região Norte, onde o câncer do colo do útero ocupa essa posição (INCA, 2023). Segundo o INCA (2022a) na mortalidade proporcional por câncer em mulheres, no período 2016-2020, os óbitos por câncer de mama ocupam o primeiro lugar no país, representando 16,3% do total.

Em muitos estudos de câncer, o principal resultado em avaliação é o tempo para um evento de interesse (BUSTAMANTE-TEIXEIRA et al., 2002). A Análise de Sobrevivência é uma técnica estatística que visa analisar dados onde a variável de interesse é o tempo até a ocorrência de um evento. Esse tempo é denominado tempo de falha, podendo ser o tempo até a morte do paciente, bem como até a cura ou reincidência de uma doença (ECHEVESTE, 1997).

Com a disponibilidade crescente de dados relevantes para o desenvolvimento de pesquisas em saúde, cada vez mais são utilizados algoritmos de inteligência artificial (*machine learning*) (SANTOS et al., 2019). Esses algoritmos se destacam por terem uma melhor capacidade na predição de resultados capturando relações complexas nos dados, bem como por sua capacidade em lidar com um grande volume de informações (SANTOS et al., 2019).

Dentre esses algoritmos um dos mais utilizados é o *Random Forest*, devido à sua simplicidade e sua capacidade de utilização tanto para tarefas de classificação quanto de regressão (DONGES, 2018). Este algoritmo é também uma boa opção no desenvolvimento de um modelo em curto espaço de tempo. Além disso, ele provê um bom indicador de importância para as variáveis em estudo (DONGES, 2018).

Dentro do *Random Forest* temos o *Random Survival Forest*, que é um método projetado para a análise de dados de sobrevivência (ISHWARAN et al., 2008). Além de operar diretamente sobre o tempo de sobrevida, uma outra grande vantagem do *Random Survival Forest* é a possibilidade de incorporação de variáveis censuradas. Dessa forma, é possível utilizar a informação censurada para construir aprendizado ao modelo, o que pode melhorar o desempenho do mesmo (OLIVEIRA, 2020).

Diante do exposto, o objetivo desse trabalho consiste na aplicação de técnicas de Análise de Sobrevivência com auxílio do *Random Survival Forest* para analisar o tempo de sobrevivência em mulheres acometidas por câncer de mama. Visando obter informações que auxiliem as autoridades responsáveis na tomada de decisão em relação ao diagnóstico e tratamento de câncer de mama.

2 MATERIAL E MÉTODOS

2.1 Material

Os dados utilizados para a análise são de mulheres diagnosticadas com câncer de mama no estado do Pernambuco, no período de 1996, quando ocorre o primeiro diagnóstico, e finalizado em 2017. Disponíveis de forma livre e gratuita no site do Instituto Nacional de Câncer (INCA[1]). O conjunto de dados é formado de 2.337 observações e 12 variáveis: sexo, faixa etária, endereço do estado, descrição da topografia, descrição da doença, meio de diagnóstico, extensão, tipo do óbito, data do óbito, data de diagnóstico, tempo em meses e status.

O tempo em meses foi definido pela diferença da data de óbito e a data de diagnóstico. Os tempos considerados censura à esquerda foram excluídos do banco de dados, sendo então trabalhado apenas o mecanismo de censura à direita, com o tipo de censura aleatória ocorrendo quando a paciente sai do estudo sem ter ocorrido a falha, ou seja, é retirada do estudo antes de um exato período (MARTINS; WERNER, 2010).

Para a análise dos dados foi utilizado o *software R* (versão 4.1.3). Para a Análise de Sobrevivência utilizou-se o pacote *survival* (THERNEAU et al., 2022) e para as técnicas de *Random Survival Forest* foi utilizado o pacote *ranger* (WRIGHT; WAGER; PROBST, 2023). As variáveis utilizadas estão descritas na Tabela 1.

Tabela 1 – Descrição das variáveis utilizadas.

Variável	**Classificação**
Faixa etária	21 a 40; 41 a 60; 61 a 80; 81 ou mais
Meio de diagnóstico	Histologia do tumor primário; histologia da metástase; SDO; pesquisa; citologia; clínico; sem informações
Extensão	Localizado; metástase; sem informações
Tempo em meses	De 1 a 229
Status	Censura = 0; óbito = 1

Fonte: Elaborada pela autora, 2023.

Na variável “Status”, a falha que designa que a paciente faleceu por câncer de mama será representada por “1” e censura; indivíduo não faleceu por câncer de mama ou não se sabe informação será representada por “0”.

2.2 Conceitos Básicos de Análise de Sobrevivência

A análise de sobrevivência é utilizada quando o tempo for o objeto de interesse, seja este interpretado como o tempo até a ocorrência de um evento ou o risco de ocorrência de um evento por unidade de tempo (PINHEIRO, 2022). Segundo Colosimo e Giolo (2006) em análise de sobrevivência a variável resposta é, geralmente, o tempo até a ocorrência de um evento de

[1] https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/registros/base-populacional

interesse. Este tempo é denominado tempo de falha, podendo ser o tempo até a morte do paciente bem como até a cura ou recidiva de uma doença (COLOSIMO; GIOLO, 2006).

