ISSN: 2317-0840

# Comparação de *Random Survival Forest* e modelo de Cox com relação a performance de previsão: Um estudo de caso

Tiago A. Oliveira[1†], Pedro Augusto F. Silva[2], Hiago José A. A. Martins[3], Lucas C. Pereira[4], Alisson L. Brito[5], Ednário B. Mendonça[6]

[1] *Departamento de Estatística - UEPB.*
[2] *UEPB - Universidade Estadual da Paraíba. E-mail: pedro1_20@hotmail.com.*
[3] *UEPB - Universidade Estadual da Paraíba. E-mail: hiago1803@gmail.com.*
[4] *UEPB - Universidade Estadual da Paraíba. E-mail: lacascp@hotmail.com.*
[5] *Programa de Pós Graduação em Estatística e Experimentação Agropecuária - UFLA. E-mail: alissonjs95@gmail.com.*
[6] *Departamento de Estatística - UEPB. E-mail: eddybarbosa92@gmail.com.*

**Resumo:** *A análise de sobrevivência é atualmente uma das ferramentas estatísticas que mais crescem no meio acadêmico. Na análise de sobrevivência existe uma sólida teoria de modelos de regressão que pode ser usada para modelar dados com observações incompletas chamadas censuras, a maioria desses modelos são paramétricos e existe também o modelo semi-paramétrico de riscos proporcionais de Cox. Neste sentido, os modelos Machine Learning em conjunto com o Random Forest em análise de sobrevivência (RSF) são uma alternativa crescente para o uso em predição. Foram ajustados 4 diferentes configurações de coariáveis no RSF, partindo de um modelo saturado com presença de interação até um modelo parcimonioso baseado em critérios própios a área de Machine Learning para escolha de variáveis. Os modelos foram comparados contra o modelo de Cox via critérios C-index e Brier Score Index - IBS. O melhor modelo ajustado para predição foi o modelo completo com todas as covariáveis sob modelagem de Random Survival Forest.*

**Palavras-chave:** Análise de Sobrevivência; Riscos Proporcionais; Machine Learning.

**Abstract:** *Survival analysis is currently one of the fastest growing statistical tools in academia. In survival analysis there is a robust regression model theory that can be used to model data with incomplete observations called censoring, most of these models are parametric, and there is also the Cox proportional hazards model. Machine Learning in conjunction with Random Forest in Survival Analysis (RSF) are an increasing alternative for use in prediction. Four different configurations of coefficients were adjusted in the RSF, starting from a saturated model with presence of interaction to a parsimonious model based on the criteria of the Machine Learning area to choose variables. The models were compared against the Cox model using the C-index and Brier Score Index (IBS) criteria. The best model adjusted for prediction was the complete model with all covariates under Random Survival Forest modeling.*

**Keywords:** Survival Analysis; Proportional risks; Machine Learning.

[†]Autor correspondente: tiagoestatistico@gmail.com.

# Introdução

Diabetes *Mellitus* é uma doença caracterizada pela elevação da glicose no sangue (hiperglicemia). Pode ocorrer devido a defeitos na secreção ou na ação do hormônio insulina. A Retinopatia Diabética é uma complicação que ocorre quando o excesso de glicose no sangue danifica os vasos sanguíneos dentro da retina. Caso o paciente não busque tratamento, com o passar do tempo, porém, a visão passa a piorar, podendo até mesmo chegar à cegueira, caso não seja tratada. A doença apresenta 4 fases, sendo a última a fase mais agressiva da doença (SOCIEDADE BRASILEIRA DE DIABETES, 2019; SOCIEDADE BRASILEIRA DE ENDOCRINOLOGIA E METABOLOGIA, 2019).

Em estudos de tempo até acontecimento de evento de interesse (*follow-up studies*) o teste *Logrank* e o modelo de regressão de Cox, são mais amplamente utilizados (Yosefian et al., 2018). O modelo semiparamétrico de Cox pode ser usado para identificar variáveis que afetam significativamente a variável resposta em estudo e apresenta seus resultados em termos da Razão de Riscos (RR), com a dificuldade de nem sempre trazerem luz as decisões que tem que ser tomadas do ponto de vista clínico. Neste sentido, quando o maior objetivo do estudo é realizar predição na taxa de sobrevivência dos indivíduos, os métodos de *Machine Learning* tem ganhado notória atenção na última década.

*Random Forest* (RF) é um método estatístico não paramétrico que não requer suposições distribucionais sobre a relação das covariaveis com a variável resposta (BREIMAN, 2001). RF é uma técnica não linear robusta que otimiza a acurácia de predição, por realizar um ajuste sobre o conjunto de árvores no intuito de estabilizar as estimativas de um modelo. *Random Survival Forests* (RSF) (ISHWARAN e KOGALUR, 2007; ISHWARAN et al., 2008) são extensões da técnica de RF de Breiman (2001), permitindo uma análise não paramétrica eficiente de dados de tempo até a ocorrência de um evento de interesse (EHRLINGER, 2016). Neste sentido, a análise de sobrevivência por *Random Forest* RF é um método de *Machine Learning* que combina agregação *bootstrap* com árvores de sobrevivência aleatórias para *follow-up studies*, com a característica de censura à direita. A abordagem não assume um modelo e assim, fornece uma alternativa flexível ao modelo de regressão de Cox (RYTGAARD e GERDS, 2018).

Os métodos de (RF) em análise de sobrevivência (RSF), tem se mostrado adequados, principalmente por trazer mais informações clínicas aos pesquisadores, por meio de regras de decisão sobre as árvores de probabilidades construídas (ISHIWARAN et al., 2008). As árvores consistem de raízes internas, ou nó filhos e nós terminais. Em um primeiro passo, todos os sujeitos são colocados em um nó raiz. Os indivíduos são categorizados dentro de dois nós filhos com máxima diferença entre eles. Isto é feito por extensiva procura entre todas as variáveis para encontrar a variável separadora (*cutoff*) que maximiza a diferença entre os indivíduos. Todas os possíveis *cutoffs* de todas as variáveis independentes são tratados para explorar os mais altos valores da estatística de *Logrank* (= menor Valor P). Quando a primeira divisão é criada, o processo se repete dentro de cada nó interno. Isto cria uma estrutura para a árvore até dividir os sujeitos em um só nó terminal.

O presente trabalho visa predizer o tempo até a cegueira de pacientes com retinoplastia diabética, por meio da técnica de *Machine Learning* e *Random Forest* aplicadas a análise de sobrevivência, neste sentido, modelo de Cox será ajustado e seus resultados serão comparados com diferentes configurações de variáveis preditoras no RSF, afim de encontrar qual modelo tem melhor capacidade preditiva. Mensurações de performance preditiva de Brie Score Index(IBS) e C-Index (Curva ROC generalizada) serão obtidas para avaliar os

modelos, além de uma gama de análise gráficas de resíduos (Cox), Predições e estimações sobre dados de *Out-Of-Bag* (OOB) de treino e predições sobre os dados de teste.

# Fundamentação Teórica

## Análise de Sobrevivência

No ramo da análise de sobrevivência, a variável resposta, é usualmente, o tempo até a ocorrência de um determinado evento de interesse. A principal característica em dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. A censura pode ocorrer por diversos motivos, sendo uma das principais características em estudos de sobrevivência. Existem diversos tipos de censuras, das quais a censura à direita, ocorre quando após o período de estudo é terminado e um ou mais indivíduos em estudo experimentam o evento de interesse (COLOSSIMO; GIOLO, 2006).

### Funções em Análise de Sobrevivência

Em análise de sobrevivência a variável resposta é o tempo até a ocorrência de um determinado evento. Esta variável é caracterizada, normalmente como sendo uma variável aleatória contínua, e associada a essa variável aleatória há uma função denominada de função de densidade de probabilidade a qual pode ser interpretada como sendo a probabilidade de um indivíduo sofrer o evento em um intervalo de tempo. Por outro lado, a função de sobrevivência é uma das funções mais importantes em estudos de sobrevivência, pois ela caracteriza a probabilidade de um indivíduo sobreviver até um certo tempo $t$, ou seja, a probabilidade deste indivíduo não falhar (ou ter o desfecho) até um tempo $t$. Ela pode ser vista como sendo o complementar da função de distribuição acumulada. A função de taxa de falha ou função de risco pode ser entendida como o risco instantâneo da ocorrência de um determinado evento, dado que este evento não tenha ocorrido até o presente momento. A Função de Risco Acumulado é outra função bastante utilizada na análise de sobrevivência. Ela descreve a taxa de falha acumulada do indivíduos (CARVALHO et al., 2005; COLOSIMO E GIOLO, 2006).

