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

SÓSTENES JERÔNIMO DA SILVA

PREDIÇÃO DOS TEMPOS ATÉ A MORTE DE MULHERES COM CÂNCER DE MAMA VIA *RANDOM SURVIVAL FOREST*

CAMPINA GRANDE - PB
2022

SÓSTENES JERÔNIMO DA SILVA

PREDIÇÃO DOS TEMPOS ATÉ A MORTE DE MULHERES COM CÂNCER DE MAMA VIA *RANDOM SURVIVAL FOREST*

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

Orientador: Prof. Dr. Tiago Almeida de Oliveira

CAMPINA GRANDE - PB
2022

S586p Silva, Sóstenes Jerônimo da.

Predição dos tempos até a morte de mulheres com câncer de mama via *Random Survival Forest* [manuscrito] / Sostenes Jeronimo da Silva. - 2022.

33 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2022.

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

1. Machine Learning. 2. Random Survival Forest. 3. Câncer de mama. 4. Análise de sobrevivência. I. Título

21. ed. CDD 519.535

Elaborada por Camile de A. Gomes - CRB - 15/559 BC/UEPB

SÓSTENES JERÔNIMO DA SILVA

PREDIÇÃO DOS TEMPOS ATÉ A MORTE DE MULHERES COM CÂNCER DE MAMA VIA RANDOM SURVIVAL FOREST

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

Trabalho aprovado em 22 de Julho de 2022.

BANCA EXAMINADORA

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

Prof. Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba (UEPB)

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

Ao meu pai e minha mãe, pela dedicação, companheirismo e amizade, DEDICO.

AGRADECIMENTOS

Primeiramente agradeço a Deus, pois sem Ele não chegaria aqui. Em seguida ao meu pai Izaias Jerônimo da Silva e minha mãe Aldira da Silva Santos pela compreensão em todas minhas escolhas e, além do mais, pelo suporte financeiro que propiciou esta formação.

Aos professores do Curso da UEPB, em especial, ao meu querido amigo Dr. Tiago Almeida de Oliveira que contribuiu ao longo desta formação com diversas oportunidades para o crescimento rico em técnicas aplicadas por um profissional em Estatística.

Aos colegas de classe pelos momentos de amizade e apoio. Cito alguns nomes com orgulho de tê-los como amigos que contribuíram muito nesse decorrer de tempo, Débora Cordeiro, Lucas Cardoso, Débora dos Santos, Mayara Cristina e Damião Everton.

Além destas pessoas em especial, meus profundos agradecimento a CAPES e ao Hospital da Fundação Assistencial da Paraíba (FAP), e os funcionários que me auxilaram, proporcionando a base de dados deste trabalho. Além do mais, me deram atenção nas diversas situações de dúvidas em como conduzir a coleta dos dados. Agradeço de coração a todos.

“Sem dados você é apenas mais uma pessoa com opinião”
(William Edwards Deming)

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SÍMBOLOS

A_i	i-ésima paciente do estudo
D	Direita
E	Esquerda
H	Hormonoterapia
Q	Quimioterapia
R	Radioterapia

SUMÁRIO

PREDIÇÃO DOS TEMPOS ATÉ A MORTE DE MULHERES COM CÂNCER DE MAMA VIA *RANDOM SURVIVAL FOREST*

Sóstenes Jerônimo da Silva*
Tiago Almeida de Oliveira†

RESUMO

A Análise de sobrevivência é uma área de pesquisa que cada vez mais vem ganhando ênfase em diversos setores produtivos e acadêmicos. Esta análise baseia-se em analisar dados cuja a variável de interesse é o tempo até a ocorrência de um evento, permitindo modelar dados com a presença de informações incompletas, denominadas censuras. Estes modelos podem apresentar distribuição paramétrica ou semi-paramétrica de riscos proporcionais de Cox. No entanto, tais modelos citados não possuem a melhor capacidade preditiva, sendo assim, os modelos de *Machine Learning* em conjunto com o *Randon Forest* via Análise de Sobrevivência (RSF), têm sido utilizado como alternativa para predições mais robustas. Neste trabalhou objetivou-se utilizar a técnica *Random Survival forest* com o intuito de prever o tempo que uma paciente diagnosticada com câncer leva até a morte, com base em alguns indicadores. Os resultados obtidos evidenciam uma boa capacidade preditiva do modelo, apresentando *C-index* de 0,8800 e IBS de 0,1559, no qual as variáveis de suma importância para a predição são: idade, o receptor de progesterona, número de quimioterapias, número de radioterapias e número de hormonoterapia.

Palavras-chaves: *Machine Learning*. *Random Survival Forest*. Câncer de mama. Análise de Sobrevivência.

ABSTRACT

Survival Analysis is an area of research that is increasingly gaining emphasis in various productive and academic sectors. Our study is based on analyzing data whose variable of interest is the time until the occurrence of an event, allowing the modeling of data with the presence of incomplete information, called censorship. These models can present a parametric or semi-parametric distribution of Cox proportional hazards. However, the aforementioned models do not have the best predictive capacity, so *Machine Learning* models together with *Randon Forest* via Survival Analysis (RSF) have been used as an alternative for predictions. more robust. This work aimed to use the technique *Random Survival forest* in order to predict the time that a patient diagnosed with cancer takes until death, based on some indicators. The results obtained show a good predictive capacity of the model, presenting *C-index* of 0.8800 and IBS of 0.1559, in which the variables of paramount importance for the prediction are: age, progesterone receptor, number of chemotherapies, number of radiotherapy and number of hormone therapy.

Keywords: *Machine Learning*. *Random Survival Forest*. Breast cancer. Survival Analysis

1 INTRODUÇÃO

Em países desenvolvidos, o câncer de mama é considerado o tipo de câncer mais prevalente no mundo, apresentando maiores taxas de incidência em países deste tipo. No Brasil, para o ano de 2022 estimou-se 66.280 novos casos de câncer mamário, representando um risco

* Sóstenes Jerônimo da Silva, Depto de Estatística, UEPB, Campina Grande, PB, sostenes.silva@aluno.uepb.edu.br
† Prof. Dr. Tiago Almeida de Oliveira, Depto de Estatística, UEPB, Campina Grande, PB, tiagoestatistico@gmail.com

de 43,74 casos a cada 100 mil mulheres (INCA, 2019). Além disso, trata-se da neoplasia com maiores taxas de mortalidade nas mulheres brasileiras, com estimativas para o ano de 2019 de 14,23/100 mil mulheres, taxa ajustada por idade pela população mundial.

As técnicas de Análise de Sobrevivência tem sido bastante empregadas para a avaliação da distribuição de frequência do tempo entre o diagnóstico do câncer de mama e o desfecho da doença, sendo este a cura ou óbito. Assim, essa área da Estatística lida com estudos que possui associação entre um evento de interesse e um intervalo de tempo a ser estimado (COLOSIMO; GIOLO, 2006). Outra técnica que é bastante utilizada nos estudos que buscam realizar predições das taxas de sobrevivência de um indivíduo é o *Machine Learning*.

Com o intuito de um melhor ajuste e, consequentemente, melhores informações clínicas, tem-se uma área de pesquisa que é a combinação da Análise de Sobrevivência com os métodos do *Machine Learning*, a *Random Survival Forest*. Essa técnica busca indicar modelos com melhor capacidade preditiva utilizando dados de sobrevivência censurados à direita (ISHWARAN et al., 2008).

Assim, o objetivo desse trabalho consistiu em analisar o tempo decorrido entre o diagnóstico do câncer de mama e o óbito de mulheres acometidas por este câncer de um hospital de referência em tratamento para o câncer no interior da Paraíba. Para isso, foi utilizada uma extensão da técnica de *Machine Learning* e *Random Forest* aplicados a Análise de Sobrevivência, a RSF.

2 METODOLOGIA

2.1 Base de dados

Neste trabalho foi coletado um banco de dados referente a pacientes do gênero feminino do Hospital Fundação Assistencial da Paraíba (FAP) que tiveram câncer de mama entre o ano de 2005 a 2014. O estudo foi realizado com autorização do comitê de ética (Certificado de Apresentação para Apreciação Ética-CAAE) da Universidade Federal de Campina, número 97198518.9.0000.5182. Na coleta de dados, os prontuários das pacientes que tiveram câncer de mama e realizaram tratamento no hospital, a identificação dos pacientes foi anonimizadas de acordo com a Lei de Proteção de Dados do Brasil (LGPD).

Por meio de consulta aos funcionários do setor de arquivo do hospital, em média, 200 mulheres com câncer de mama, por ano, iniciam o tratamento no hospital. A quantidade de homens com câncer de mama é muito inferior, por isso foi considerado só as pacientes do gênero feminino. Com esta informação, foi calculado que no intervalo do estudo a população seria de aproximadamente 2000 pacientes. Utilizando a técnica de amostragem aleatória simples levamos em consideração um nível de confiança de 95% e uma margem de erro de 5 pontos percentuais, em que o tamanho mínimo da amostra seria de 220 pacientes, no qual neste estudo foram coletadas 221 observações de pacientes.

As variáveis coletadas em cada prontuário médico foram: endereço do prontuário, data da primeira consulta, data da última consulta, data da morte do paciente, localidade do câncer (mama esquerda ou direita), idade, número de sessões de quimioterapia, número de sessões de hormonoterapia, número de sessões de radioterapia, receptor de estrogênio, receptor de progesterona, proteína Ki-67, proteína de tumor p53 e o C-erb-b2.

Com o intuito de aumentar o conhecimento sobre as variáveis, utilizou-se o método *feature engineering* que permite criar novas variáveis com base nas oriundas da planilha de dados. Assim, a partir das variáveis coletadas, foram originadas as seguintes novas variáveis: tempo de consulta, tempo de vida, tempo de morte, indicador de morte e terapia adjuvante. As variáveis categóricas do status mediano e médio separam os dados em dois grupos, sendo eles

abaixo e acima do tempo mediano e médio. A variável dependente do estudo foi denominada "tempo de vida", que se refere ao tempo entre o diagnóstico e o óbito da paciente com câncer de mama.