Técnicas de análise de sobrevivência podem ser utilizadas em diversas áreas do conhecimento, como a saúde, para estimar o tempo de recuperação de pacientes diagnosticados com determinado tipo de doença, etc (RAMOS, 2022).

2.2.1 Censura

Segundo Colosimo e Giolo (2006) censura é o registro parcial do tempo de falha, devido à perda ou retirada de um elemento do estudo. Em casos clínicos pode ocorrer, por exemplo, quando o paciente perde o contato com o pesquisador ou quando o paciente falece por algum motivo diferente do estudado.

2.2.1.1 Tipos de Censura

Nos estudos clínicos há três tipos de censuras que são mais comuns:

- Censura do tipo I - O estudo termina após um tempo pré-estabelecido.
- Censura do tipo II - O estudo é encerrado após ocorrer uma quantidade pré-estabelecida de falhas no evento de interesse.
- Censura aleatória - Ocorre com frequência na área médica; o indivíduo sai do estudo sem ter ocorrido o evento de interesse.

Segundo Strapasson (2007), o mecanismo de censura pode ser classificado em censura à direita, censura à esquerda e censura intervalar. A censura à direita é a mais utilizada na qual o tempo de ocorrência do evento de interesse está à direita do tempo registrado. A censura à esquerda é aquela em que o indivíduo ou objeto já experimenta o evento de interesse no início do estudo. E a censura intervalar é aquela em que não se sabe o tempo exato em que a falha ocorreu, sabe-se apenas que se deu em um intervalo de tempo.

Para a análise dos dados de sobrevivência, os tempos dos indivíduos observados t_i sendo t o tempo e i $(i = 1, \ldots, n)$ os indivíduos observados, a variável indicadora de falha ou censura δ_i é representada da seguinte forma:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo de censura} \end{cases}$$

2.2.2 Função de Sobrevivência

Estima a probabilidade de um indivíduo sobreviver por mais do que um determinado tempo t, segundo o evento de interesse. é uma das funções mais utilizadas no estudo da análise de sobrevivência. Segundo Pinheiro (2022) é definida por:

$$S(t) = P(T > t).$$

Note que, a função de sobrevivência pode ser obtida em termos da função de distribuição acumulada. Nesse contexto, a função de distribuição acumulada pode ser entendida como a probabilidade de uma observação não sobreviver ao tempo t. Segundo Pinheiro (2022) a função é definida por:

$$F(t) = 1 - S(t).$$

2.2.3 Função de Risco

A função de risco $\lambda(t)$ é definida como a probabilidade de um indivíduo sofrer o evento em um intervalo de tempo, dado que ele sobreviveu até o tempo t, ou seja, é uma probabilidade de falhar durante um intervalo de tempo muito pequeno. Seja T uma variável aleatória que corresponde o tempo até a ocorrência de um evento, a função é então definida como:

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

2.2.4 Função de Risco Acumulado

Outra função importante é a função de risco acumulado, ela fornece a soma de todas as taxas de falhas $\lambda(u)$ dos indivíduos até o tempo t, propriamente dita, a taxa de falha acumulada. Esta é uma função que não possui uma interpretação direta, mas fornece informação no que se refere à função taxa de falha. É dada por:

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

2.3 Técnicas Não-Paramétricas

Para a estimação da função de sobrevivência pode se considerar técnicas paramétricas e técnicas não-paramétricas. As técnicas não-paramétricas podem ser utilizadas para verificar se o modelo paramétrico está bem ajustado pois não estabelecem pressupostos sobre a distribuição dos dados. Segundo Colosimo e Giolo (2006) existem técnicas não-paramétricas para estimar parâmetros em análise de sobrevivência, obtendo a opção de ajustar os dados utilizando-se os modelos paramétricos probabilístico para tempo de falha. Há dois estimadores não-paramétricos mais utilizados: Kaplan-Meier e Nelson-Aalen. Neste trabalho será utilizado o estimador de Kaplan-Meier.

2.3.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier foi proposto por Kaplan e Meier (1958), é o estimador mais utilizado em estudos clínicos de análise de sobrevivência. Esse modelo é aplicado por ser não viciado para amostras grandes e também por permitir a estimativa no tempo mesmo possuindo casos censurados. Pelo estimador de Kaplan-Meier é possível comparar os tempos distintos de falhas através da curva de sobrevivência. Sua função de sobrevivência estimada é uma função "escada" pelo qual os "degraus" correspondem aos tempos distintos de falhas observados. Considerando:

- $t_1 < t_2 < \ldots < t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em t_j, $j = 1, \ldots, k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é definido por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right).$$

Breslow e Crowley (1974) destacam algumas propriedades desse estimador:

- é fracamente consistente;
- converge assintoticamente para um processo gaussiano e
- é estimador de máxima verossimilhança de $S(t)$.

2.3.2 Comparação de Curvas de Sobrevivência

A comparação de curvas de sobrevivência tem sido muito procurada principalmente na área médica (BUSTAMANTE-TEIXEIRA et al., 2002). Estudos como este possibilitam a comparação entre diversas categorias de uma única variável utilizando as curvas de sobrevivência de uma técnica não-paramétrica. Na área médica, como foi mencionado, o interesse principal é avaliar se dois ou mais tratamentos são estatisticamente iguais.