### Estimador limite-produto de *Kaplan-Meier*

Proposto por Kaplan e Meier em 1958, o estimador não paramétrico de *Kaplan-Meier* é atualmente o mais utilizado na literatura para estimar a função de sobrevivência na presença de censura. Segundo Colossimo e Giolo (2006), este estimador considera intervalos de tempo como os tempos de falha da amostra, bem como o número de falhas distintas. O estimador produto de *Kaplan-Meier* é definido como

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

em que $t_1 < t_2 < \cdots < t_k$ são os $k$ tempos distintos e ordenados de falha, $d_j$ é o número de falhas em $t_j, j = 1,...,k$ e $n_j$ é o número de indivíduos sob risco em $t_j$.

**Comparação entre as curvas de sobrevivência**

Muitas vezes em análise de sobrevivência estamos interessados em testar se as curvas de sobrevivência entre dois ou mais grupos são iguais ou não. Algumas estatísticas podem ser utilizadas para comparação entre essas curvas, dentre elas está o teste *logrank* no qual é bastante utilizado na análise de sobrevivência. Este teste é apropriado para populações que possuam a propriedade de riscos proporcionais, onde a hipótese nula é que $S_1(t) = S_2(t)$. A estatística do teste é dada pela seguinte equação

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

em que $d_{2j}$ caracteriza a falha dos indivíduos do grupo 2 no tempo $j$. $w_{2j}$ e $V_j$ é a média e a variância de $d_{2j}$ respectivamente, obtidos a partir da distribuição de $d_{2j}$.(COLOSIMO; GIOLO, 2006)

**O Modelo de Cox**

Conforme Colossimo e Giolo (2006), o modelo de regressão de Cox permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando por covariáveis. No caso especial em que a única covariável é um indicador de grupos, o modelo de Cox assume a sua forma mais simples. Abaixo é apresentado o modelo de Cox para este caso.

$$\lambda(t) = \lambda_0(t)\exp\left\{\mathbf{X}'\beta\right\} \tag{3}$$

**Método de Estimação**

Para que seja possível fazer inferência por meio do modelo de Cox é preciso que se tenha um método para podermos estimar os parâmetros do modelo. Um dos métodos de estimação mais utilizado em modelos de regressão é o método da máxima verossimilhança. Contudo, este método é ineficiente para o modelo de Cox já que esse possui um componente não-paramétrico na função de verossimilhança.

Uma estratégia proposta por Cox foi o método da máxima verossimilhança parcial que consiste em condicionar a construção da função de verossimilhança ao conhecimento da história passada de falhas e censuras para eliminar esta função de perturbação da verossimilhança. Deste modo, a função de verossimilhança parcial para o modelo de Cox é dada da seguinte forma (COLOSIMO e GIOLO, 2006),

$$L(\beta) = \prod_{i=1}^{k}\frac{\exp\mathbf{X}_i'\beta}{\sum_{j\in R(t_i)}\exp\mathbf{X}_j'\beta} = \prod_{i=1}^{n}\left(\frac{\exp\mathbf{X}_i'\beta}{\sum_{j\in R(t_i)}\exp\mathbf{X}_j'\beta}\right)^{\delta_i} \tag{4}$$

Esta função assume que os tempos de sobrevivência são contínuos e não considera a possibilidade de empates nos valores observados. Porém, prática podem ocorrer empates nos tempos de falha ou de censura devido à escala de medida.

Assim, a função de verossimilhança parcial deve ser modificada para incorporar as observações empatadas quando estas estão presentes. A solução da função de verossimilhança é obtida por meio do vetor escore e solução da matriz hessiana.

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[ x_i - \frac{\sum_{j \in R(t_i)} \exp \mathbf{X}_j' \hat{\beta}}{\sum_{j \in R(t_i)} \exp \mathbf{X}_j' \hat{\beta}} \right] = 0. \quad (5)$$

em que $\delta_i$ é o indicador de falha. Os valores de $\beta$ que maximizam a função de verossimilhança parcial, $L(\beta)$, são obtidos resolvendo-se o sistema de equações defifinidos pela equação (5), em que $U(\beta)$ é o vetor de escores de derivadas de primeira ordem da função $l(\beta) = log(L(\beta))$.

## Random Survival Forest

*Random Survival Forest* é um método de conjunto de arvores aleatórias para a análise de dados de sobrevivência censurados à direita. Breiman (2001), mostrou que as árvores de decisão podem ser melhoradas, injetando randomização no processo básico em conjunto com a aprendizagem de máquina (*Machine Learning*), por um método chamado *Random Forests* (BREIMAN, 2001). *Random Survival Forest* (RSF) é estreitamente modelado após a abordagem de Breiman. Em *Random Forest*, a randomização é introduzida de duas formas. Primeiro, uma amostra de *bootstrap* desenhada aleatoriamente faz com que os dados sejam usados para o crescimento da árvore (ramos ou galhos). Por sua vez, a árvore de decisão é criada dividindo os nós em preditores selecionados aleatoriamente. Embora à primeira vista a *Random Forest* possa parecer um procedimento incomum, considerável evidência empírica mostrou que é altamente eficaz. Duas características que merecem destaque são: primeiro é fácil de usar, os parâmetros precisam ser definidos com número de preditores selecionados aleatoriamente, o número de árvores cultivadas na floresta e a regra de divisão a ser usada; segundo é altamente adaptável aos dados e o modelo é livre de pressupostos, ou seja, é especialmente útil na análise de sobrevivência, pois as análises padrões frequentemente dependem de hipóteses restritivas, como exemplo a de risco proporcional no modelo semi-paramétrico.

Além disso, com tais métodos há sempre a preocupação se associações entre preditores e os riscos foram modelados apropriadamente, e se os efeitos não-lineares ou interações de ordem superior para os preditores devem ou não ser incluídos. Por outro lado, tais problemas são tratados de forma transparente e automática dentro de uma abordagem em *Random Forest*. Nota-se que é necessário um método abrangente com um software de acompanhamento. Neste sentido, existe um pacote o *randomForestSRC* do *software R* para implementar *Random Survival Forest*. O pacote *randomForestSRC* (ISHWARAN e KOGALUR, 2014) é um tratamento unificado de *random forest* para problemas de sobrevivência, regressão e classificação (EHLINGER, 2016), sendo descrito pelos seguintes passos:

**1º passo:** Obtenha $n$ amostras de *bootstrap* formando $n$ árvores a partir dos dados originais;

**2º passo:** Cultive uma árvore para cada conjunto de dados amostrado. Em cada nó da árvore selecione aleatoriamente preditores (covariáveis) de divisão. Assim, usando um critério de divisão de sobrevivência onde o nó é dividido nesse preditor que maximiza diferenças de sobrevivência entre nós-filhas;

**3º passo:** Cresça a árvore até o tamanho máximo sob a restrição que um nó terminal não deve ter um tamanho menor do que nó de mortes únicas;

**4º passo:** Calcule uma estimativa de risco acumulado combinando informações das $n$-árvores. Uma estimativa para cada indivíduo nos dados é calculada;

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

As regras de divisões de nó são fatores cruciais para o algoritmo. Diante disso, o pacote *Random Survival Forest SRC* fornece duas regras diferentes de divisão de sobrevivência para o utilizador. Estes são:

(i) Uma regra de divisão de *logrank*, a regra de divisão padrão, chamada pela opção *splitrule = "logrank"*;

(ii) Uma regra de pontuação *logrank*, *splitrule = "logrankscore"*;

**Estatísticas de Desempenho**

Tradicionalmente, no ramo de *machine learning*, a precisão da predição é fundamental para validar certos modelos. Em dados de sobrevivência com censura à direita, tal predição para as medidas padrões são avaliadas conforme a área dos tempos dependentes de acordo com a característica operacional, como também para os tempos dependentes em *Brier score*. Desse modo, quando o objetivo é avaliar a perfomance de predição em *Random Forest*, uma possibilidade é utilizar os dados *Out-Of-Bag*(OOB). Por outro lado, quando o intuito é comparar a performance de modelos *Random Forests* com modelos de predição em análise de sobrevivência (RSF), faz-se necessário realizar um *loop* de validação cruzada para estimar a precisão da predição. Neste sentido, a estatística *C-index* é utilizada nos dados *Out-Of-Bag*, ao qual visa avaliar a precisão no conjunto de dados treino. Por sua vez, a estatística de *Brier score* (*IBS*) é aplicada no conjunto de dados teste para avaliar a qualidade da predição.