Para a execução da análise utilizou-se o *software* R Core Team (2021) na interface do RStudio na versão 4.1.0. Para a realização da Análise de Sobrevivência foi utilizado o pacote *Survival* (THERNEAU; LUMLEY, 2015). Para o desenvolvimento das técnicas *Random Survival Forest* utilizou o pacote *randomForestSRC* (ISHWARAN; KOGALUR, 2007).

2.2 Análise de sobrevivência

A Análise de sobrevivência é considerada uma área de pesquisa bastante utilizada no ramo da Saúde, com a finalidade de observar o tempo até que determinado evento de interesse venha a ocorrer. Dois termos devem ser levados em consideração neste tipo de análise, o tempo e a censura.

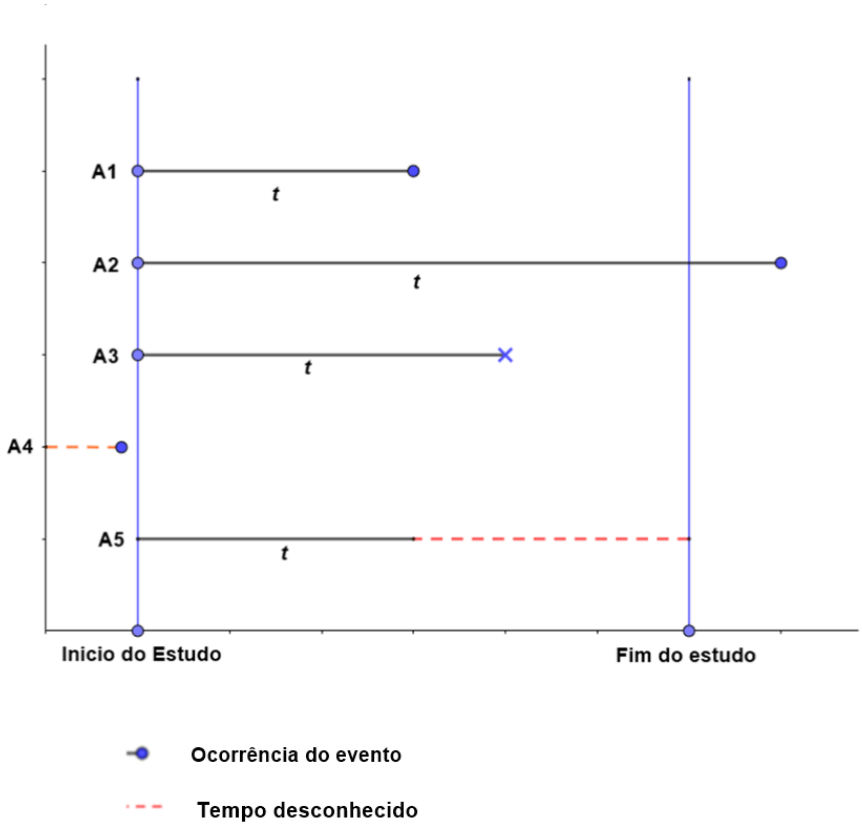
- O **tempo** t consiste na variável resposta do modelo, observado desde o início do estudo até a ocorrência do evento de interesse;
- A **censura** acontece quando há dados parciais ou incompletos. Isto ocorre quando o evento de interesse realiza-se por natureza diferente do estudo em questão. Tendo definido o início e/ou fim do estudo, dois tipos de censura podem ocorrer: a censura à esquerda e a censura à direita. A censura à esquerda acontece quando o evento ocorreu antes do início do estudo e a censura à direta ocorre quando o tempo da observação é superior ao tempo estabelecido no estudo, em que a mesma pode ser definida em três tipos.
    a) Censura tipo I: É determinado um fim para o estudo e a informação posterior será ignorada;
    b) Censura tipo II: É determinado quantos eventos precisam ocorrer e quando a meta de ocorrência é atingida o estudo é finalizado, em que os eventos posteriores são ignorados;
    c) Censura aleatória: Acontece quando um indivíduo abandona o estudo ou o evento ocorre em uma determinada natureza diferente.

Algebricamente, o tempo e a censura podem ser representados de maneira simples, de acordo com Colosimo e Giolo (2006), em que T é a variável aleatória referente ao tempo de vida de uma paciente com câncer de mama, e C independente a T, representado o status de censura ou óbito das pacientes. Logo a representação do tempo com a censura não informativa associada pode ser expressa da seguinte maneira:

$$t = \min(T, C).$$

Alguns tipos de censura são apresentados na Figura 1.

Figura 1 – Gráfico referente a representação dos tipos de censura.



Fonte: Produzido pelos autores.

A Figura 1 ilustra o tempo de vida dos pacientes até que o evento ocorra, no qual pode-se observar que os pacientes fora do intervalo do estudo são censurados, além disso, foi considerado o tempo de vida dos pacientes que estão dentro do intervalo do estudo, visto que os pacientes que não tiveram o evento de interesse representado pelo circulo azul, tem status de censura no estudo.

A variável aleatória não negativa T, representa o tempo de vida do paciente em questão, logo é possível calcular a probabilidade de um paciente sobreviver até o tempo t. Algebricamente a função de sobrevida pode ser definida na seguinte expressão:

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

Com a função de sobrevida é possível definir a função de distribuição acumulada, expressa, a seguir:

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

Para calcular o risco relativo de uma observação e descrever a distribuição do tempo de vida de pacientes, utiliza-se a função de risco definida da seguinte maneira:

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

.

O planejamento do estudo deve ser estruturado antes da aplicação aos dados, tendo em vista que o inicio do estudo precisa ser definido, assim como o final. Este tipo de analise pode

ser aplicada em diversas situações, no qual o interesse é descrever a probabilidade de ocorrência de determinada doença reincidir. Também se aplica em estudos não só na área da saúde, pois esta área é ampla, podendo ser aplicada em estudos voltados a descrever o tempo de vida útil de determinados objetos. Em suma, a análise de sobrevivência foi originada com o intuito de descrever a variável de interesse com base no seu tempo. Dentro da análise de sobrevivência, uma série de técnicas foram criadas, mesmo se tratando de uma área de estudo nova, em que uma das mais utilizadas é o estimador de *Kaplan-Meier* que foi desenvolvido a fim de mensurar os erros. Esta técnica permite comparar grupos a partir de um teste estatístico, o qual é destacado na seção, a seguir.

2.2.1 O Estimador de Kaplan-Meier

A técnica não-paramétrica desenvolvida por Kaplan e Meier (1958), que objetiva estimar a função de sobrevivência, foi desenvolvida para avaliar o segmento da observação com base no tempo. Trata-se de uma técnica bastante utilizada em estudos clínicos. Tal técnica denominada *Kaplan-Meier* diminui o erro à medida que a amostra cresce, condizendo com o Teorema do Limite Central (COLOSIMO; GIOLO, 2006). Além do mais, esta técnica permite censuras no estudo e comparações de grupos. Considerando:

- $t_1 < t_2 < \ldots < t_k$, os k com tempos distintos e ordenados de falha;
- d_j o número de falhas em t_j, $j = 1, \ldots, k$;
- n_j o número de individuos sob risco em t_j

Com isso, a função de sobrevivência para o estimador é expressa a seguir

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

A curva de *Kaplan-Meier* tem aparência de escada, em que é possível avaliar as falhas ao longo do tempo. Quando se tem por objetivo comparar as curvas das funções de sobrevivência obtidas por meio do estimador não-paramétrico *Kaplan-meier*, faz-se o uso de testes não-paramétricos como o teste de *Log-rank*, que é o mais comum em Análise de Sobrevivência como afirma Colosimo e Giolo (2006). Este teste é indicado em casos que ao se comparar grupos, a razão das funções de risco são aproximadamente constantes. De acordo com Fay e Shaw (2010), o teste de *Log-rank* é bastante eficiente para dados censurados à direita quando é válida a suposição de riscos proporcionais, além disso, quando o mesmo é da forma ponderada, concede valores constantes.

Ainda segundo Colosimo e Giolo (2006), a estatística de teste para o *Logrank* concerne na diferença obtida entre a quantidade observada de falhas de cada grupo sob estudo e o número esperado de falhas mas sob a hipótese nula de que tais curvas não diferem estatisticamente, isto é:

$$\begin{cases} H_0 = \text{as curvas não apresentam diferença significativa;} \\ H_1 = \text{as curvas apresentam diferença significativa.} \end{cases}$$

Outro estimador que faz parte da literatura de sobrevivência e surge como alternativa para o Kaplan-Meier é o estimador de *Nelson-Aalen* (NELSON, 1982), apresentando resultados semelhantes, visto que sua função de sobrevida é baseada na função de risco acumulada definida, a seguir:

$$S(t) = exp\{-\Delta(t)\}$$

O estimador apresenta resultados na probabilidade de sobrevivência iguais ou maiores ao *Kaplan-Meier* para todo tempo t de acordo com Lima (2019).

2.3 Aprendizado de Máquina (*Machine Learning*)

No campo de estudo das Ciências da Computação, o ramo da Inteligência Artificial (IA) tem recebido grande atenção e dentro desta área de pesquisa o *Machine Learning* tem se destacado. Até então tal termo tem a proposta de ensinar máquinas, ou dispositivos, a se aproximar do pensamento humano, ou seja, raciocinar, resolver questões, prever resultados, etc. Nos últimos anos essa técnica já vem sendo bem visível em aplicativos, um exemplo é o reconhecimento fácil em vários aparelhos espalhados no mundo. Logo, neste estudo trabalhamos com um dos conceitos da IA denominado *Machine Learning* (ML), traduzido como Aprendizado de Máquina. O ML trata-se de um subconjunto de técnicas da IA, visto que seus algoritmos são baseado na matemática e estatística (ESCOVEDO; KOSHIYAMA, 2020). Essa técnica tem ganhado bastante visibilidade em análises que objetiva a realização de predições na taxa de sobrevivência dos indivíduos. Para trabalhar com essa técnica, são utilizadas informações do conjunto de treino de maneira que seja possível prever algo determinado no conjunto de teste.