2.3.2.1 Logrank

O teste de logrank proposto por Mantel (1966) foi utilizado nesse estudo para comparar curvas de sobrevivência. Este teste é particularmente apropriado quando a razão das funções de risco dos grupos a ser comparados é aproximadamente constante.

Hipóteses para o teste:

$$\begin{cases} H_0: \text{ não há diferença entre as curvas de sobrevivência.} \\ H_1: \text{ há diferença entre as curvas de sobrevivência.} \end{cases}$$

Suponha uma comparação entre duas curvas de sobrevivência $S_1(t)$ e $S_2(t)$. Considere ainda como $t_1 < t_2 < \ldots < t_k$ sendo os tempos distintos de falha obtidos pela combinação de duas amostras, d_j o número de falhas, n_j o número de indivíduos sob risco inferior a d_j na amostra combinada e respectivamente d_{ij} e d_{ij} na amostra $i; i = 1,2$ e $j = 1,\ldots,k$. A estatística de teste logrank é dada por:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2},$$

sendo que para cada tempo distinto, d_{2j} representa o número observado de falhas no grupo dois, $w_{2j} = n_{2j}d_j n_j^{-1}$ corresponde a média de falhas para o grupo dois e $(V_j)_2$ as variâncias para o mesmo grupo. O teste tem uma distribuição qui-quadrado com 1 grau de liberdade e é baseado na hipótese nula, ou seja, $H_0 : S_1(t) = S_2(t)$.

2.4 Aprendizado de Máquina *(Machine Learning)*

O *Machine Learning* é uma área derivada da Inteligência Artificial. Por meio de algoritmos, dá aos computadores a capacidade de identificar padrões em dados massivos e fazer previsões. Essa aprendizagem permite que os computadores efetuem tarefas específicas de forma autônoma, ou seja, sem necessidade de serem programados (IBERDROLA, 2023). Cada algoritmo é responsável por um comando diferente e é a combinação entre eles que permite aos computadores tomarem decisões de acordo com as situações ou informações que encontrarem (NEOWAY, 2020).

2.4.1 Random Survival Forest (RSF)

Random Survival Forest (RSF) é um método desenvolvido para estender o algoritmo de *Random Forest* (RF) ao cenário de dados de sobrevivência censurados à direita. O RSF calcula uma RF usando o teste log-rank como critério de divisão. Calcula os riscos cumulativos dos nós das folhas em cada árvore e a média deles no seguinte conjunto. A árvore cresce até o tamanho máximo sob a condição de que cada nó terminal não tenha menos que um número pré-especificado de mortes (LEE et al., 2018). As amostras que estão fora são então usadas para calcular o erro de previsão da função de risco cumulativo do conjunto. Tendo como estratégia geral os seguintes passos:

- Passo 1. Desenhe B amostras bootstrap.
- Passo 2. Crie uma árvore de sobrevivência com base nos dados de cada uma das amostras bootstrap $b = 1,...,B$:

  - (a) Em cada nó da árvore, selecione um subconjunto das variáveis preditoras.
  - (b) Entre todas as divisões binárias definidas pelas variáveis preditoras selecionadas em (a) , encontre a melhor divisão em dois subconjuntos (os nós filhos) de acordo com um critério adequado para dados censurados à direita, como o teste log-rank.
  - (c) Repita (a)-(b) recursivamente em cada nó filho até que um critério de parada seja satisfeito.

- Passo 3. Agregar informações dos nós terminais (nós sem divisão adicional) das árvores de sobrevivência B para obter um conjunto de previsão de risco.

O conjunto é construído pela agregação de estimadores Nelson-Aalen baseados em árvores. Em cada nó terminal de uma árvore, a função de risco cumulativo condicional é estimada usando o Nelson-Aalen usando os dados "in-bag" (ISHWARAN et al., 2008).

$$\widehat{H}_b(t|x) = \int_0^t \frac{\widetilde{N}_b^*(ds,x)}{\widetilde{Y}_b^*(s,x)}$$

Sendo $\widetilde{N}_b^*(ds,x)$ os eventos não censurados até o tempo s e $\widetilde{Y}_b^*(s,x)$ é o número em risco no tempo s. A função de sobrevivência do conjunto da floresta de sobrevivência aleatória é

$$\widehat{S}^{rsf}(t|x) = exp\left(-\frac{1}{B}\sum_{b-1}^{B}\widehat{H}_b(t|x)\right).$$

Com o objetivo de avaliar a precisão de um modelo, temos as estatísticas de desempenho. Dentre elas temos Brier Score e C-index, que são calculadas com base na função de risco acumulado de Nelson-Aalen para todas as árvores. Brier Score se aplica no conjunto de dados de teste objetivando a avaliação da qualidade das predições, já a estatística C-index refere-se a avaliação da precisão no conjunto de dados de treino e, para isso, é utilizada nos dados Out-Of-Bag (OOB), que são dados separados ao se calcular as amostras com o intuito de utilizá-los para medir o erro de previsão.