**C-Index**

Seja $(T_{1,h};\delta_{1,h}),(T_{2,h};\delta_{2,h}),...,(T_{n,h};\delta_{n,h})$ os tempos de sobrevivência e os status de censura para os $n$ indivíduos em um nó terminal $h$. Além disso, seja $t_{1,h} < t_{2,h} <...< t_{m,h}$ o $m$-ésimo tempo de evento distinto em um nó terminal $h$. Em seguida, defina $d_{1,h}$ e $Y_{1,h}$ como o número de mortes e indivíduos em risco no tempo $t_{1,h}$. Desse modo, a função de risco acumulada para estimar o nó terminal $h$ é o estimador de Nelson-Aalen definido como

$$\hat{H_h}(t) = \sum_{t_{1,h} \leq t} \frac{d_{1,h}}{Y_{1,h}} \tag{6}$$

para o indivíduo $i$ com uma covariável $x_i$ $d$-dimensional tem-se que

$$H(t|x_i) = \hat{H}_h(t), \text{ se } x_i \in h. \tag{7}$$

Por sua vez, nos procedimentos em RSF, para estimar a função de risco acumulada para o indivíduo $i$, se considera $I_{i,b} = 1$ caso $i$ pertença à OOB, na $b$-ésima amostra via *bootstrap*, Caso contrário, $I_{i,b} = 0$. Além disso, para a árvore gerada na $b$-ésima amostra via *bootstrap*, se considera $H_b(t|x_i)$ a função de risco acumulada para o indivíduo $i$. Desse modo, a função de risco acumulada conjunta para $i$ é,

$$H^*(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}. \quad (8)$$

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

$$\sum_{l=1}^{m} H(t_l|x_i) > \sum_{l=1}^{m} H(t_l|x_j), \quad (9)$$

em que, $t_1 < t_2 <...< t_m$ são os tempos de eventos únicos no conjunto de dados.

Neste sentido, em RSF, a função de risco acumulada($H^*(T|X)$) conjunta é mais utilizada do que $H(T|x)$. Um valor de 0,5 para a estatística *C-index* não é considerado melhor do que, por exemplo, um resultado de cara e coroa. Por outro lado, o valor 1 indica uma habilidade discriminatória completa.

**IBS**

O *Brier score* até o tempo $t$ é definido por,

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{(0 - \hat{S}(t|x_i))^2}{\hat{G}(t_i)} I(T_i \leq t, \sigma_i = 1) + \frac{(1 - \hat{S}(t|x_i))^2}{\hat{G}(t)} I(T_i > t) \right\} \quad (10)$$

onde $\hat{G}(t) = P(C_i > t)$ é o estimador de *Kaplan-Meier* para a função de sobrevivência.

A curva do erro de previsão é obtida a partir do *Brier Score* através dos tempos. Desse modo, o *Brier Score* integrado é a curva do erro de previsão acumulada sobre o tempo, e é dada por,

$$IBS = \frac{1}{max(t_i)} \int_0^{max(t_i)} BS(t)dt. \quad (11)$$

Neste sentido, quanto mais próximo de 1 for o IBS, pior é a capacidade preditiva do modelo avaliado. Por outro lado, valores próximos de 0 indicam que o modelo tem uma melhor capacidade preditiva.

# Materiais e Métodos

O conjunto de dados utilizado neste trabalho é derivado do trabalho feito por Blair et al. em 1976, na Irlanda do Norte. A base de dados contém 394 observações de 197 pacientes com retinopatia diabética que faziam um tratamento de fotocoagulação à *laser*. Para cada paciente foi aleatorizado um dos olhos para receber o tratamento e o outro olho foi tido como controle. As variáveis presentes no banco de dados são: *id* (que é uma variável identificadora do indivíduo), *olho*, *status*, *tratamento*, *idade*, *tipo de laser* e *tipo de diabetes* e *Grau de Risco* para cegueira, que é uma pontuação de risco para o olho. Este subconjunto de alto risco é definido como uma pontuação de 6 ou maior em pelo

menos um olho. É possível ter acesso à esses dados através do comando `data(rms)` no software R.

Neste trabalho, foi estimado a curva de risco acumulado de Kaplan-Meier, o teste *Logrank* para comparação entre estes riscos. Observou-se o tempo até a cegueira total de 394 pacientes com Retinopatia Diabética que foram submetidos a um tratamento à *laser* para melhoria da doença. Para tanto, inicialmente foi utilizado o estimador limite-produto de Kaplan-Meier, teste *Logrank* e o modelo de regressão de Cox (Saturado) entre os grupos estudados eq. (12), todas as inferências feitas ao nível de 5% de significância.

$$\lambda(t|X) = \lambda_0(t) \exp x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + ... + x_{10}\beta_{10} \quad (12)$$

em que, $x_1$:Tipo de *Laser* (0: Argon; 1: Xenon); $x_2$:Olho (0: direito; 1: esquerdo); $x_3$:Tipos de Diabetes (0: Juvenil; 1: Adulto); $x_4$:Tratamento(0: Controle; 1:Tratado); $x_5$:Grau de Risco (6 à 12); $x_6$: Idade (1 à 58); $x_7 = x_1x_3$: Tipo de Diabetes × Tipo de Laser; $x_8 = x_3x_4$: Tipo de Diabetes × Tipo de Laser; $x_9 = x_2x_3$: Olho × Tipo de Diabetes; $x_{10} = x_1x_2$: Tipo de Laser × Olho.

Realizou-se um estudo diagnóstico via resíduos de Shoenfeld, Deviance e Escore. O critério de informação de Akaike (AIC), foi utilizado para selecionar a melhor configuração de covariáveis que explicassem o risco de cegueira total por retinopatia diabética.

Após a escolha do modelo de Cox, selecionou-se diferentes modelos ajustados via *Random Surivival Forest* com diferentes critérios de escolha entre os modelos estudados. Os modelos ajustados via RSF foram: (M1) RF - Saturado, com a presença de interação entre Tratamento e Tipo de Diabetes; (M2) RF - Completo, com a presença de todas as covariáveis do banco de dados; (M3) RF - Cox, modelo com a configuração do modelo de Cox ajustado neste trabalho; (M4) RF - Corrente, modelo com as variáveis preditoras ajustadas segundo a escolha via Fator de variância importância e o Modelo (M5) modelo de Cox clássico ajustado via verossimilhança parcial, na configuração selecionada via critério de AIC e análise diagnóstica dos resíduos.

Foram consideradas 1000 amostras bootstrap para compor as florestas aleatórias, e utilizou-se a divisão do conjunto de dados em `Treino` 70% e `Teste` 30%, além do uso das amostras *Out-Of-Bag*, para compor a análise de precisão do *Random Survival Forest*.

As análises foram feitas por meio do software estatístico R na versão 3.5.0, de uso livre e amplamente utilizado em âmbito acadêmico, abrangendo também a modelagem via técnicas de *Machine Learning* com o algoritmo do tipo *Random Forest*, nos pacotes `randomForestSRC`, `pec` e as análises clássicas pelos pacotes `survival`, `muhaz` `rms`.

# Resultados e Discussão

O banco de dados compreende um total de 394 observações referentes à 197 pacientes avaliados. No total 155 olhos experimentaram o evento de interesse (cegueira devido à retinopatia diabética). Na Tabela 1, tem-se as informações descritivas para as variáveis independentes coletadas e as estatísticas obtidas por meio do estimador de Kaplan-Meier e na Figura 1 tem-se os gráficos para as variáveis contínuas e para a variável Grau de Risco nos grupos de indicadora de Falha/Censura.

O número de eventos para o grupo controle foi praticamente o dobro quando comparado ao número de eventos do grupo tratado. O que significa dizer que, durante o período de estudo o número de pacientes que cegaram para o grupo controle foi consideravelmente

maior que do grupo tratado. O que nos leva a entender que provavelmente o tratamento efetuado nos pacientes parece exercer um resultado positivo a cerca da cegueira obtida por meio de Retinopatia Diabética.

Tabela 1: Resumo estatístico das informações e estimativas de Kaplan-Meier para as variáveis do banco de dados.

| Variáveis Preditoras | Níveis | Número (porcentagem %) | Eventos | Mediana | LI | LS |
|---|---|---|---|---|---|---|
| Laser | Xenon/Argon | 200 (50,76)/194 (49,24) | 76/79 | NA/NA | 58,8/43,7 | NA/NA |
| Olho | Esquerdo/Direito | 178(45,17)/216(54,83) | 60/95 | NA/61,8 | NA/46,2 | NA/NA |
| Tipo | 1/2 | 228(57,86)/166(42,14) | 87/68 | NA/63,3 | 48,9/48,3 | NA/NA |
| Tratamento | Controle/Tratado | 197(50)/197(50) | 101/54 | 43,7/NA | 31,6/NA | 59,8/NA |
| Censura/falha | 0/1 | 239(60,66)/155(39,34) | NA | NA | NA | NA |

Observa-se que não foi possível obter todos os valores para o tempo mediano de sobrevivência dos grupos em estudo. Isso ocorreu porque neste estudo não foi atingido este tempo mediano na quase totalidade dos grupos, ou seja, o estudo terminou sem que 50% dos pacientes tenham experimentado o evento de interesse (cegueira), dentro de cada grupo. Consequentemente não foi possível realizar o cálculo do intervalo de confiança para essa medida, nesta situação. Porém, nota-se que, o tempo em que 50% dos pacientes do grupo controle cegaram foi de 43,7 meses com o intervalo de confiança (31,6; 59,8). As casas decimais obtidas para o tempo em meses deve-se a escala de tempo utilizada para a observação desses pacientes no estudo em que, provavelmente foi considerado as semanas.