O sistema utilizado para o aprendizado de máquina trata-se de um sistema supervisionado, visto que os dados são fornecidos ao algoritmo. Além disso, o algoritmo possibilita a detecção de padrões nos dados de treinamento e, assim, é criado um modelo preditivo para resolução do problema.

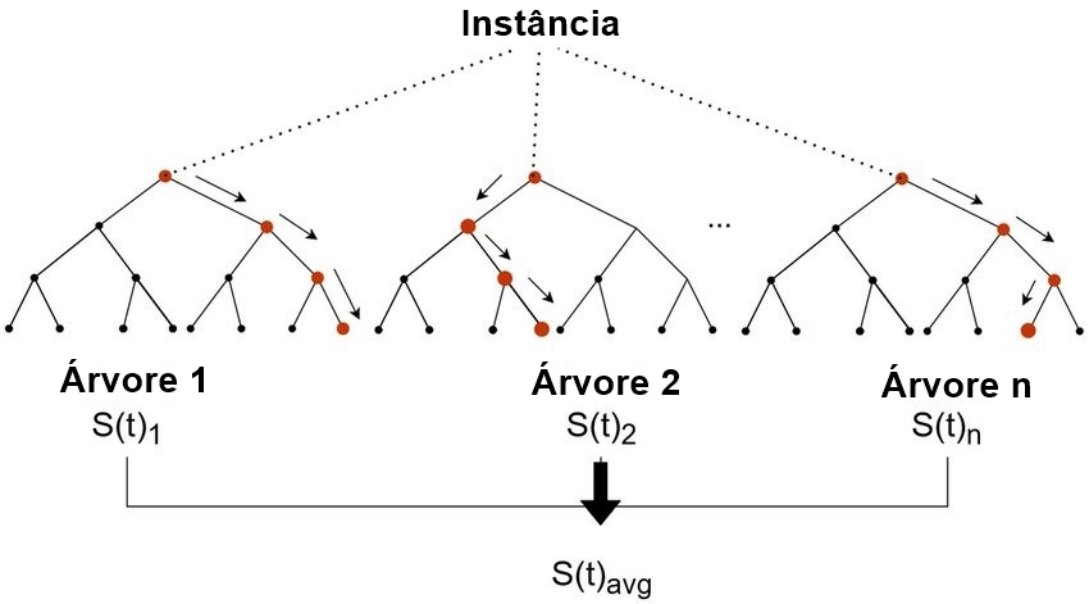
2.3.1 Random Survival Forest (RSF)

Quando lidamos com dados que possuem tempo de falha com censura à direita e há o intuito de realizar uma análise, pode-se aplicar um método conhecido como *Random Survival Forest* (RSF) (ISHWARAN et al., 2008). Este método é derivado da *Random Forest*, no qual o algoritmo RSF foi inicialmente implementado por Breiman (2001). O método é introduzido de duas formas, no primeiro passo são sorteadas amostras de *Bootstrap*, no qual cada amostra se tornará uma árvore de sobrevivência. No segundo passo é selecionado em cada nó uma das variáveis do conjunto das variáveis do estudo pretendentes para a previsão, e, estes nós são definidos pelo método de classificação do *log*, de maneira que é selecionado a variável que apresenta a maior profundidade nas subárvores, a divisão do nó é realizada com base no teste *Log-Rank*. Em seguida é utilizado a função de risco de *Nelson-Aalen* com ela se calcula a função de risco acumulada para uma árvore, para logo calcular-se a média da função de risco acumulada para todas as árvores resultando em ricas classes com mensuração no erro de previsão (MIAO et al., 2015). Estas previsões são denominadas índices de riscos de mortalidades. Utilizando métodos semelhantes, a RSF trás uma abordagem simples permitindo descobrir estruturas difíceis por meio do estudo superintendente do diagnóstico de um paciente.

Na realização da RSF, os parâmetros *status*, tempo e censura são essenciais para a análise de sobrevivência, no qual os nós das arvores são divididos embasados no tempo e nas covariáveis, resultando em estimativas não paramétricas para a função de sobrevida, em que será possível a escolha de quais variáveis possuem efeitos significativos na sobrevivência.

Em suma, o *Machine Learning*, por sua vez, surge como uma ferramenta alternativa para aperfeiçoar as árvores de decisão que formam a floresta aleatória por meio dos processos de aleatorização. Na Figura 2 é exemplificado de forma simples como funciona a RSF.

Figura 2 – Método da *Random Survival Forest*



Fonte: Snider e McBean (2022).

Para selecionar as variáveis de impacto positivo para o modelo final, se utiliza a medida *variable importance* (VIMP), que retorna um valor no qual quanto maior for mais indicação de que aquela variável apresenta uma boa capacidade preditiva, enquanto que valores baixos indica que a variável deve ser filtrada. De modo geral, a estatística VIMP, segundo Ishwaran e Kogalur (2007), mede o aumento ou a queda no erro de classificação nos dados teste.

Para a validação do método algumas estatísticas de desempenho são apresentadas, a seguir, no qual umas das medidas importantes para verificar o poder do modelo será o *C-index*, calculado com base nas amostras de *bootstrap* separadas anteriormente.

2.3.2 Estatísticas de Desempenho

Com o intuito de buscar a precisão das predições realizadas para validar um dado modelo em estudo, surgem algumas estatísticas de desempenho, entre elas as que serão utilizadas nesse estudo, a *Brier Score* e a *C-index*, que são calculadas com base na função de risco acumulado de *Nelson-Aalen* para todas as árvores. De maneira geral, a *Brier Score* (BS) aplica-se no conjunto de dados de teste objetivando a avaliação da qualidade das predições, por sua vez, 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 refere-se a dados separados ao se calcular as amostras com o intuito de utilizá-los para medir o erro de previsão.

A estatística C-*Index*, segundo Leger et al. (2017), é uma generalização da área da curva de sobrevivência, possibilitando validar o modelo preditivo, em que o valor de C-*Index* quando próximo a $0,5$ indica que o modelo não é preciso o suficiente para se utilizar, entretanto, quando o valor é próximo de 1 significa que as previsões são precisas. Nesse sentido, de acordo com Oliveira et al. (2019), para se obter o valor de C-*Index* seja $(T_{1,h};\delta_{1,h}),(T_{2,h};\delta_{2,h}),\ldots,(T_{n,h};\delta_{n,h})$ os tempos de sobrevivência com os status de censura para os n indivíduos da amostra. Logo $t_{1,h},t_{2,h},\ldots,t_{m,h}$, possui o m-ésimo tempo de um determinado evento distinto em um nó na árvore de decisão h. Posteriormente, o número de mortes $d_{(1,h)}$ deve ser definido e o número de indivíduos em risco $Y_{(1,h)}$ no tempo $t_{(1,h)}$. De maneira tal que é possível definir a Função do

Risco Acumulado para estimar o nó terminal h, baseado no Estimador de *Nelson-Aalen* expresso a seguir.

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

No entanto, quando temos covariáveis x_i d-dimensionais para cada indivíduo i tem-se então.

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

Na RSF para estimar a Função de Risco Acumulada para uma observação i, considera-se $I_{(i,b)} = 1$ caso i pertença ao *Out of Bag* (OOB), b-ésima amostra por *bootstrap*, caso contrário, $I_{(i,b)} = 0$. Contudo, a função de risco acumulada mais utilizada é expressa, a seguir, para cada observação i.

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

Mas vale salientar que na RSF o valor C-*index* é calculado com o somatório de todas observações $\sum_{l=1}^{m} H(t_l|x_i)$, em que $t_1 < t_2 < \ldots < t_m$ são os tempos dos eventos únicos no banco de dados.

Quanto à estatística de performance *Brier Score*, esta é calculada em um determinado tempo t, em que a pontuação de *Brier* é definida como a diferença quadrada entre o status e sobrevivência e uma previsão também realizada no mesmo tempo t, entretanto o valor utilizado para tomada de decisão é a pontuação de *Brier* esperada em toda simulação na RSF calculado pela seguinte expressão

$$BS(t, \hat{S}) = E\left[S_i(t) - \hat{S}(t|X_i)^2\right]$$

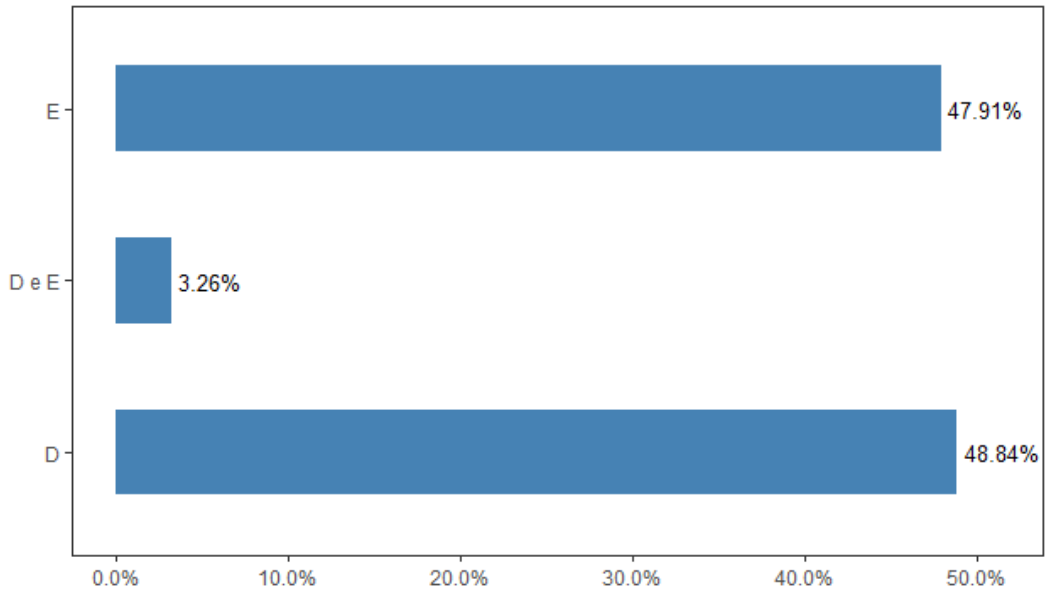
Em que Y_i é o verdadeiro status para o individuo i, $\hat{S}(t|X_i)^2$ é a probabilidade de sobrevivência prevista no tempo t para o individuo i com X_i variáveis preditoras. Para cálculo da probabilidade do erro do modelo, utiliza-se o IBS, identificado como CRPS na saída do modelo, expresso pela seguinte função.