2.4.2 C-index

O índice de concordância ou C-index é uma generalização da área sob a curva ROC (AUC) que pode levar em conta dados censurados. Representa a avaliação global do poder de discriminação do modelo: é a capacidade do modelo de fornecer corretamente uma classificação confiável dos tempos de sobrevivência com base nos escores de risco individuais (UNOA et al., 2011). Pode ser calculado com a seguinte fórmula :

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

Onde:

- η_i é a pontuação de risco de uma unidade de i
- $1_{T_j<T_i} = 1$ se $T_j < T_i$, caso contrário, é igual a 0
- $1_{\eta_j>\eta_i} = 1$ se $\eta_j > \eta_i$, caso contrário, é igual a 0

Similar à curva ROC, C-index = 1 corresponde a melhor previsão do modelo, e C-index = 0,5 corresponde a uma previsão aleatória.

2.4.3 Brier Score

O Brier Score é usado para avaliar a precisão de uma função de sobrevivência prevista em um determinado momento t; representa as distâncias quadradas médias entre o estado de sobrevivência observado e a probabilidade de sobrevivência prevista e é sempre um número entre 0 e 1, sendo 0 o melhor valor possível.

Dado um conjunto de dados de N amostras, $\forall i \in [\![1,N]\!], (\vec{x}_i, \delta_i, T_i)$ é o formato de um ponto de dados e a função de sobrevivência prevista é $\hat{S}(t,\vec{x}_i), \forall t \in R^+$: Na ausência de censura correta, o Brier Score pode ser calculada de modo que:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}(1_{T_i>t} - \hat{S}(t,\vec{x}_i))^2$$

No entanto, se o conjunto de dados contiver amostras censuradas à direita, será necessário ajustar a pontuação ponderando as distâncias ao quadrado usando o método de probabilidade inversa de pesos de censura. Sendo $\hat{G}(t) = P[C > t]$ o estimador da função de sobrevivência condicional dos tempos de censura calculados pelo método de Kaplan-Meier, onde C é o tempo de censura.

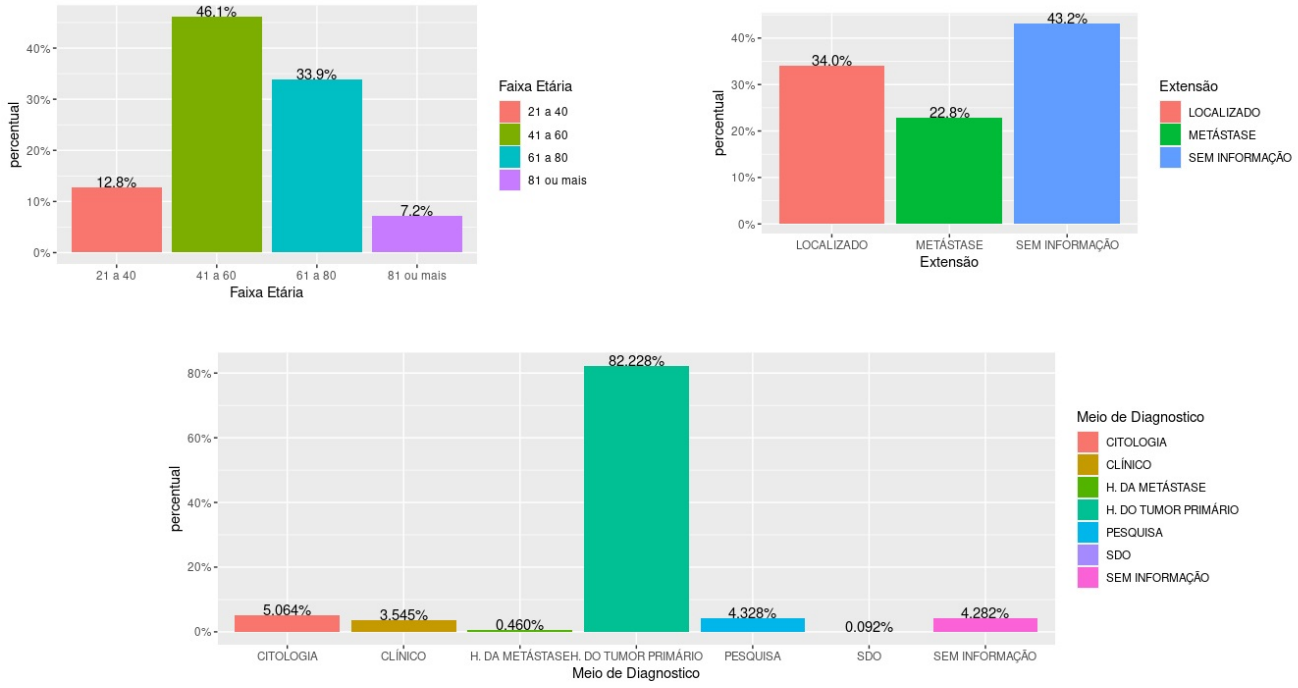
$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(0-\hat{S}(t,\vec{x}_i)\right)^2 \cdot 1_{T_i\leq t,\delta_i=1}}{\hat{G}(T_i^-)} + \frac{\left(1-\hat{S}(t,\vec{x}_i)\right)^2 \cdot 1_{T_i>t}}{\hat{G}(t)}\right)$$

Como referência, um modelo útil terá um Brier Score abaixo de 0,25.