Os resultados obtidos via histogramas (Figuras $1a_1$ e $1a_2$), demonstram que neste estudo os indivíduos falharam em grande quantidade até pouco mais de 10 meses e que o percentual de censura foi mais alto a partir dos 35 meses. A diferença entre a mediana de grau de risco para a censura e a falha (Figura 1b) para quem tinha pelo menos um olho em alto risco foi de um grau, indicando que um maior grau de risco afeta no risco de retinopatia diabética.

Com relação a idade de diagnóstico de diabetes, Figura 1c, percebe-se que tanto para a idade de forma geral, quanto para ela estratificada por falha e censura o comportamento foi o mesmo, com uma possível sugestão de dois picos um por volta dos 10 anos e outro por volta dos 45 anos, porém a maior concentração foi até os 20 anos de idade.

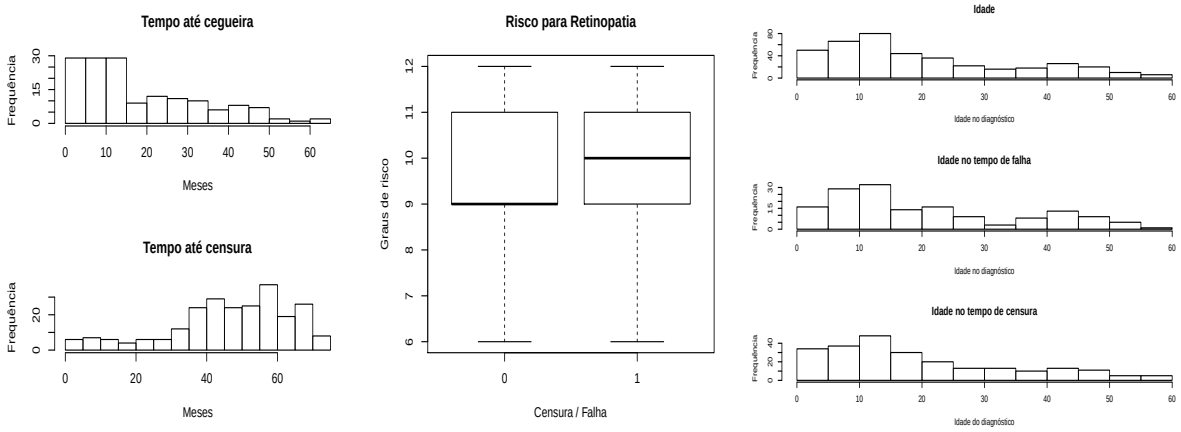


Figura 1: Histograma para o tempo: até a cegueira(a.1) e censura (a.2). Boxplot para o grau de risco de retinopatia diabética em ao menos um olho (b). Histograma para a idade: no diagnóstico (c.1), estratificada no tempo de falha (c.2) tempo de censura (c.3).

O estimador de Kaplan-Meier é o mais utilizado na literatura em dados de sobrevivência para caracterizar curvas de sobrevivência e risco. Neste sentido, pela Figura 2 observa-se as curvas de risco acumulado obtidas por meio do estimador de Kaplan-Meier para os grupos em estudo. Em que, é possível observar a diferença do tempo de sobrevivência entre esses grupos. Dessa forma, é possível corroborar o que foi exposto na Tabela 1 a cerca do tempo mediano de sobrevivência dos pacientes do grupo tratado. Percebe-se que, o risco de cegueira no grupo tratado é menor do que no grupo controle, ou seja, o grupo que não fez uso do tratamento tem uma probabilidade maior de cegar, e assim o tratamento à *laser* surtiu efeito nos pacientes.

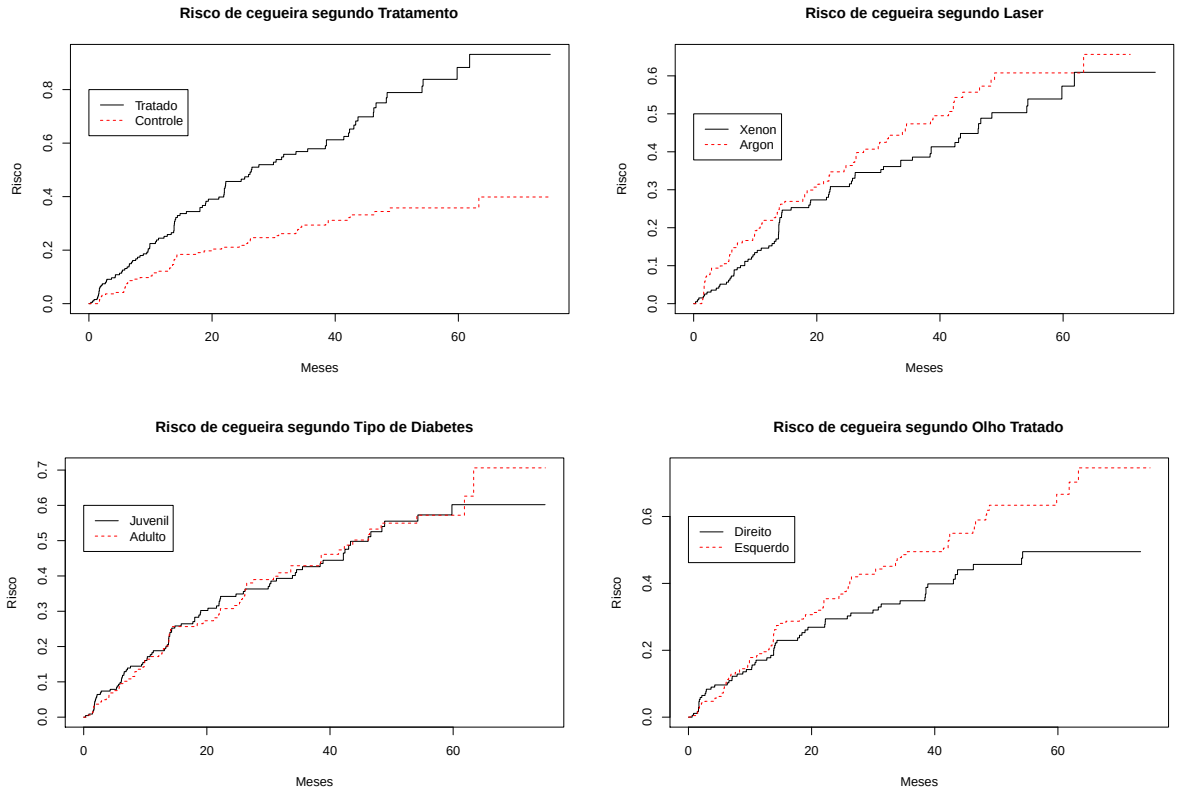


Figura 2: curvas de risco acumulado obtidas por meio do estimador de Kaplan-Meier para os grupos em estudo.

Para verificar se existe diferença significativa entre as curvas de sobrevivência entre o grupo tratado e o grupo controle utilizou-se o teste *logrank* partindo do pressuposto de que os riscos entre os grupos são proporcionais e obteve-se um valor P $< 0{,}001$. Deste modo, rejeitou-se a hipótese nula e concluiu-se que as curvas de sobrevivência entre os grupos diferem significativamente.

É também importante saber se há diferença nas curvas de sobrevivência entre pacientes que tinham Diabetes do Tipo 1 e Diabetes do Tipo 2 (Figura 2c) e também se a sobrevivência dos pacientes do grupo tratado que utilizavam certo tipo de *laser* (argon ou xenon) diferem significativamente. Foi utilizado o teste *logrank* para este fim. Para o primeiro caso foi obtido um valor P $< 0{,}001$, levando a conclusão de que o tempo até a cegueira total de pacientes com Diabetes do Tipo 1 e do Tipo 2 são diferentes significativamente, a Retinopatia Diabética se mostrou mais agressiva nos pacientes que tinham

Diabetes do Tipo 1 do grupo tratado. Já do grupo controle, os pacientes que tinham Diabetes do Tipo 1 foram mais favorecidos, obtendo probabilidades de sobrevivência maiores que dos pacientes com Diabetes do Tipo 2. Quanto ao resultado obtido para o tipo de *laser* utilizado, obteve-se um valor P $<$ 0,001, estatisticamente significativo. O que nos faz concluir que a sobrevivência entre pacientes que utilizavam os *lasers* argon ou xenon diferem entre si, ou seja, o *laser* xenon é mais eficiente quanto a prevenção à cegueira de pacientes com Retinopatia Diabética.