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

3 RESULTADOS E DISCUSSÃO

Inicialmente os resultados apresentados referem-se a análise exploratória dos dados de câncer de mama em mulheres que fizeram tratamento no Hospital da FAP. Serão expostas algumas informações preliminares em relação as variáveis sob estudo, por meio dos gráficos e tabelas, abaixo.

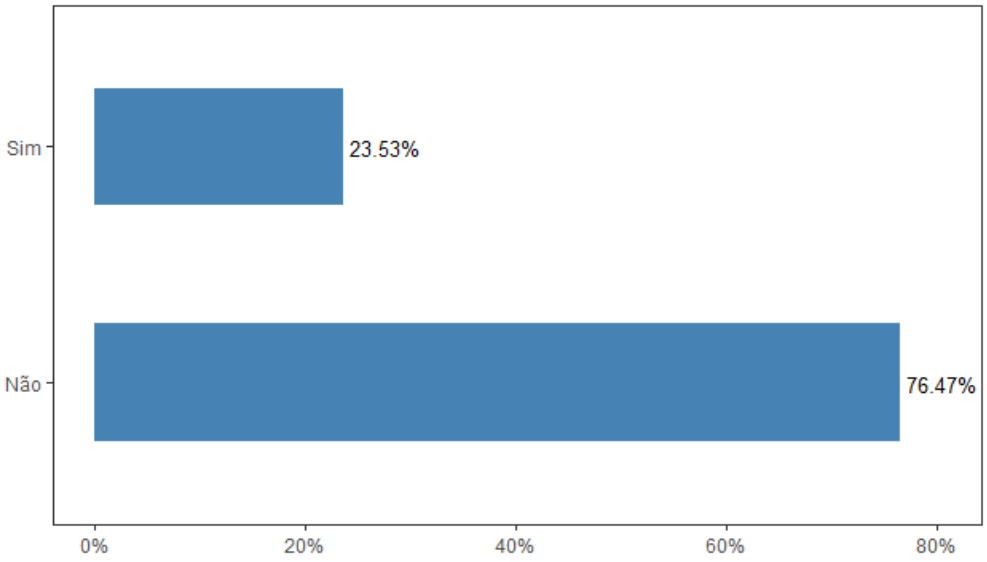
Figura 3 – Gráfico referente a mama que possui o câncer.



Fonte: Produzido pelos autores.

De acordo com a Figura 3, podemos observar que o percentual de pacientes que possuem o câncer na mama direita ou esquerda é muito semelhante, em que 48,84% destas pacientes têm o câncer na mama direita, enquanto que 47,91% das mulheres possuem tal doença na mama esquerda e apenas 3,26% destas possuem nas duas mamas simultaneamente. Visto que nesta amostra não há uma maior prevalência de câncer por lado da mama. Na Figura 4 é apresentado o percentual de dados censurado e falhas. Sendo a falha definida como ter ocorrido o evento de interesse de estudo, neste caso, a morte devido ao câncer de mama.

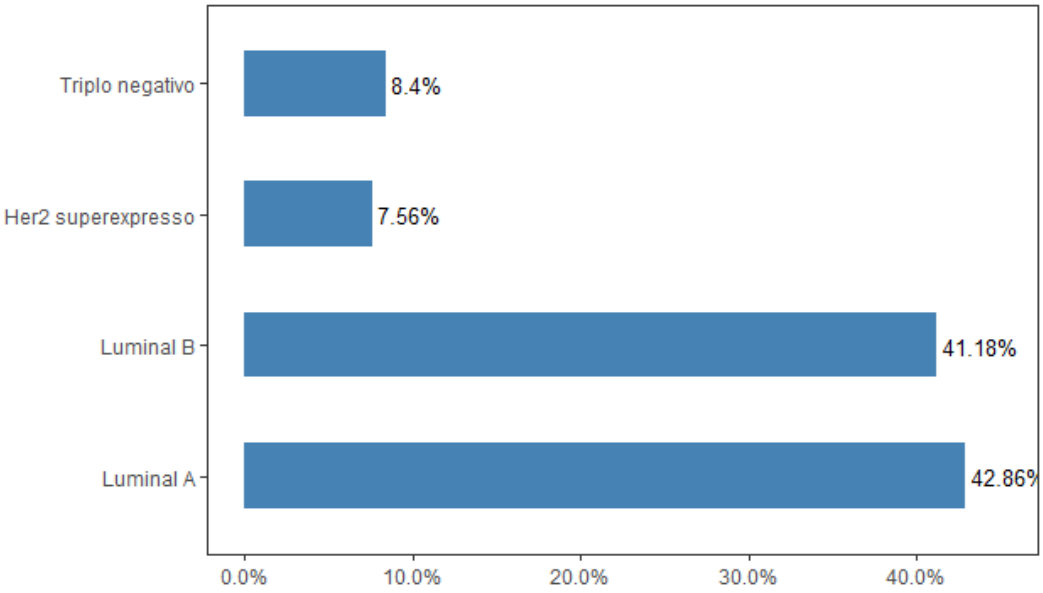
Figura 4 – Gráfico referente a morte das pacientes.



Fonte: Produzido pelos autores.

Em respeito ao percentual de mortes das mulheres acometidas pelo câncer de mama, observa-se que apenas cerca de 23% destas chegaram a óbito, estas são as pacientes que tiveram o evento de interesse do estudo, as demais pacientes apresentaram um quadro clínico de alta ou desistiram do tratamento para o câncer. Foi levantado os subtipos moleculares apresentados na Figura 5.

Figura 5 – Gráfico referente ao Subtipo Molecular.

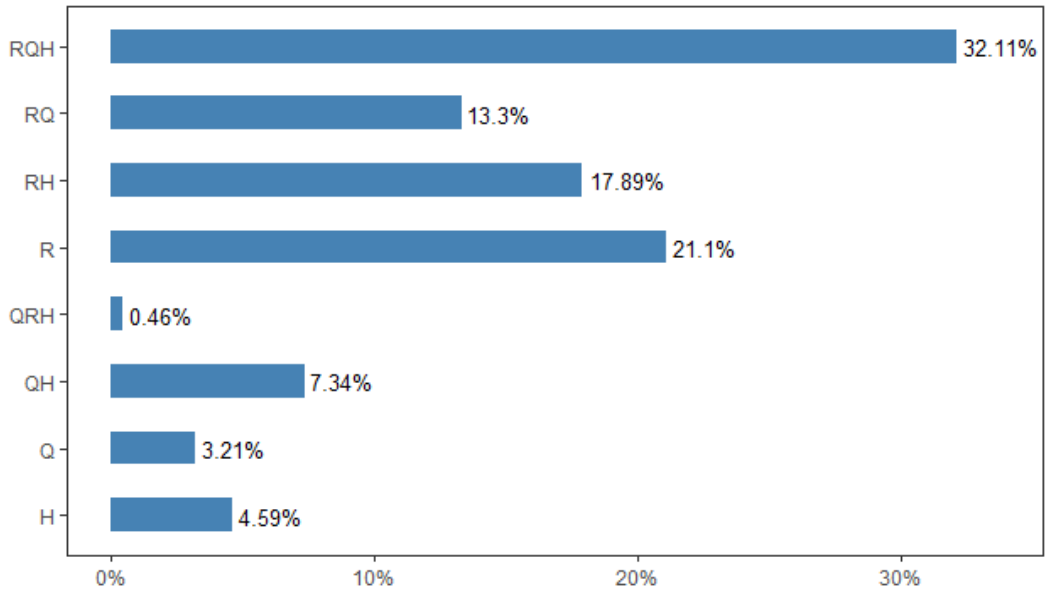


Fonte: Produzido pelos autores.

Referente aos quatro tipos de câncer de mama, que são classificados de acordo com as características das células cancerosas, pode-se notar na Figura 5, que os cânceres do tipo Luminal A e B possuem os maiores percentuais. Nesse sentido, 42,86% das pacientes possuíam tumores mamários do tipo Luminal A, enquanto que cerca de 41% tinham tumores do tipo Luminal B, corroborando com o que afirmam Cirqueira et al. (2011) em relação a estes tipos serem os mais comuns, enquanto a porcentagem de pacientes que apresentaram um Subtipo Molecular Triplo negativo e Her2 superexpresso foi de 8,4% e 7,56% respectivamente, podendo-se concluir que uma pequena porcentagem das pacientes apresentam um subtipo molecular de câncer agressivo.

Sabe-se, segundo o que afirmam Nicolussi e Sawada (2011), que quando mulheres são acometidas pelo câncer de mama, no período de tratamento da doença realiza-se procedimentos em busca da cura objetivando eliminar toda e quaisquer célula cancerígena presente no corpo, para que se impeça a remissão de tal câncer. Desse modo, para impedir esse retorno, bem como auxiliar na cura, existem tratamentos complementares chamados de terapia adjuvante.

Figura 6 – Gráfico referente a Terapia Adjuvante.



Fonte: Produzido pelos autores.

Sobre o percentual de pacientes que realizaram algum desses tratamentos, nota-se na Figura 6 que 32,11% das mulheres se submeteram aos tratamentos de Radioterapia (R), Quimioterapia (Q) e Hormonoterapia (H) simultaneamente, bem como 21,10% destas utilizaram a Radioterapia como tratamento. Além disso, com relação as terapias adjuvantes, submetidas de forma simultânea nas pacientes, Radioterapia e Hormonoterapia, Radioterapia e Quimioterapia, observa-se percentuais de 17,89% e 13,3%, respectivamente. Esses percentuais também podem ser observados na Tabela 1.