3 RESULTADOS E DISCUSSÃO

Primeiramente foi efetuada uma análise descritiva afim de verificar o comportamento dos dados. Na Figura 2 temos a representação gráfica das variáveis faixa etária, extensão do tumor e meio de diagnóstico. Podemos observar que cerca de 46% das mulheres diagnosticadas com câncer de mama estão na faixa etária de 41 a 60 anos, aproximadamente 34% estão na faixa de 61 80 anos, 12,8% de 21 a 40 anos e 7,2% de 81 anos ou mais. Considerando a extensão do tumor, vemos que 34% das mulheres foram diagnosticadas no grau de tumor localizado, 22,8% no grau de metástase (avançado) e cerca de 43% não se tem informação. Observamos também que cerca de 82% foram diagnosticadas pelo meio de histologia do tumor primário.

Figura 1 – Frequência de casos de câncer de mama por faixa etária, extensão do tumor e meio de diagnóstico respectivamente.



Fonte: Elaborada pela autora, 2023.

Na Tabela 2 podemos observar o percentual de censuras e óbitos para as variáveis em análise. Em relação a variável extensão das mulheres diagnosticadas com câncer de mama que possuíam um tumor localizado, 89,02% (737) faleceram de câncer de mama e aquelas que possuíam um tumor em metástase, 94,65% (495) faleceram pelo mesmo motivo. Na variável faixa etária, 95,52% (1.001) das mulheres que tinham idades de 41 a 60 anos faleceram e 88,14% (156) das que tinham 81 anos ou mais falharam. Na variável meio de diagnóstico, todas as mulheres que foram diagnosticadas por meio de análise das células (citologia), vieram a óbito.

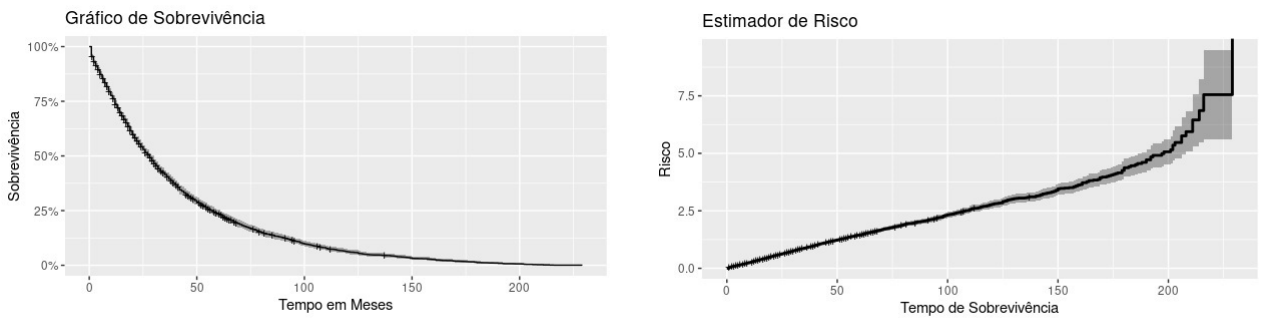
Tabela 2 – Percentual de censuras e óbitos para as covariáveis em estudo dentro de cada categoria.

Variável	**Classificação**	**Censuras + Óbitos**	**Censuras (em %)**	**Óbitos (em %)**
Extensão	1: Localizado	829	10,98%	89,02%
	2: Metástase	523	5,35%	94,65%
	3: Sem informação	985	4,67%	95,33%
Faixa etária	1: 21 a 40	288	3,12%	96,87%
	2: 41 a 60	1.048	4,48%	95,52%
	3: 61 a 80	824	10,68%	89,32%
	4: 81 ou mais	177	11,86%	88,14%
Meio de diagnóstico	1: Citologia	110	0%	100%
	2: Clínico	79	2,53%	97,47%
	3: Histologia da metástase	11	9,09%	90,91%
	4: Histologia do tumor primário	1.935	7,70%	92,30%
	5: Pesquisa	104	9,62%	90,38%
	6: Somente por declaração de óbito - SDO	2	0%	100%
	7: Sem informação	96	3,13%	96,87%

Fonte: Elaborada pela autora, 2023.

Na Figura 3 temos a curva de sobrevivência e o estimador de risco para as mulheres diagnosticadas com câncer de mama. A partir da análise gráfica e dos dados fornecidos pelo modelo de sobrevivência foi observado que o tempo mediano de sobrevivência foi de 28 meses, indicando que 50% (1.168) das pacientes faleceram antes do 28º mês e 50% (1.169) depois. Observamos também que o risco de ocorrência da falha, isto é, o risco de morte, aumenta consideravelmente ao decorrer do tempo.

Figura 2 – Curva de sobrevivência e estimador de risco (Kaplan-Meier) para mulheres diagnosticadas com câncer de mama



Fonte: Elaborada pela autora, 2023.