Outra observação que se desejou fazer neste estudo é se há diferença entre os tempos de sobrevivência de pacientes que faziam tratamento no olho direito ou no esquerdo (Figura 2d). E assim, utilizou-se o gráfico de risco acumulado via Kaplan-Meier, é possível observar que o risco de cegar nos pacientes que faziam tratamento no olho direito é menor que dos pacientes que faziam tratamento no olho esquerdo independentemente do tipo de *laser* utilizado, isto provavelmente se deve a um maior desenvolvimento deste olho em relação ao esquerdo (olho dominante).

Após observadas as diversas características expostas nas estatísticas descritivas a cerca do tempo de sobrevivência até a cegueira total de pacientes com Retinopatia Diabética e que faziam determinado tratamento, verificou-se a influência de covariáveis na cegueira destes pacientes, para tanto, foi-se em busca de um modelo que pudesse descrever estes dados. O ajuste do modelo completo com o efeito dos fatores: Riscos, Tratamento, Diabetes e efeitos de interações dois à dois, realizou-se a seleção *stepwise* afim de se escolher o melhor modelo. Após checagem via AIC, o modelo da Tabela 2 foi o modelo selecionado, o qual inclui os efeitos de Risco, Tratamento, Tipo de Diabetes e a interação entre Tipo de diabetes e Tratamento. Na Tabela 2, tem-se as estimativas obtidas para cada covariável em estudo e seus respectivos erros padrão.

Tabela 2: Estimativas do modelo de Cox para Retinopatia Diabética.

| Covariaveis | Coef. | R.R. | E.P. | Valor Z | Valor P |
|---|---|---|---|---|---|
| Tratamento à *laser* | -0,411 | 0,663 | 0,218 | -1,889 | 0,059 |
| Risco | 0,153 | 1,166 | 0,056 | 2,722 | 0,006 |
| Diabetes Tipo 2 | 0,371 | 1,449 | 0,200 | 1,860 | 0,063 |
| Tratamento× Tipo2 | -0,882 | 0,414 | 0,351 | -2,512 | 0,012 |

R.R.: Razão de Riscos - exp(Coef.); E.P.: Erro Padrão.

Percebe-se que todas as variáveis foram significavas ou marginalmente significativas para o modelo ao nível de 5% de significância (à 10%), sendo as variáveis Tratamento à *laser* e Diabetes Tipo 2 marginalmente significativas (significativas à 10% apenas), porém como foram selecionadas pelo critério de informação de Akaike (AIC) e por fazerem parte do efeito de interação, optou-se assim por mantê-las. Veja que a cada unidade da variável Tratamento à *laser* há um aumento de 0,41 meses no tempo até a cegueira total dos pacientes, ou em outras palavras, uma redução de 66% no risco de cegueira para os pacientes que faziam tratamento e tinha Diabetes do Tipo 2 (efeito de interação). E que a cada unidade de aumento do fator Grau de Risco, há um maior risco de acontecer à cegueira (16% mais risco de cegueira).

Após o ajuste do modelo, partiu-se para a etapa de análise dos resíduos com o objetivo de investigar a consistência nas estimativas dos coeficientes do modelo ajustado. Dessa forma, a Figura 3 tem-se os resíduos de *Schoenfeld* obtidos para cada uma das covariáveis abordadas no modelo.

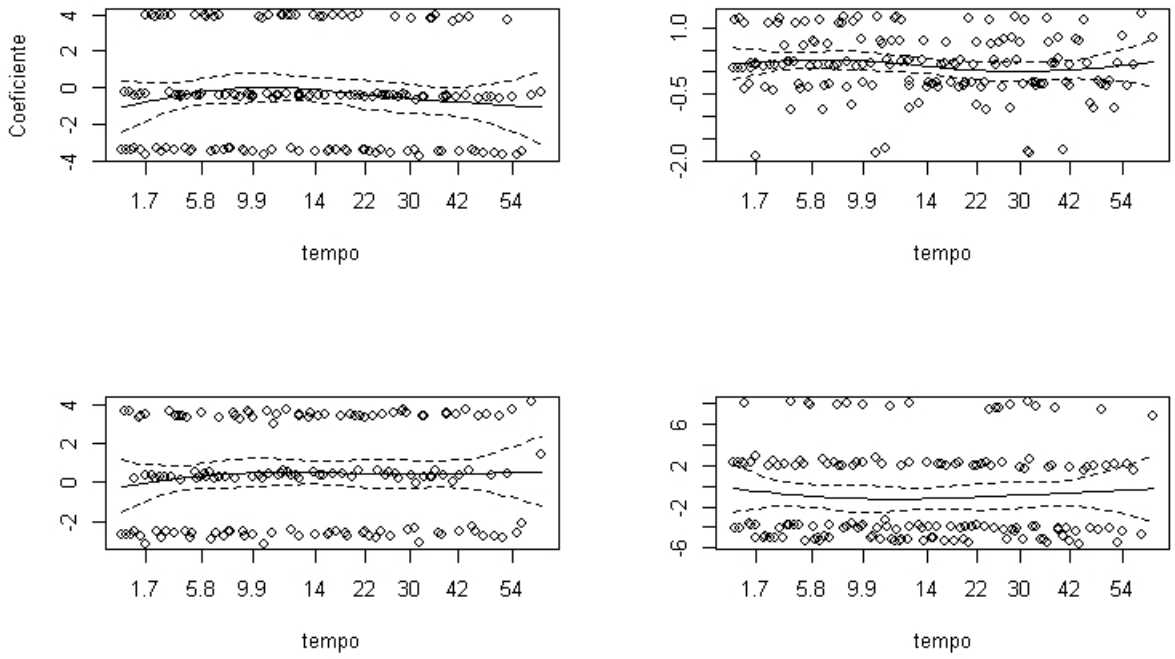


Figura 3: Gráfico dos resíduos de Shoenfeld para os dados de Retinopatia Diabética.

Os resíduos de *Shoenfeld* servem para verificar a proporcionalidade dos riscos entre os grupos, pela análise da Figura 3 indica se que o modelo de sobrevivência de Cox selecionado se ajusta a esses dados. Pois, é possível ver um resultado intuitivo da proporcionalidade dos risco entre os grupos. Nesta figura espera-se que haja uma linha reta em torno do zero indicando proporcionalidade entre as curvas de risco, também deve observar se não há nenhuma "tendência"nos resíduos, tal tendência não se verifica com a inspeção gráfica. E conclui-se que os riscos entre os grupos são proporcionais.

É possível observar uma pequena variabilidade entre os resíduos Deviance das quatro variáveis estudadas (Figura 4). Ver-se que, o valor máximo de resíduos atingindo por todas as covariáveis não ultrapassou o valor 1, indicando que o modelo gerou boas estimativas com baixos resíduos. Todas as variáveis atingiram praticamente o valor 0 para representar a mediana. Em geral, os resíduos obtidos para o modelo ajustado não apresentaram grandes valores ou valores atípicos, não violando as pressuposições necessárias para validação do modelo, o que indica que, o mesmo gerou boas estimativas (Figura 4). Para que se fosse possível analisar se existem valores influentes que possivelmente estariam causando distúrbios nas estimativas dos coeficientes, gerou-se um gráfico com esses valores para checar essa influência sob as estimativas. Assim, a Figura 5 vê-se os resultados obtidos neste seguimento, por meio dos resíduos escore.