Tabela 1 – Frequências Absolutas e Relativas.

Variáveis	Categorias	n	%	% Válido
Morte	Não	169	76,50	76,50
	Sim	52	23,50	23,50
Local	D	105	47,50	48,80
	D e E	7	3,20	3,30
	E	103	46,60	47,90
	NA	6	2,70	NA
Terapia adjuvante	H	10	4,50	4,60
	Q	7	3,20	3,20
	QH	16	7,20	7,30
	QRH	1	0,50	0,50
	R	46	20,80	21,10
	RH	39	17,60	17,90
	RQ	29	13,10	13,30
	RQH	70	31,70	32,10
	NA	3	1,40	NA
Receptor de estrogênio	Negativo	22	10,00	17,20
	Positivo	106	48,00	82,80
	NA	93	42,10	NA
Receptor de progesterona	Negativo	41	18,60	32,50
	Positivo	85	38,50	67,50
	NA	95	43,00	NA
Ki - 67	< 15%	53	24,00	48,60
	15% - 50%	41	18,60	37,60
	> 50%	15	6,80	13,80
	NA	112	50,70	NA
Proteína P53	Positivo	42	19,00	36,20
	Negativo	74	33,50	63,80
	NA	105	47,50	NA
C - erb - b2	Negativo	48	21,70	40,70
	Positivo	70	31,70	59,30
	NA	103	46,60	NA
Subtipo Molecular	Luminal A	51	23,10	42,90
	Luminal B	49	22,20	41,20
	Her2 superexpresso	9	4,10	7,60
	Triplo negativo	10	4,50	8,40
	NA	102	46,20	NA
Total		221	100,00	

Fonte: Produzido pelos autores.

Na Tabela 1 observam-se informações, além das expressas anteriormente por meio dos gráficos. Dessa forma, para os pacientes que realizaram a biópsia das células cancerígenas, a porcentagem de paciente com o receptor de estrogênio positivo foi de 82,80%, enquanto que 67,5% apresentaram diagnóstico positivo para o receptor de progesterona. De acordo com Pachnicki et al. (2012), as pacientes que têm pelo menos um dos receptores positivos podem realizar o tratamento com hormonoterapia, diminuindo os índices de estrogênio nas células

cancerígenas, pois os mesmos resultam em um desenvolvimento do câncer.

As proteínas Ki-67, p53 e C-erb-b2 podem ser obtidas através do exame imuno histoquímico. Neste exame, como afirmam Salles et al. (2009), é possível observar se alguma mutação ocorreu na célula com base nos indicadores. A proteína Ki-67 mede o nível de agressividade do câncer, em que a medida que a porcentagem da proteína cresce indica um câncer com efeito agressivo na paciente, observando que apenas 13,8% das pacientes apresentaram um estado extremamente agressivo do câncer. Além disso, a proteína p53 tem a finalidade de prevenir a proliferação de genes duplicados, no qual observa-se que apenas 36% das pacientes apresentaram positividade para essa proteína. Sobre a proteína C-erb-b2, esta influencia no crescimento do tumor, observando na Tabela 1 que aproximadamente 60% das pacientes apresentam resultado positivo neste exame e, portanto, necessitam de cuidados médicos para o tratamento do câncer. Para as demais variáveis sob estudo, tem-se na Tabela 2 algumas medidas descritivas.

Tabela 2 – Medidas descritivas.

Variáveis	Min	Max	1º Q.	Mediana	3º Q.	Média	D. padrão
Nº de Hormonoterapia	0,00	109,00	0,00	12,00	58,00	26,21	28,60
Nº de Quimioterapia	0,00	67,00	0,00	4,00	14,00	8,91	12,67
Nº de Radioterapia	0,00	90,00	25,00	25,00	30,25	25,65	14,58
Tempo Morte	21,00	3121,00	301,00	876,00	1446,00	995,07	830,98
Tempo Consulta	7,00	4809,00	208,00	921,50	1842,00	1124,52	1026,65
Idade	30,00	89,00	49,00	59,00	68,00	58,71	12,82

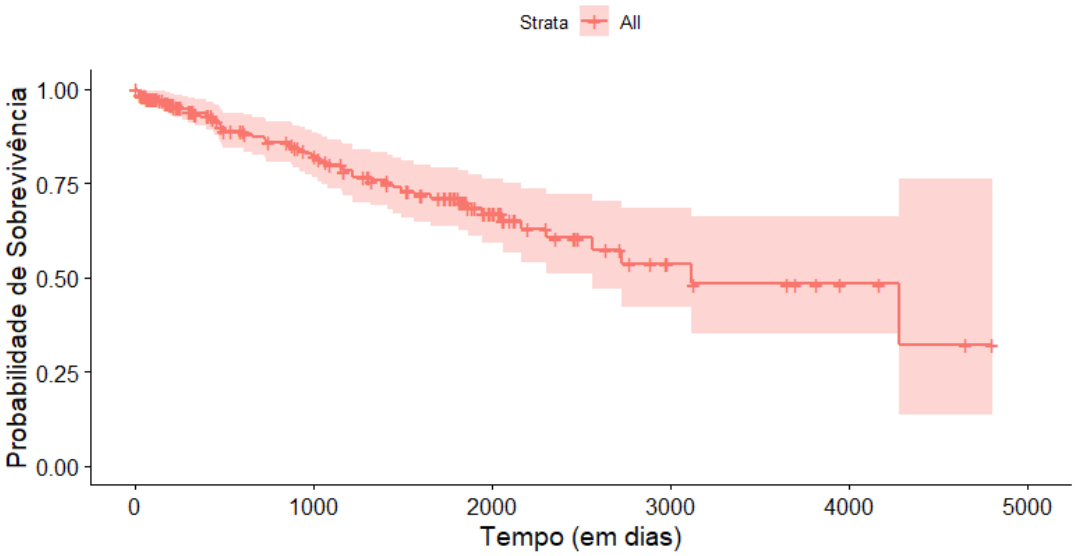
Fonte: Produzido pelos autores.

Na Tabela 2 é possível observar que as pacientes se submeteram, em média, a 26 sessões de Hormonoterapia, cerca de 8 sessões de Quimioterapia, bem como em torno de 25 Radioterapias, no qual tais pacientes possuíam em média 58 anos. Em relação aos tempos, vale salientar que as mulheres acometidas pelo câncer mamário tiveram cerca de 995 dias que referem-se a 2,73 anos, em média, até o óbito. O tempo médio que tais pacientes ficaram sob acompanhamento foi de 1124,52 dias, que representam cerca de três anos.

3.1 Gráficos das Curvas de Sobrevivência

Nesta seção é apresentada a análise com relação as curvas de sobrevivência. Na Figura 7 é apresentado o decaimento da curva à medida que o evento de interesse ocorre ao logo do tempo, isto é, o óbito das pacientes.

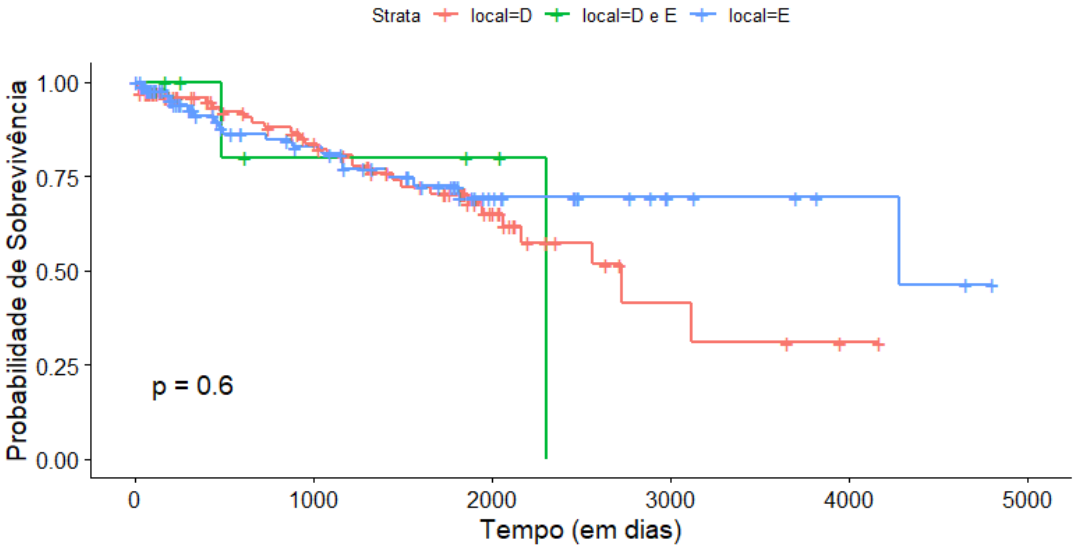
Figura 7 – Gráfico referente a curva de sobrevivência.



Fonte: Produzido pelos autores.

De acordo com a Figura 7, percebe-se que em 3000 dias 50% dos pacientes tiveram o evento de interesse ocorrido ou censurado. Nas próximas figuras são apresentados os respectivos *p-valores* referente ao teste de *Log-rank* que permite comparar as curvas de *Kaplan-Meier*. Na Figura 8, por sua vez, são apresentadas as curvas de sobrevivência para a verificar se o local do câncer influencia no desfecho da análise.

Figura 8 – Gráfico referente a curva de sobrevivência para local.