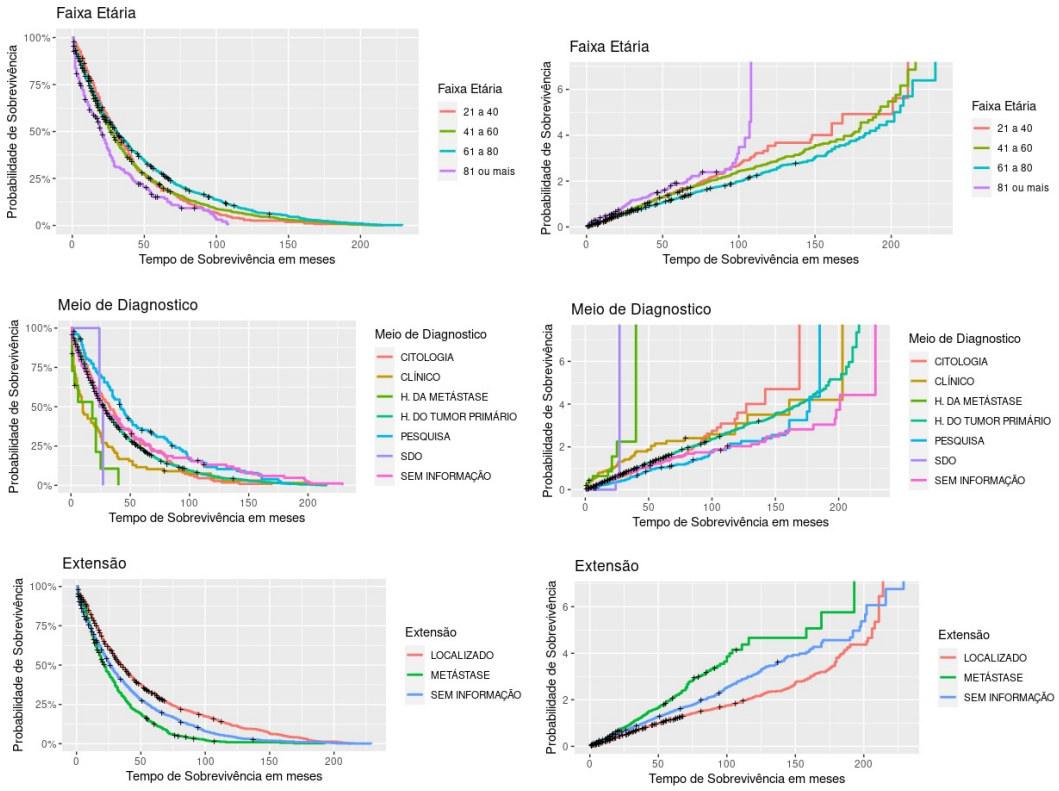
Analisando a Figura 4, podemos observar que as pacientes na faixa etária de 21 a 40 anos possuem maior probabilidade de sobrevivência até o 28º mês. A partir desse mês, pessoas com faixa etária de 41 a 60 anos tiveram uma probabilidade de sobrevivência maior comparada às demais curvas. As pessoas com 81 anos ou mais possuem o tempo de sobrevivência menor e risco elevado de morte.

As pessoas diagnosticadas através da pesquisa tiveram uma probabilidade de sobrevivência superior em comparação com as demais curvas, Até o 40º mês o risco de obter o evento de interesse oscila entre pessoas diagnosticadas pela histologia da metástase e exames clínicos.

As mulheres diagnosticadas com tumor localizado possuem um maior tempo de sobrevivência e risco menor de morte, enquanto as pacientes com mestástase possuem menor tempo de sobrevivência e risco elevado de morte.

Foi aplicado o teste *logrank* com o objetivo de comparar as curvas de sobrevivência de cada grupo de variáveis (faixa etária, meio de diagnóstico e extensão). Considerando as hipóteses H_0 : não há diferença entre as curvas de sobrevivência e H_1 : há diferença entre as curvas de sobrevivência, foi verificado nas variáveis em estudo,faixa etária, meio de diagnóstico e extensão, que o p-valor foi menor que 0,05, sendo assim, rejeitamos a hipótese nula H_0 : e podemos afirmar que há diferença significativa entre as curvas de sobrevivência das variáveis em estudo.

Figura 3 – Comparação entre curvas de sobrevivência e estimador de risco (Kaplan-Meier) por covariável para mulheres diagnosticadas com câncer de mama



Fonte: Elaborada pela autora, 2023.

3.1 Random Survival Forest

Nesta seção serão apresentados os resultados obtidos com a aplicação das técnicas de *Random Survival Forest* (RSF). Foram utilizados os conceitos básicos de *Machine Learning* para divisão dos dados, sendo 70% para treino e 30% teste.

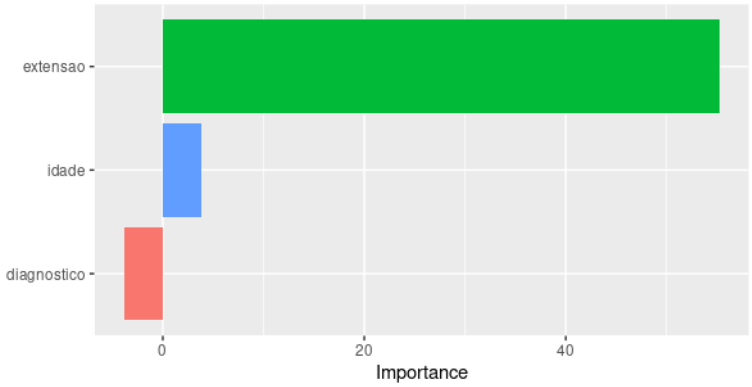
Tabela 3 – Saída do modelo de RSF

Tipo	Survival
Número de árvores	1000
Número de variáveis independentes	3
OOB erro de performance (1-C)	0.45581
Brier score	0.09

Fonte: Elaborada pela autora, 2023.