No gráfico da Figura 5, são apresentados as observações que obtiveram os maiores valores para os resíduos, com destaque para a observação 98. Entretanto, fica claro, na obervação dessa figura de influência global que mesmo tendo a observação 98 obtido um valor muito acima da média dos demais, este valor não é de alta influência para a estimação dos coeficientes do modelo, tendo portanto um pequeno desvio. Os pontos 98,

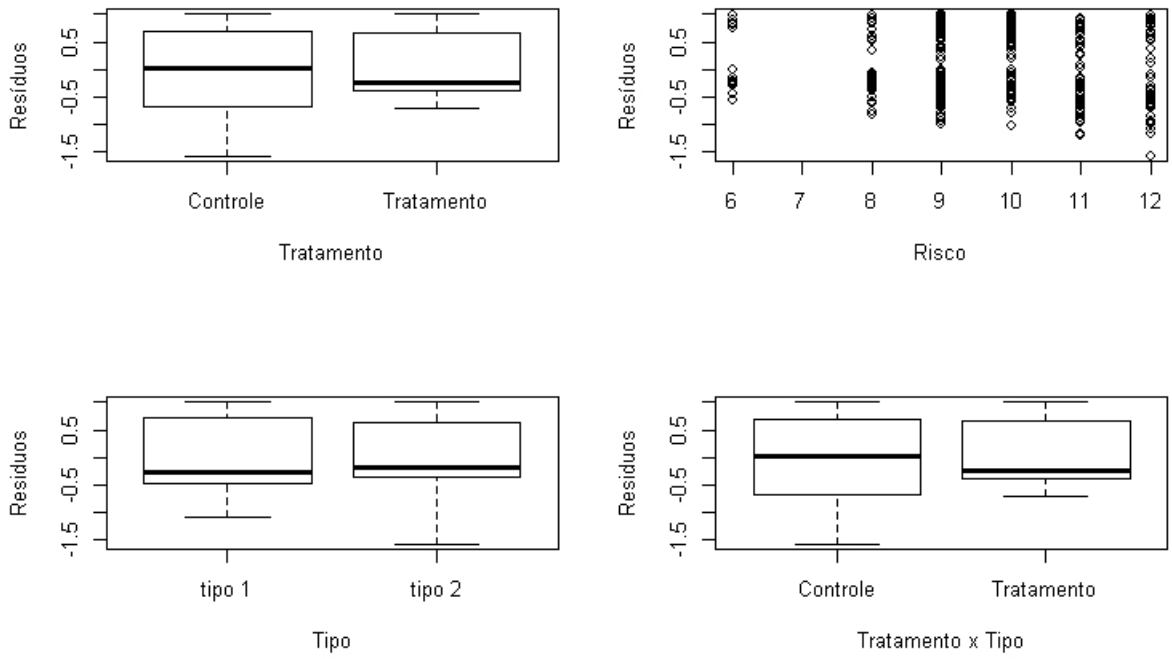


Figura 4: Resíduos Deviance do modelo de Cox.

190, 14, 27, 369, identificados no gráfico da Figura 5, foram retirados e ao ser realizado o modelo sem estes pontos, não se percebeu modificações nos erros padrão das estimativas dos parâmetros, no valor de AIC dos modelos sem estes pontos, e nos gráficos dos resíduos deviance.

Para realizar a predição selecionou-se o melhor modelo ajustado via riscos proporcionais e denominou-se de Modelo 5 (M5), e realizou-se o ajuste via *Random Forest* para análise de sobrevivência (RSF). Os principais resultados são vistos nas Figuras de 6 à 9, bem como pelo resultado da comparação entre os modelos na Tabela 3. De inicio foi ajustado o RSF, com todas covariáveis para 1000 amostras *bootstrap* e calculou-se a taxa de erro e o fator de importancia de variância (*Variable importance*; Figura 6).

Foi verificado por meio da Figura 6, que há uma estabilização na taxa de erro de previsão com os dados OOB, por volta de 37% a partir de 800 arvores de decisão, neste conjunto de dados foi realizado o crescimento de 1000 arvores na floresta aleatória, de modo que obteve-se convergência nesta simulação. A Figura 7 (b), percebe-se que tanto a variável Olho, quanto a variável *laser*, apresentam contribuição baixa e negativa para a discriminação dos sujeitos dentro da arvore, fazendo com que se gera-se diferentes configurações de RSF: com estas variáveis e sem estas variáveis (M2 e M4).

O resumo da Tabela 3, o C-index que mede o erro de previsão foi maior no modelo M1, M2 e M3, no modelo de Cox ele foi otimizado em relação aos demais (M5), isto aplicado ao conjunto de treino, porém quando se olha a capacidade preditiva do modelo de Cox no conjunto de teste, percebe-se que para os dados de teste o desempenho pior foi para este modelo (M5), com uma das mais alta variações (19,34%) e menor variação para o modelo completo (M2).

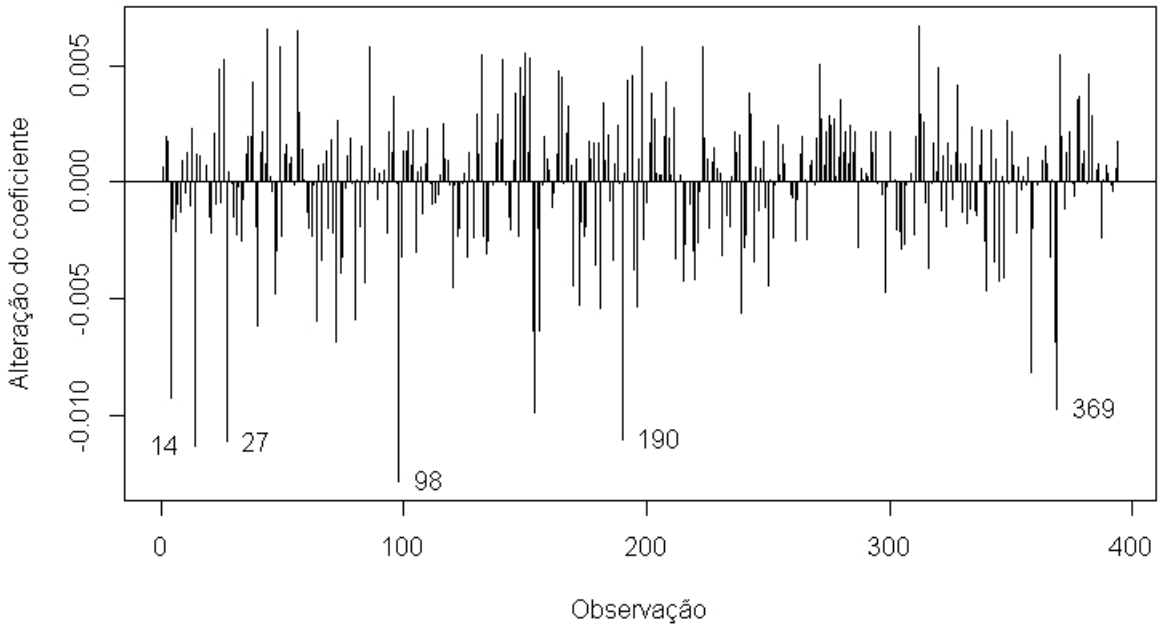


Figura 5: Gráfico de valores influentes para o modelo de Cox.

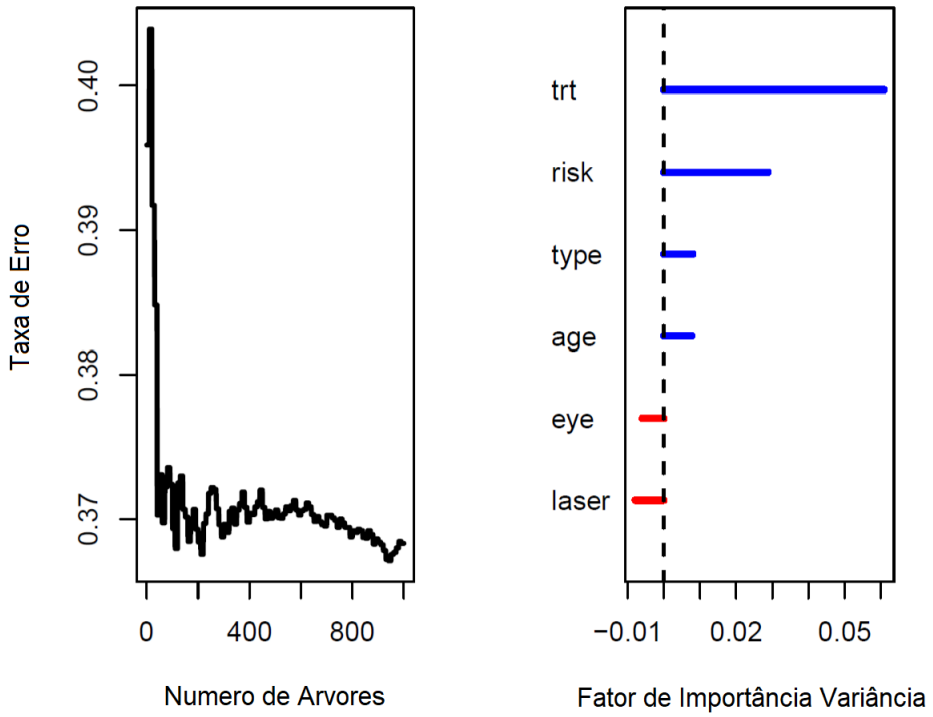


Figura 6: Taxa de erro e Fator de importancia da variável obtidas via RSF.