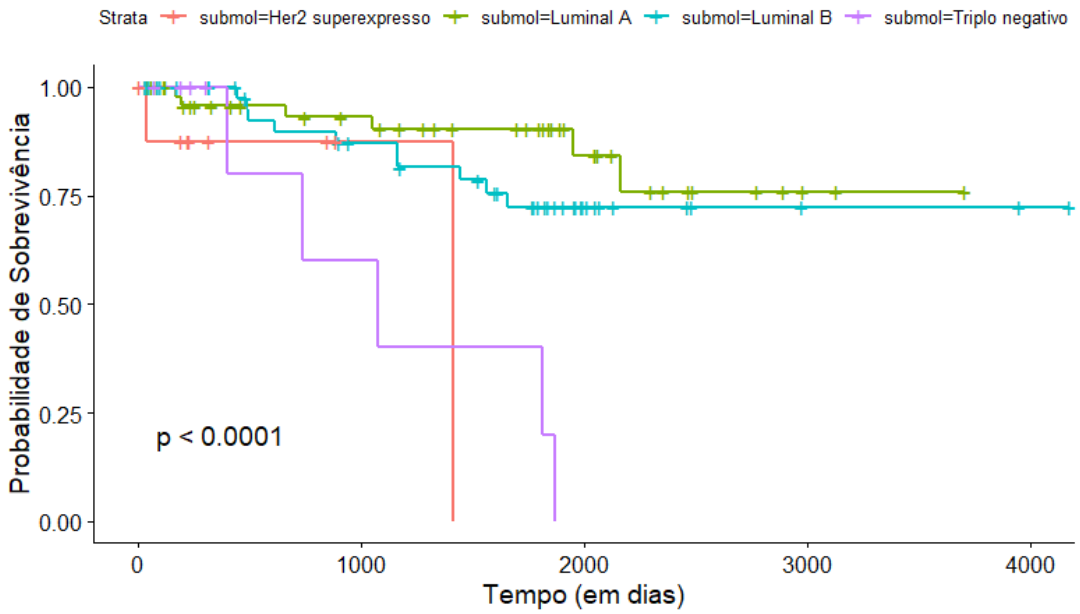


Fonte: Produzido pelos autores.

Na Figura 8 observa-se um *p-valor* não significativo evidenciado pelo teste de *Log-rank*, isto indica que não há diferença significativa entre os grupos, a um nível de confiança de 5%, ou seja, entre as curvas de sobrevivência da localidade do câncer de mama, confirmando o que

foi apresentado anteriormente na análise descritiva. Com respeito ao subtipo molecular que as pacientes possuíam, pode-se observar na Figura 9.

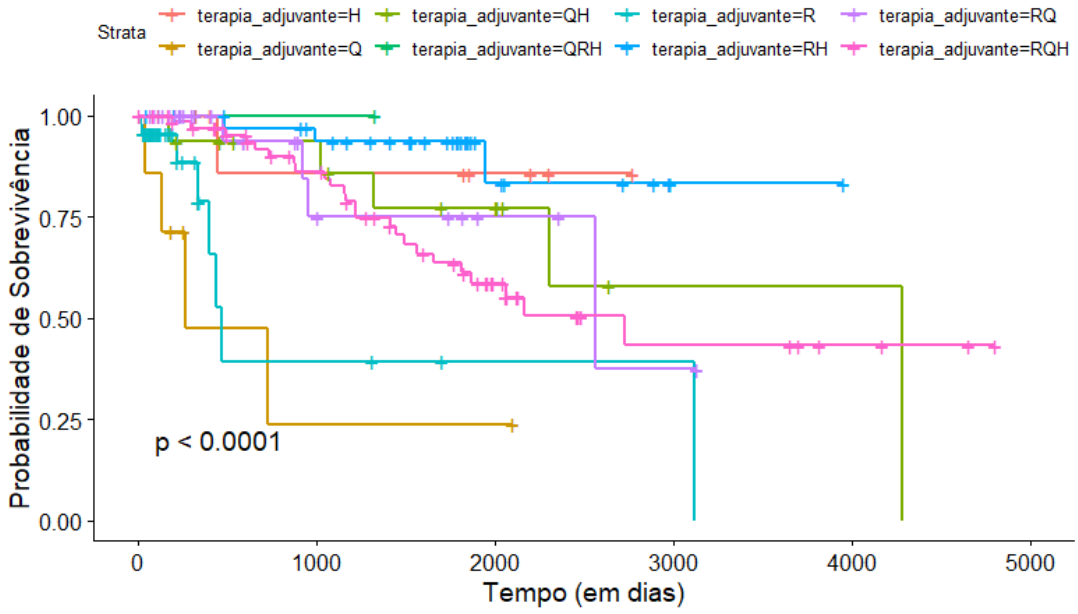
Figura 9 – Gráfico referente a curva de sobrevivência para Subtipo Molecular.



Fonte: Produzido pelos autores.

Na Figura 9 é possível observar um *p-valor* significativo evidenciado pelo teste de *Log-Rank*, indicando que há diferença em pelo menos um dos grupos para o subtipo molecular das células cancerígenas, observando que na análise visual do gráfico, os pacientes que apresentaram um subtipo molecular Triplo Negativo e um Her2 superexpresso possuem uma probabilidade de sobrevivência inferior após mil dias, se tratando também de um subtipo não comum, quando comparado com os demais. Por sua vez, na Figura 10 verifica-se a influência das terapias adjuvantes na morte das pacientes.

Figura 10 – Gráfico referente a curva de sobrevivência para a terapia adjuvante.



Fonte: Produzido pelos autores.

É apresentado na Figura 10 as curvas de sobrevivências de acordo com a terapia adjuvante, assim, verifica-se que há diferenças significativas pelo teste de *Log-Rank*, isto retrata que pelo menos um dos grupos apresenta diferença entre si, no qual observa-se que o decaimento da curva de sobrevida para os pacientes que realizaram a terapia adjuvante só com a Radioterapia representa uma expectativa de vida inferior aos demais, isto vale também para as pacientes que realizaram o tratamento só com a Quimioterapia.

3.2 RSF

O intuito desta seção é apresentar o resultados obtidos com a aplicação da técnica de *Random Survival Forest*, em que utilizou-se um dos conceitos básicos de *Machine Learning* para dividir a amostra em treino e teste. O conjunto de dados treino foi originado de 70% do banco de dados original, isso equivale a 155 observações sobre as pacientes, observando na Figura 11 que 36 dos mesmos tiveram o evento de interesse.

Figura 11 – Saída referente ao Modelo Completo.

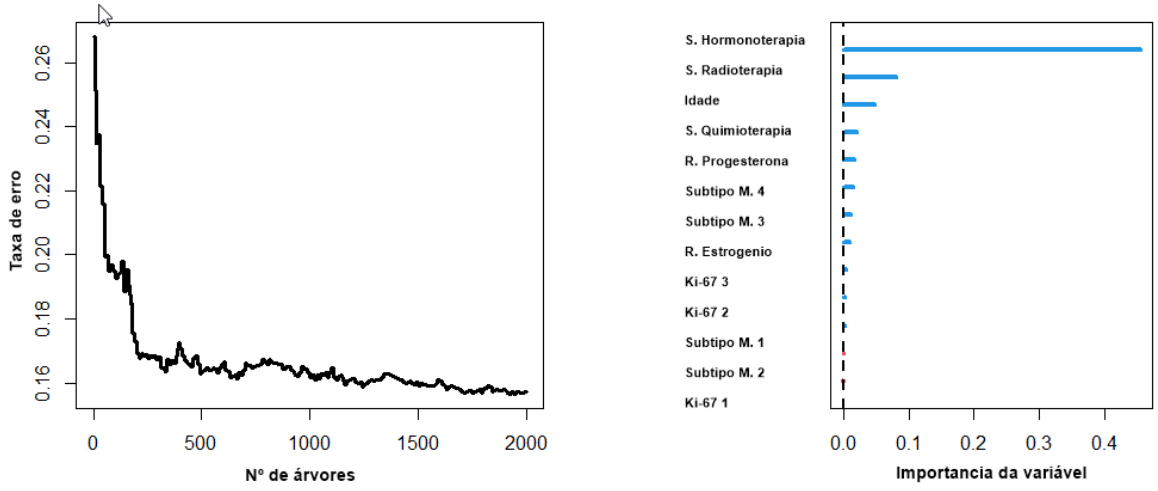
```
                         Sample size: 155
                    Number of deaths: 36
                    Was data imputed: yes
                     Number of trees: 2000
          Forest terminal node size: 15
      Average no. of terminal nodes: 6.276
No. of variables tried at each split: 3
            Total no. of variables: 13
      Resampling used to grow trees: swor
   Resample size used to grow trees: 98
                           Analysis: RSF
                             Family: surv
                   Splitting rule: logrank *random*
      Number of random split points: 10
                         (OOB) CRPS: 0.15124188
(OOB) Requested performance error: 0.15729167
```

Fonte: Produzido pelos autores.

Na Figura 11 observa-se que foi realizado 2000 árvores com amostras de *bootstrap*. Para cada amostra foi retirado uma média de 37% das informações que são armazenadas de fora e utilizadas para compor o *Out of Bag* OOB, estes valores servem para medir o erro de predição em cada árvore. Na regra de divisão dos nós foi utilizado o teste de *Log-Rank*. O CRPS é a medida de desempenho *Integrate Brier Score*, que resulta em uma probabilidade de 0,1512 (15,12%) para erro de previsão, indicando um bom desempenho para o modelo preditivo. Outra medida observada é o erro de performance do modelo, que com base na métrica *C- index* o erro associado foi de 0,1573 (15,73%), complementando que o valor de *C- index* foi de 0,8427 ou 84,27% que é uma taxa que representa uma boa preditividade do modelo completo.

Para selecionar as variáveis para o modelo final apresentam-se nas Figuras 12 e 13 os resultados de acordo a estatística VIMP.

Figura 12 – Gráficos de seleção de variáveis pela estatística VIMP.



Fonte: Produzido pelos autores.

É possível observar na Figura 12 que no gráfico esquerdo a convergência da árvore foi em torno 1000 mil árvores. Além do mais, no gráfico à direita observa-se um gráfico de colunas, em que cada barra representa de acordo com a estatística VIMP o valor de impacto das variáveis no modelo, observando que o número de hormonoterapia, de radioterapia e a idade das pacientes apresentaram mais impacto do que as demais variáveis. Os valores apresentados nestes gráficos são originados pela lista da Figura 13.