Na Tabela 3 temos a saída do modelo de RSF, que foi ajustado com todas covariáveis para 1000 amostras bootstrap. Vemos que o erro associado à performance do modelo é de aproximadamente 0,45, com isso temos que a taxa C-index, que mede a preditividade do modelo, ou seja, o quanto o modelo previu corretamente, com base nos dados de treino e teste, é de aproximadamente 0,55 (55%). No cálculo do Brier Score, que é uma maneira de quantificar o quão precisa é a previsão, foi obtido o valor de 0,09 indicando um bom desempenho para o modelo preditivo, pois este deve estar abaixo de 0,25 para indicar uma boa precisão do modelo.

Figura 4 – Gráfico de seleção de variáveis pela estatística VIMP.

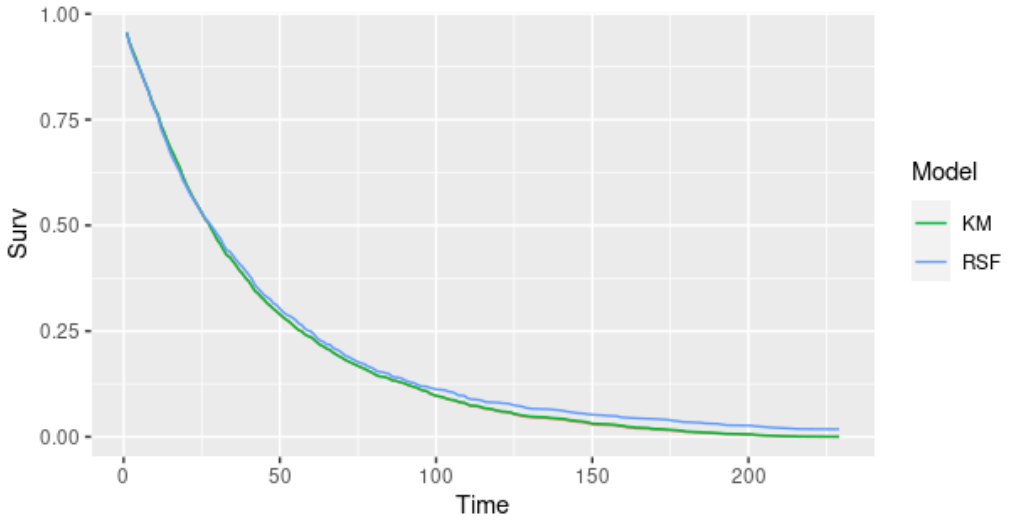


Fonte: Elaborada pela autora, 2023.

Podemos observar na Figura 6 o gráfico de seleção de variáveis pela estatística VIMP, onde cada barra representa o valor de impacto das variáveis no modelo. Temos que a variável que indica a extensão do tumor apresenta um maior impacto no modelo, enquanto a variável que indica o meio de diagnóstico apresenta um impacto ligeiramente negativo. Ao retirar a variável meio de diagnóstico foi observado valores piores para as métricas C-Index e Brier Score, desse modo optou-se por permanecer com o modelo RSF completo. Na Figura 7 temos uma

comparação entre curvas de sobrevivência utilizando os modelos de Kaplan Meier (KM) e RSF. Vemos que a curva do modelo de RSF até o 100º mês possui a previsão próxima ao modelo de Kaplan Meier, após esse período apresenta uma ligeira diferença entre as curvas.

Figura 5 – Comparação entre curvas de Kaplan Meier e Random Survival Forest



Fonte: Elaborada pela autora, 2023.

No artigo de Santos et al. (2019) se discute sobre a capacidade preditiva dos modelos de *Machine Learning*, onde os modelos ajustados não obtiveram boas performances preditivas. Entre outras características, esse cenário pode estar relacionado à disponibilidade de um número reduzido de observações, sobretudo de observações com desfecho presente, ou de preditores para o treinamento e, mais frequentemente, ao sobreajuste do modelo para os dados existentes, no caso de algoritmos mais flexíveis (BUTCHER; SMITH, 2020). Esse é o caso desde estudo, em que apresentou um erro de performance de 0,45 (45%).

É importante destacar que mesmo um modelo preditivo com bom poder discriminatório e bem calibrado pode não se traduzir em melhores cuidados à saúde, pois uma predição acurada não diz o que deve ser feito para modificar o desfecho sob análise (CHEN; ASCH., 2018). Além disso, modelos preditivos de óbito, bem como de doenças crônicas podem basear-se não só em fatores de risco modificáveis, mas também em características biológicas não modificáveis, como idade e sexo, que, embora contribuam para a performance preditiva do modelo, podem não ser relevantes em estratégias de prevenção ou controle (MENA et al., 2012).