Tabela 3: Desempenho de diferentes configurações RFS e modelo de Cox clássico utilizando o conjunto de dados de treino e teste.

| Modelos Analisados | C-index | | | IBS | | |
|---|---|---|---|---|---|---|
| | Dados Treino | Dados Teste | Porcentagem % | Dados Treino | Dados Teste | Porcentagem % |
| M1: RF - Saturado | 0,6390 | 0,5590 | 14,18 | 0,196 | 0,205 | 4,59 |
| M2: RF - Completo | 0,6316 | 0,5721 | 10,40 | 0,198 | 0,203 | 2,52 |
| M3: RF- Cox | 0,6385 | 0,5413 | 17,95 | 0,190 | 0,214 | 12,63 |
| M4: RF - Corrente | 0,6477 | 0,5111 | 26,72 | 0,197 | 0,217 | 10,15 |
| M5: Cox Clássico | 0,6447 | 0,5402 | 19,34 | 0,188 | 0,213 | 13,29 |

Em relação ao IBS, houve uma piora em todos os modelos quando se passou do conjunto de treino para o conjunto de teste, sendo a menor variação para o modelo RF-completo (M2) (2,52%). Neste sentido, apesar de o RF- Saturado(M1), RF-Cox (M3) e Cox Classico (M5) terem apresentado o menor erro, quando aplicado ao conjunto de treino o modelo completo (M2) apresenta o melhor desempenho na variação entre conjunto de treino e teste, além de um ganho de 6% na variação entre o modelo de Cox (M5) no C-index para o conjunto de teste e de 6% no conjunto de teste no IBS, indicando ser o melhor modelo para realizar previsão.

A Figura 7, os valores do erro de predição por Brier Score para os quatro modelos RSF *versus* o modelo de Cox são apresentados no conjunto de teste. Percebe-se que de forma geral os modelos de *Random Forest* (M1,M2,M4,M5) tem desempenho melhor que o modelo de Cox selecionado por meio do método de *stepwise* nas quatro figuras (M3). O modelo de riscos proporcionais de Cox foi melhor apenas que a curva de referência que neste caso se trata da obtida pela predição via Kaplan - Meier, percebe-se que nos tempos iniciais a predição é praticamente a mesma entre a curva de referência, modelo de regressão de Cox (M3) e as configurações de *Random Survival Forest* (M1,M2,M4,M5).

O gráfico de mortalidade *versus* o tempo permite praticamente diferenciar os indivíduos que falharam valores azuis com os indivíduos que censuraram em preto, ao longo do tempo. Neste sentido, dos 4 modelos de *Random Survival Forest* ajustados apenas, o modelo completo, configuração com todas as covariáveis, (M2) é apresentado na Figura 8 para o conjunto de treino e teste com os dados OOB. Percebe-se a predição das curvas de sobrevivência de todos os indivíduos OOB, com a curva de kaplan meier centralizada, bem como um alto grau de predição registrada pela curva ascendente de Brier Score OOB, com os valores estando sempre acima do limiar 50.

A curva de sobrevivência predita via conjunto de dados de teste e taxa de falha *versus* tempo é apresentado na Figura 9. Por meio desta figura é possível ver cada individuo em teste e sua respectiva curva de sobrevida predita ao longo do tempo, nota-se uma maior predição de indivíduos que falharam (cegueira por retinopatia diabética) acima do tempo 40, indicando que RSF, tem uma capacidade de discriminação entre censura e falha bem alta a partir deste tempo.

No artigo de lançamento do pacote Random Survival Forest (RSF package) de Ishwaran e Kogalur (2007) foi apresentado a implementação da técnica em linguagem R detalhada, bem como os principais tópicos a serem trabalhados (*Prediction Error, Variance Importance*, etc.) em RSF.O método foi devidamente apresentado a comunidade científica no ano seguinte por ISHWARAN et al.,(2008) e tal método foi comparado com o modelo de regressão de Cox, com um artigo mais detalhado e com aplicações, em que se realizou uma grande gama de experimentos, desde o uso de dados reais bem como de dados simulados, para se conhecer a acurácia de predição da técnica de RSF. Foi identificado que todos os configurações de RSF, foram estáveis na presença de variáveis de ruído. E o

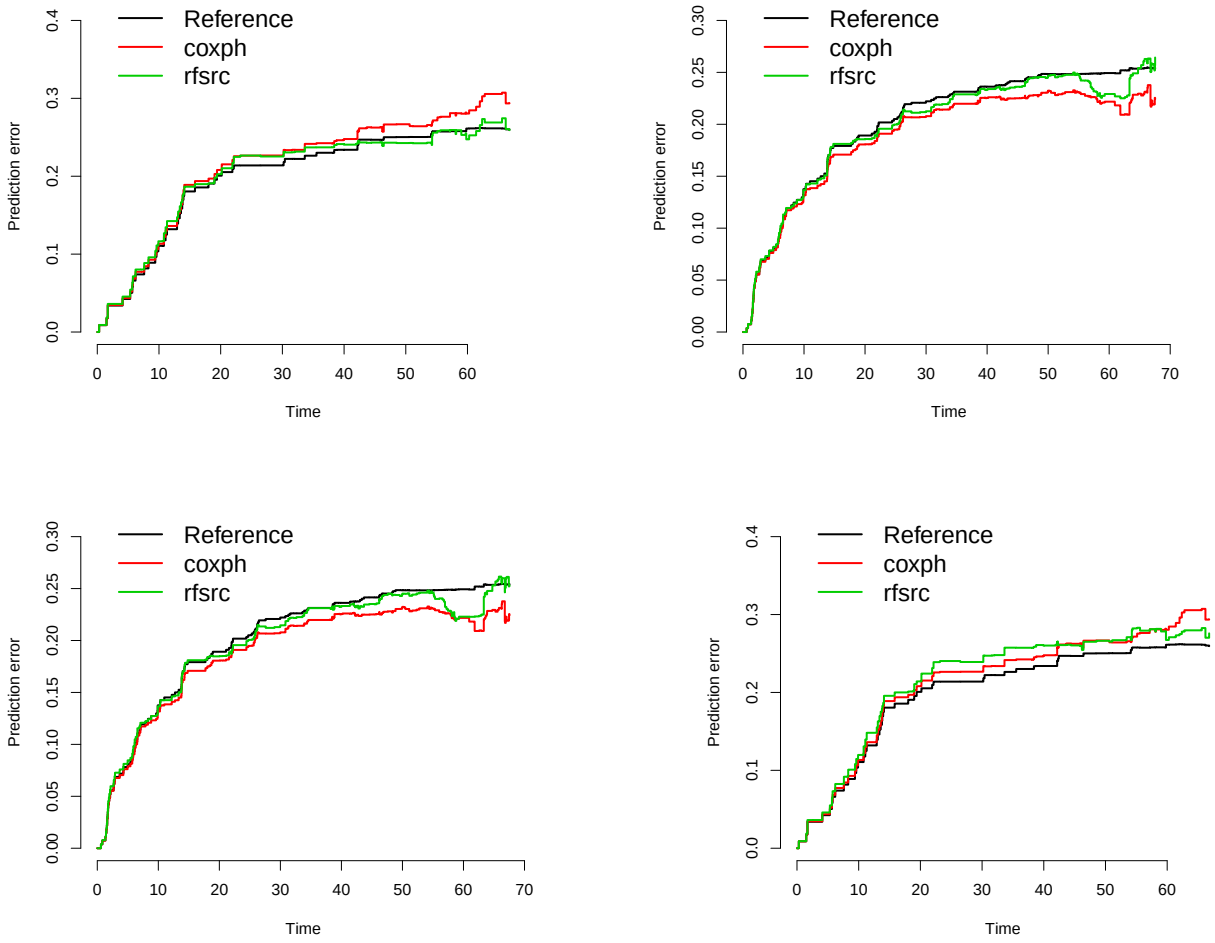


Figura 7: Valores do erro de predição por Brier Score para os quatro modelos RSF *versus* o modelo de Cox.

modelo de Regressão de Cox, em contraste, tornou-se progressivamente pior à medida que o número de variáveis de ruído aumentava, resultado também encontrado para os dados de retinopatia diabética (Tabela 3 e Figura 7). Foi discutido também em ISHWARAN et al.,(2008), que o RSF tem melhor consistência, ou ao menos é tão bom quanto os melhores dos métodos. E desde que a técnica de *Random Forest* (RF) foi introduzida na comunidade de *machine learning*, tem-se dado atenção em documentar cientificamente a sua performance. Desta forma, os resultados encontrados no trabalho original de Ishwaran et al., (2008) e os aqui encontrados corroboram por confirmar o que tem se geralmente encontrado, que o RSF produz alta acurácia no conjunto de preditores. Bou-Hamad et al., (2011) utilizaram informação de 312 pacientes que sofriam de cirrose biliar primária do fígado. O número de variáveis independentes foi 12. Eles compararam o modelo de Cox, *PT, bagging trees*, e RSF, em termos de IBS. e *10- fold cross-validation* foi aplicada para acessar a performance dos modelos. Os resultados foram apresentados graficamente e sugeriram melhores resultados para RSF, seguido por *bagging*. *PT* apresentou o pior resultado. E a performance do modelo de regressão de Cox foi intermediária. Mogensen