Figura 13 – Saída do modelo RSF para seleção de variáveis.

```
---------------------------------
family              : surv
var. selection      : Minimal Depth
conservativeness    : medium
x-weighting used?   : TRUE
dimension           : 13
sample size         : 155
ntree               : 2000
nsplit              : 10
mtry                : 3
nodesize            : 15
refitted forest     : FALSE
model size          : 5
depth threshold     : 3.3197
PE (true OOB)       : 15.7292


Top variables:
                     depth  vimp
n_de_hormonio        2.183 0.456
idade                2.481 0.047
n_de_radio           2.608 0.080
n_de_quimio          3.118 0.020
recptor_progesterona 3.215 0.016
```

Fonte: Produzido pelos autores.

Na Figura 13 no *depth*, traduzido de profundidade, o Número de hormonoterapia e da idade foi inferior aos demais, indicando que estas duas variáveis têm uma ramificação maior, quando comparada as outras, após a divisão do nó. Os valores da *Variable importance* são apresentados de maneira decrescente, no qual observamos que o Número de hormonoterapia, radioterapia, quimioterapia, a idade do paciente e o receptor de progesterona foram variáveis com maior impacto preditivo para evento de interesse, ou seja, o óbito do paciente. Com estas informações, é possível definir as variáveis que irão compor o modelo RSF apresentado na Figura 14.

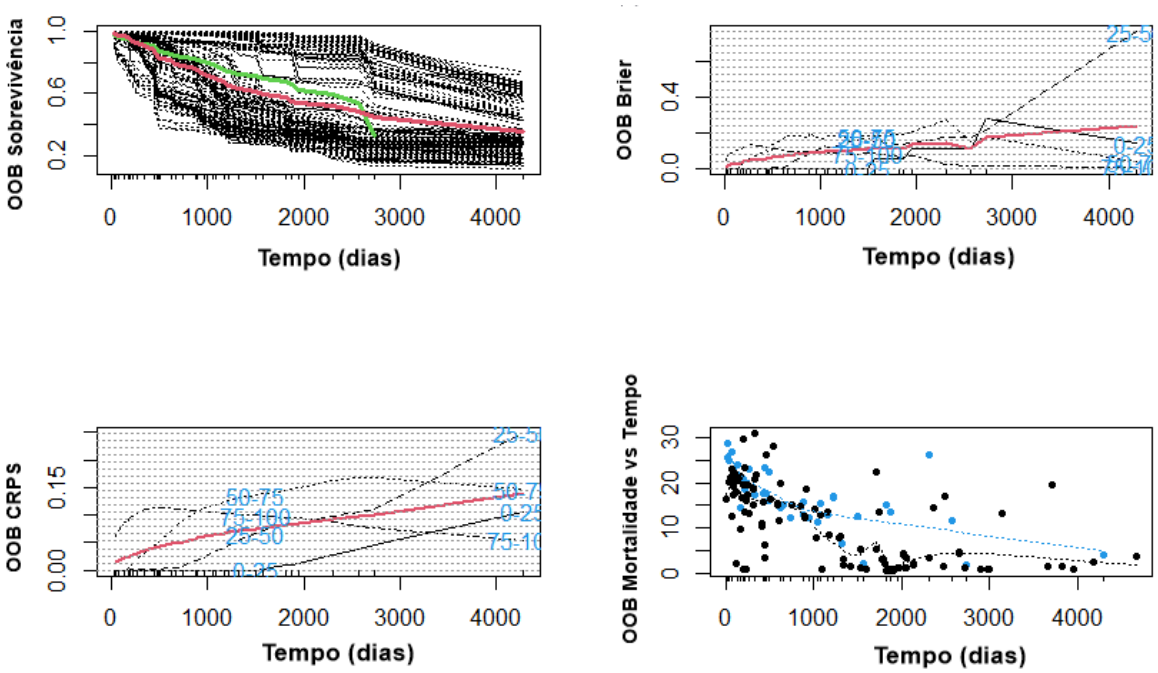
Figura 14 – Saída referente ao Modelo com variáveis selecionadas.

```
                         Sample size: 155
                  Number of deaths: 36
                  Was data imputed: yes
                   Number of trees: 2000
          Forest terminal node size: 15
      Average no. of terminal nodes: 6.588
No. of variables tried at each split: 3
              Total no. of variables: 5
      Resampling used to grow trees: swor
   Resample size used to grow trees: 98
                           Analysis: RSF
                             Family: surv
                   Splitting rule: logrank *random*
      Number of random split points: 10
                         (OOB) CRPS: 0.13878145
   (OOB) Requested performance error: 0.12291667
```

Fonte: Produzido pelos autores.

Na Figura 14 é apresentada a saída referente ao modelo selecionado com um total de cinco variáveis preditoras. Observa que os resultados apresentados foram mais satisfatórios, no qual obteve-se o valor do *Integrate Brier Score* de 0,1387 (13,88%) para o desempenho preditivo do modelo, sendo assim menor que o do modelo completo. Além disso, o valor do erro de performance do modelo foi de 0,1229 (12,29%), de maneira que o modelo obteve uma boa capacidade preditiva no banco de dados treino. Por fim, 87,71% de assertividade representa o valor *C-index*. O comportamento das métricas para a avaliação do modelo é apresentado na Figura 15.

Figura 15 – Gráfico referente ao Modelo selecionadas-OOB.



Fonte: Produzido pelos autores.

Na Figura 15, no gráfico superior a esquerda, é possível observar a curva de sobrevivência global na cor vermelha para todo conjunto, enquanto que a linha verde representa o estimador de *Nelson-Aalen*, com isso nota-se que as duas curvas estão bem próximas, indicando que a função de sobrevida estimada pela RSF se conforma com a curva de sobrevivência real. Além disso, é possível afirmar que em 2000 mil dias de diagnóstico do câncer de mama, aproximadamente 40% dos pacientes já tiveram o evento de interesse, com base na curva sobrevivência global. O gráfico superior direito é relacionado ao comportamento do erro de performance ao longo do tempo e pode-se observar que a curva ficou em torno de 0,15, o mesmo vale para o CRPS que apresenta um bom comportamento na curva em torno de 0,15. No gráfico inferior da direita, nota-se uma concentração de eventos de interesse (cor azul) que ocorreram entre 0 e 1000 dias, enquanto que os pontos na cor preta representam as censuras. Tendo escolhido as variáveis, na Figura 16 são apresentados os resultados da predição com o modelo final.

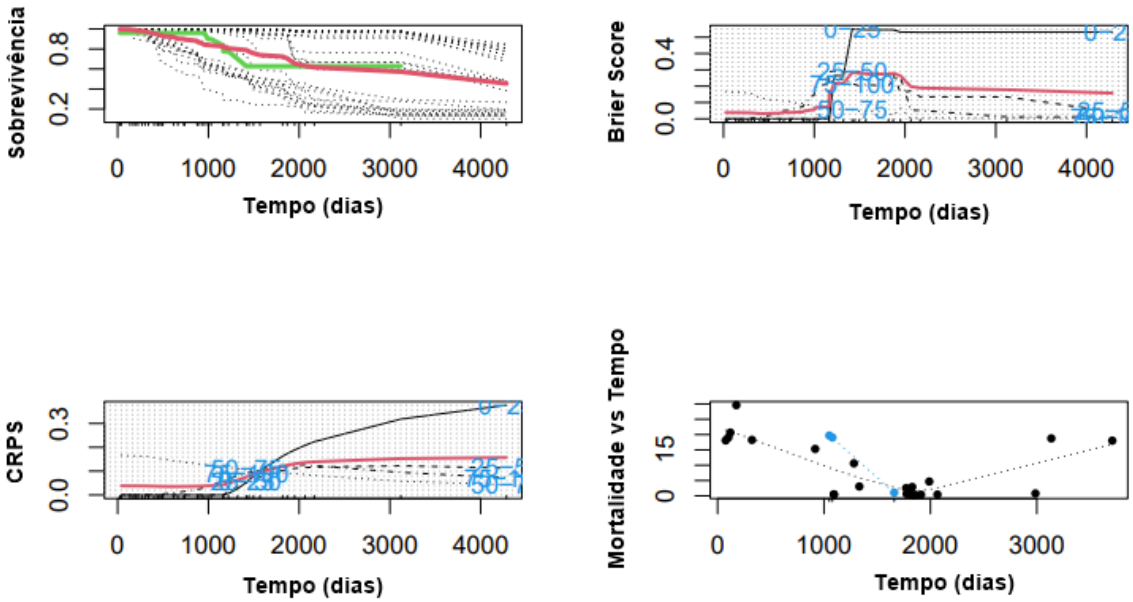
Figura 16 – Predição utilizando o modelo final.

```
Sample size of test (predict) data: 26
                 Number of grow trees: 2000
Average no. of grow terminal nodes: 8.3925
       Total no. of grow variables: 5
    Resampling used to grow trees: swor
  Resample size used to grow trees: 16
                           Analysis: RSF
                             Family: surv
                               CRPS: 0.155948
        Requested performance error: 0.12
```

Fonte: Produzido pelos autores.

Na Figura 16 apresenta-se o resultado da predição realizada com o banco teste, com 26 observações de pacientes para associar o acerto do modelo, em que pode-se observar um CRPS de 0,1559 (15,59%) de erro de previsão do modelo, no qual também foi obtido um erro de performance de 0,12 (12%) ao longo do tempo. As métricas de avaliação são apresentadas na Figura 17.