4 CONCLUSÃO

Através da análise de sobrevivência e estimativas de Kaplan Meier foi possível concluir que o risco de falecimento por câncer de mama aumenta com o passar dos meses e esse risco aumenta consideravelmente quando a paciente apresenta um tumor em metástase. As mulheres com idade acima de 80 anos possuem um maior risco de morte enquanto mulheres com idade entre 21 e 40 anos possuem maior probablidade de sobrevivência até o 28º mês após o diagnóstico.

Utilizando as técnicas de Random Survival Forest foi possível constatar que a extensão do tumor possui um maior impacto na sobrevivência das pacientes. Além disso podemos ver que o modelo de RSF pode ser um método interessante quando é preciso analisar um grande conjunto de dados, pois apesar de ter apresentado um valor de 0,45 para o erro de performance do modelo foi obtido uma previsão próxima ao modelo de Kaplan Meier até o 100º mês de diagnóstico.

REFERÊNCIAS

BRESLOW, N.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, JSTOR, p. 437–453, 1974. Citado na página 15.

BUSTAMANTE-TEIXEIRA et al. Técnicas de análise de sobrevida. *Cadernos de Saúde Pública*, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 18, n. 3, p. 579–594, May 2002. ISSN 0102-311X. Disponível em: <https://doi.org/10.1590/S0102-311X2002000300003>. Citado 2 vezes nas páginas 11 e 15.

BUTCHER, B.; SMITH, B. J. Feature engineering and selection: A practical approach for predictive models. *The American Statistician*, 2020. Citado na página 23.

CHEN, J. H.; ASCH., S. M. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med*, 2018. Citado na página 23.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1. ed. São Paulo: Blucher, 2006. Citado 3 vezes nas páginas 12, 13 e 14.

DONGES, N. Aprendendo em uma floresta aleatória. *Towards Data Science*, 2018. Disponível em: <https://towardsdatascience.com/>. Citado na página 11.

ECHEVESTE, S. S. Análise de sobrevivência: Um estudo na Área educacional. 1997. Citado na página 11.

IBERDROLA. *O QUE É 'MACHINE LEARNING'?: Conheça os principais benefícios do 'Machine Learning'*. 2023. Disponível em: https://www.iberdrola.com/inovacao/o-que-e-machine-learning. Citado na página 16.

INCA, I. N. de C. *Atlas da mortalidade.* 2022. Disponível em: https://www.inca.gov.br/app/mortalidade. Citado na página 11.

INCA, I. N. de C. *Câncer de mama: vamos falar sobre isso?* 7. ed. Rio de Janeiro, 2022. Citado na página 11.

INCA, I. N. de C. *Estimativa 2023 Incidência de Câncer no Brasil*. 7. ed. Rio de Janeiro, 2023. Citado na página 11.

ISHWARAN, H. et al. Random survival forests. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841 – 860, 2008. Disponível em: <https://doi.org/10.1214/08-AOAS169>. Citado 2 vezes nas páginas 11 e 17.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 457–481, 1958. Citado na página 15.

LEE, C. et al. Deephit: A deep learning approach to survival analysis with competing risks. 2018. Citado na página 16.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports*, v. 50, p. 163–170, 1966. Citado na página 15.

MARTINS, V. L. M.; WERNER, L. Análise não paramétrica de falhas ao longo do calendário para alto-falantes. *Produto Produção*, v. 11, n. 3, 2010. Citado na página 12.

MENA, L. J. et al. Machine learning approach to extract diagnostic and prognostic thresholds: application in prognosis of cardiovascular mortality. *Comput Math Methods Med*, 2012. Citado na página 23.

NEOWAY. *Machine Learning: Conceitos, definição e mais*. 2020. Disponível em: https://blog.neoway.com.br/machine-learning/. Citado na página 16.

OLIVEIRA, D. B. S. Estimativa de sobrevida de pacientes com glioblastoma por meio de algoritmos baseados em random forests. 2020. Citado na página 11.

PINHEIRO, N. M. Entenda o que é análise de sobrevivência e como utilizar essa técnica em projetos de data science. *Data Hackers*, 2022. Citado 3 vezes nas páginas 12, 13 e 14.

RAMOS, M. K. de F. AnÁlise de sobrevivÊncia de mulheres do estado de pernambuco diagnosticadas com cÂncer de mama. 2022. Citado na página 13.

SANTOS, H. G. d. et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de são paulo, brasil. *Cadernos de Saúde Pública*, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 35, 2019. Disponível em: <https://doi.org/10.1590/0102-311X00050818>. Citado 2 vezes nas páginas 11 e 23.

STRAPASSON, E. *Comparação de Modelos com Censura Intervalar em Análise de Sobreviência*. Tese (Doutorado em Agronomia) — Universidade de São Paulo - Escola Superior de Agricultura “Luiz de Queiroz”, 2007. Citado na página 13.

THERNEAU, T. M. et al. *ranger: A Fast Implementation of Random Forests*. [S.l.], 2022. R package version 3.4-0. Disponível em: <https://cran.r-project.org/web/packages/ranger>. Citado na página 12.

UNOA, H. et al. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, v. 30, 2011. Citado na página 17.

WRIGHT, M. N.; WAGER, S.; PROBST, P. *survival: Survival Analysis*. [S.l.], 2023. R package version 4.1.3. Disponível em: <https://cran.r-project.org/web/packages/survival>. Citado na página 12.