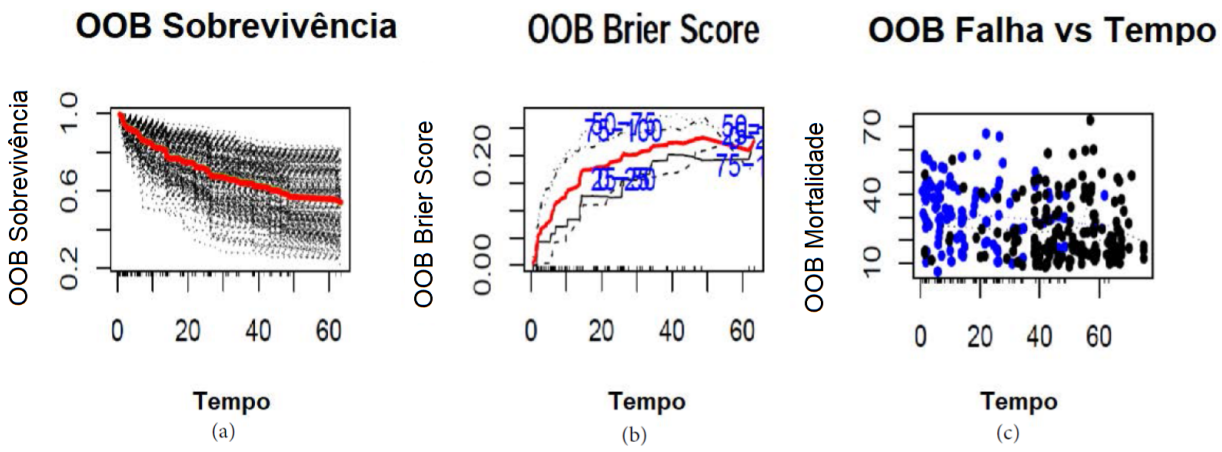


Figura 8: Sobrevivência Estimada, Indice de Brier Score e Falha *versus* tempo para o conjunto OOB no modelo (M2).

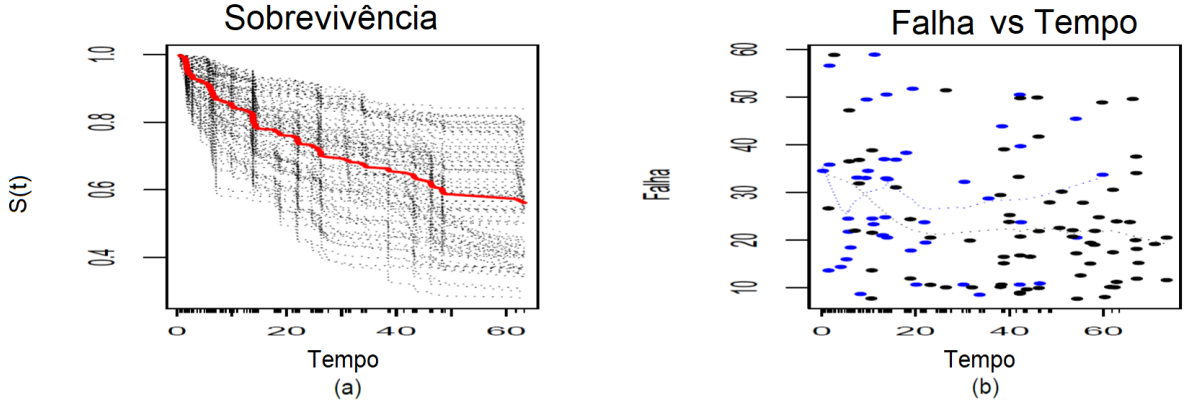


Figura 9: Curvas de sobrevivência preditas e Gráfico de taxa de falha *versus* tempo, para o modelo 2 - RF - Completo.

et al., (2012) avaliaram *Random Forest* para Análise de sobrevivência (RSF) utilizando Curvas de erro de predição, os autores utilizaram dados de acidente vascular encefálico com o uso do pacote `pec` para comparar *Random Forest* e o modelo de regressão de Cox obtido por meio da seleção de *stepwise*. Os resultados demonstraram que para 1000 amostras *bootstrap* o RSF foi melhor na acurácia de predição que o modelo de regressão de Cox, resultado que também foi encontrado neste estudo, com a ressalva que neste estudo aqui apresentado, foi considerado diferentes configurações de variáveis independentes para o RSF. Yosefian, Farkhani, Baneshi (2015), compararam diferentes métodos aplicados a análise de sobrevivência (*Saturated tree*, *Pruned tree* e RSF), aplicando a um banco de dados de 607 pacitentes com infarto agudo do miocárdio, atendidos no Imam Reza Hospital Mashhad no Iran em 2007, por meio da comparação o RSF teve melhores valores de Brier Score (IBS) e C-index, obtendo uma melhor performance de previsão para o método de RSF em relação aos demais métodos. Os autores recomendaram também o uso de conjunto de dados de treino e teste para melhorar a performance de previsão do RSF.

Rytgaard e Gerds (2018), discutem os aspectos teóricos do RSF, com uma ampla explanação dos principais tópicos (*Algorithms and Implementations Variance Importance Fator, Predictive Accuracy*) e como a técnica de *machine learning* no RSF é uma alternativa aos métodos clássicos de regressão de Cox (*risk competitives*).

# Conclusão

O modelo de Cox ajustado para os dados de retinopatia, indicaram que as covariáveis Tratamento, Grau de Risco, Tipo de Diabetes e a interação Tratamento e Tipo de Diabetes. Observou-se também que pacientes que tinham Diabetes Tipo 2 e faziam o tratamento à *laser* tinham um aumento no tempo até a cegueira, um aumento na probabilidade de sobrevivência. O que também foi mostrado no teste *logrank* e que o acréscimo de uma unidade no Grau de Risco, aumentava em 16% o risco de cegueira. Porém o melhor modelo de Cox ajustado aos dados de retinopatia diabética, apesar de trazer luz aos fatores de prognóstico que afetam o risco de cegueira total, não foi o mais adequado para realizar predições. Os modelos RSF tiveram maior acurácia na predição dos valores observados em contraste com o modelo de Cox convencional. Assim, os resultados indicaram que o uso dos modelos de *Random Forest* revelaram a importância de todas as covariáveis para a predição (M2). Sendo assim, fica claro neste trabalho a importância de considerar para a predição a correta configuração do modelo, ajustando o com a presença de todas as covariáveis.

# Agradecimentos

O presente trabalho foi realizado com o apoio da Pró-reitoria de Pós-graduação e Pesquisa - PRPGP da Universidade Estadual da Paraíba - UEPB, por meio de projeto aprovado no Edital PROPESQ 2017.

# Referências bibliográficas

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

CARVALHO, M.S.; ANDREOZZI, V.L.; CODEÇO, C.T.; BARBOSA, M.T.S.; SHIMAKURA, S.E. *Análise de sobrevida: teoria e aplicações em saúde.* Rio de Janeiro: Editora Fiocruz, 2005.

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

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

ISHWARAN, H.; KOGALUR, U.B. Random survival forests for R. *R news*, v.7, n.2, p.25-31, 2007.

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

ISHWARAN, H; KOGALUR, U. B. Random Forests for Survival, Regression and Classification (RF-SRC). *R package version 1.6.*"URL http://CRAN.R-project.org/package= randomForestSRC, 2014.

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

BOU- HAMAD, I., LAROCQUE, D., BEN-AMEUR, H. A review of survival trees. *Statistics Surveys*, v.5, p.44-71, 2011.

*Sociedade Brasileira de Diabetes:* Tipos de Diabetes. Disponível em: <http://www.diabetes.org.br/publico/diabetes/tipos-de-diabetes>. Acesso em: 09 de abr. de 2019.

*Sociedade Brasileira de Endocrinologia e Metabologia:* O que é Diabetes. Disponível em: <https://www.endocrino.org.br/o-que-e-diabetes/>. Acesso em: 09 de abr. de 2019.

RYTGAARD, H.C.; GERDS, T.A. Random Forests for Survival Analysis. *Wiley StatsRef: Statistics Reference Online*, p.1-8, 2018.

R CORE TEAM. *R: A language and environment for statistical computing.* **R Foundation for Statistical Computing**, Vienna, Austria. 2019. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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