Figura 17 – Gráfico referente as métricas para o Modelo preditivo

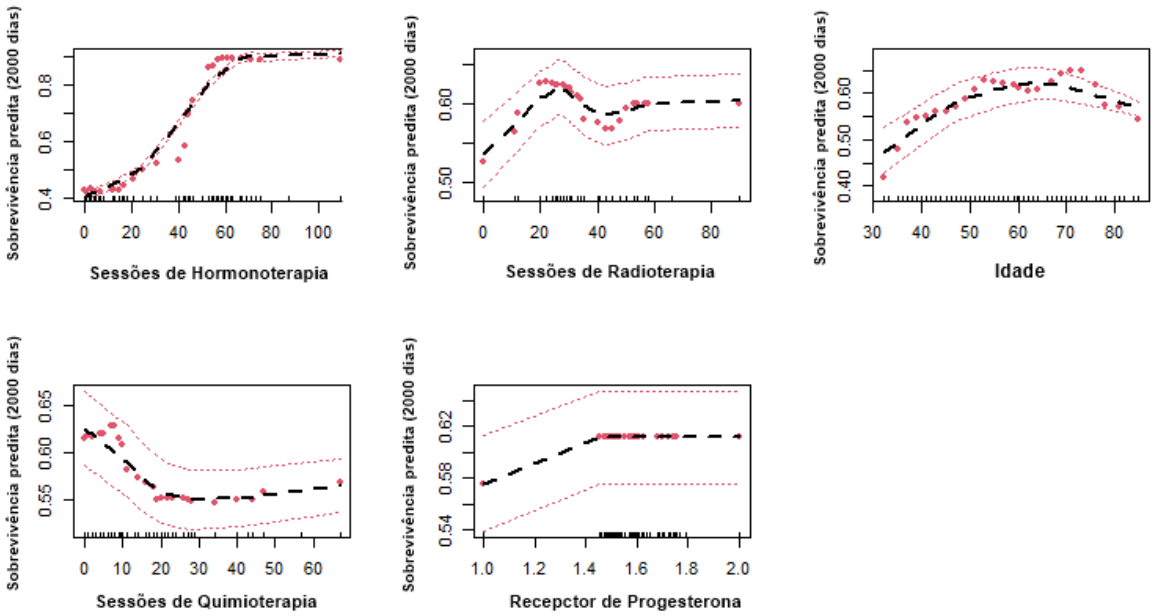


Fonte: Produzido pelos autores.

Na Figura 17 apresentam-se os resultados relacionados a predição utilizando o modelo selecionado, em que se pode observar no gráfico superior a esquerda, que a curva média de sobrevivência global indica que em mil dias de diagnóstico a paciente tem um risco de chegar à óbito em torno de 20%. O gráfico superior à direita é referente ao erro de predição cometido pelo modelo, ficando abaixo de 0,15 (15%). Para o gráfico inferior esquerdo, tem-se a curva do erro

(IBS), no qual esta curva ficou em torno de 0,1 (10%). No gráfico inferior à direita, observa-se que os pacientes tiveram o evento de interesse em torno de 1000 dias. Na Figura 18 apresenta-se o risco associado a cada variável do paciente vir a óbito, calculado com base nas demais.

Figura 18 – Predição utilizando o modelo final para cada variável.



Fonte: Produzido pelos autores.

Com base nas variáveis selecionadas para o modelo final, gerou-se os gráficos parciais baseados no modelo RSF que incluiu as cinco covariáveis. O eixo vertical representa a sobrevivência em 2 mil dias para uma determinada variável, com base no ajuste das demais. O eixo horizontal representa a quantidade de concentração de cada variável. Logo, observa-se que no gráfico para o Número de Hormonoterapia, a reta traçada de acordo os pontos preditos indica que o risco da paciente chegar óbito diminui a medida que o número de tratamentos cresce, em que partir de 60 sessões a sobrevivência predita de aproximadamente 0,95. O mesmo se aplica para a variável Número de radioterapia, pois os efeitos colaterais para pacientes que fazem uso deste tratamento são mais suaves, do que os demais tratamentos, visto que o a sobrevivência predita aumenta 0,10 nas primeiras 20 sessões e permanece em uma pseudo constância. Em contrapartida, a medida que os pacientes realizam sessões de quimioterapia o risco destes sobreviverem diminui. As pacientes que foram diagnosticadas com câncer em torno de 30 anos, tem um risco menor de chegar a óbito de aproximadamente 0,40 e este risco tende a subir para as pacientes com idade até 60 anos.

4 CONCLUSÃO

A Análise de Sobrevivência torna-se importante e crucial para prever diagnósticos clínicos com base no tempo. Neste trabalho o evento de interesse foi o óbito das pacientes, em que a previsão foi realizada com base em indicadores coletados nos prontuários do Hospital da FAP. Observou-se que em comparação com os grupos das pacientes diagnosticadas com o subtipo molecular Triplo Negativo e Her2 superexpresso as mesmas apresentaram uma probabilidade menor de sobrevivência no tratamento após mil dias, isto acontece pelo fato destes tumores serem

extremamente agressivos. Entretanto, o grupo de pacientes que realizaram só o tratamento com a quimioterapia e a radioterapia apresentaram um risco maior de vir a óbito quando comparado com pacientes que fizeram uso de tratamento simultaneamente ou só com a hormonoterapia.

Quando o intuito é realizar predições e se tem uma grande quantidade de variáveis preditoras, o RSF é um método interessante, uma vez que apresenta resultados simples com bases nas árvores de decisões. Além do mais, o pacote *randomForestSRC* do *Software R* traz funções simplórias que facilitam todo desenvolver da aplicação do método. Com isso, obteve-se um modelo com as variáveis Número de sessões de Hormonoterapia, Quimioterapia, radioterapia e a idade do paciente, sendo as que apresentaram mais impacto preditórios para prever a probabilidade de óbito do paciente, com base no tempo. Assim, o modelo apresentou um erro *CRPS* de 15,59%, e ao realizar a predição com o banco de dados teste, o mesmo teve um erro de performance de 12%. Com isso, concluindo que uma paciente com 1000 dias de diagnóstico tem uma probabilidade em torno de 20% de ir a óbito.

Para cada variável em particular concluiu-se que, a quantidade de sessões de hormonoterapia e radioterapia contribuem para diminuir o risco da paciente ir a o óbito, isto acontece por esses tratamentos ajudarem a melhorar a qualidade de vida da paciente, além de diminuir o tumor, aliviando uma série de efeitos malignos causados pelo mesmo. Em relação ao número de quimioterapia observou-se que a medida que o número de sessões cresce o risco da paciente vir a óbito aumenta, isto pode acontecer pelo fato dos efeitos serem bastantes agressivos e contínuos ao longo do tratamento. No mais, a quimioterapia proporciona a redução dos tumores e combate as células doentes. A variável idade indica que as pacientes com idade em torno de 30 anos têm uma probabilidade maior de sobreviver após ser diagnosticada e se submeter ao tratamento indicado pelo médico.

REFERÊNCIAS

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 15.

CIRQUEIRA, M. B. et al. Subtipos moleculares do câncer de mama. *Femina*, 2011. Citado na página 19.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006. Citado 3 vezes nas páginas 11, 12 e 14.

ESCOVEDO, T.; KOSHIYAMA, A. *Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise*. [S.l.]: Casa do Código, 2020. Citado na página 15.

FAY, M. P.; SHAW, P. A. Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. *Journal of statistical software*, NIH Public Access, v. 36, n. 2, 2010. Citado na página 14.

INCA. *A situação do câncer de mama no Brasil: síntese de dados dos sistemas de informaçãol*. Rio de Janeiro, 2019. Disponível em: <https://www.inca.gov.br/publicacoes/livros/situacao-do-cancer-de-mama-no-brasil-sintese-de-dados-dos-sistemas-de-informacao>. Citado na página 11.

ISHWARAN, H.; KOGALUR, U. B. Random survival forests for r. *R news*, v. 7, n. 2, p. 25–31, 2007. Citado 2 vezes nas páginas 12 e 16.

ISHWARAN, H. et al. Random survival forests. *The annals of applied statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841–860, 2008. Citado 2 vezes nas páginas 11 e 15.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 14.

LEGER, S. et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific reports*, Nature Publishing Group, v. 7, n. 1, p. 1–11, 2017. Citado na página 16.

LIMA, P. A. Estimadores não paramétricos para a função de sobrevivência aplicada em um plano de saúde. DECAT-Departamento de Estatística e Ciências Atuariais–Ciências Atuariais ..., 2019. Citado na página 15.

MIAO, F. et al. Risk prediction of one-year mortality in patients with cardiac arrhythmias using random survival forest. *Computational and mathematical methods in medicine*, Hindawi, v. 2015, 2015. Citado na página 15.

NELSON, W. B. *Applied life data analysis*. [S.l.]: John Wiley & Sons, 1982. Citado na página 14.

NICOLUSSI, A. C.; SAWADA, N. O. Qualidade de vida de pacientes com câncer de mama em terapia adjuvante. *Revista gaúcha de enfermagem*, SciELO Brasil, v. 32, p. 759–766, 2011. Citado na página 19.

OLIVEIRA, T. A. et al. Comparação de random survival forest e modelo de cox com relação a performance de previsão: Um estudo de caso. *Sigmae*, v. 8, n. 2, p. 490–508, 2019. Citado na página 16.

PACHNICKI, J. P. A. et al. Avaliação imunoistoquímica dos receptores de estrogênio e progesterona no câncer de mama, pré e pós-quimioterapia neoadjuvante. *Revista do Colégio Brasileiro de Cirurgiões*, SciELO Brasil, v. 39, p. 86–92, 2012. Citado na página 21.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Citado na página 12.

SALLES, M. d. A. et al. Contribuição da imuno-histoquímica na avaliação de fatores prognósticos e preditivos do câncer de mama e no diagnóstico de lesões mamárias. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, SciELO Brasil, v. 45, p. 213–222, 2009. Citado na página 22.

SNIDER, B.; MCBEAN, E. Assessing the impact of pipe rehabilitation on decreasing watermain break rates using random survival forest models. *Water Resources Management*, 06 2022. Citado na página 16.

THERNEAU, T. M.; LUMLEY, T. Package 'survival'. *R Top Doc*, v. 128, n. 10, p. 28–33, 2015. Citado na página 12.