UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests

Danilo Barbosa da Silva de Oliveira

São Carlos – SP

Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests

Danilo Barbosa da Silva de Oliveira

Orientador: **Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho**

Monografia final de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação.
Área de Concentração: Inteligência Artificial, Aprendizado de Máquina

USP – São Carlos
Junho de 2020

Oliveira, Danilo Barbosa da Silva de
Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests / Danilo Barbosa da Silva de Oliveira. - São Carlos - SP, 2020.
53 p.; 29,7 cm.

Orientador: André Carlos Ponce de Leon Ferreira de Carvalho.
Monografia (Graduação) - Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos - SP, 2020.

1. Aprendizado de máquina. 2. Random Survival Forests. 3. Glioblastoma. 4. Análise de sobrevida. I. Carvalho, André Carlos Ponce de Leon Ferreira de. II. Instituto de Ciências Matemáticas e de Computação (ICMC/USP). III. Título.

Este trabalho é dedicado aos engenheiros e cientistas que,
por sua curiosidade, mudaram como a humanidade interage
com o universo.

AGRADECIMENTOS

Agradeço principalmente à minha família, minha mãe Eleonora, meu pai João Carlos e meu irmão Nicholas, que sempre me apoiaram em minhas escolhas e me instigaram a sonhar mais alto. Agradeço não só pelo imenso suporte que recebi durante o ciclo da graduação, mas também por cada momento que passamos juntos.

À minha namorada, Gabriela, por cada palavra de carinho e companheirismo durante todos esses anos.

Ao professor orientador André Ponce e ao Renato, que me instruíram e não pouparam atenção para a conclusão desse trabalho.

Aos meus grandes amigos feitos durante meus anos em república, aos quais guardo muito carinho e admiração, minha segunda família.

Aos amigos feitos durante os anos que passei no grupo Zenith, por compartilharem do mesmo sonho e todos os bons momentos juntos.

Aos meus amigos de sala, que estiveram juntos comigo durante os bons e maus momentos.

Aos meus amigos de longa data, que se mantiveram próximos mesmo em outras cidades.

À todos que direta ou indiretamente contribuíram para minha evolução até hoje.

“The people who are crazy enough
to think they can change the world,
are the ones who do.”
(Steve Jobs)

RESUMO

OLIVEIRA, D. B.. **Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests**. 2020. 53 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

O câncer cerebral do tipo Glioblastoma é um dos mais agressivos na atualidade, com alta taxa de mortalidade e expectativa média de sobrevivência de 15 meses após diagnóstico. Ao mesmo tempo, novos algoritmos especializados em análise de sobrevida foram descritos nos últimos anos, possibilitando trabalhar com os principais desafios encontrados em bases de dados médicos: grande número de variáveis com baixa quantidade de amostras e censura de dados. Esse trabalho tem o objetivo de avaliar o desempenho do *Random Survival Forests* (RSF) e de sua modificação mais recente, o *Maximally Selected Rank Statistics Random Forests* (MSR-RF), aplicados numa base de dados com informação genética (mRNA) de pacientes de Glioblastoma, considerando dois aspectos: capacidade de distinção de risco de pacientes (*C-index*) e precisão das curvas de sobrevivência estimadas (*Brier Score*). Ambos podem ser considerados adaptações do famoso algoritmo de aprendizado de máquina *Random Forests*, mas procuram maximizar a diferença de sobrevivência ao fazer um *split* num nó. Foi desenvolvida também uma biblioteca que encapsula diversas funções da análise de sobrevivência, além de métodos de avaliar a importância de variáveis e seleção de preditores, chamada de *SurvivalLib*. O MSR-RF apresentou melhor resultado para o C-index, 0,869, contra 0,727 do RSF. Para o *Brier Score*, ambos foram muito parecidos, com pontuação de 0,128 para o RSF e 0,123 para o MSR-RF. Gráficos das curvas de sobrevivência estimadas ao longo do tempo são apresentados, para todos os pacientes da base de testes. Neste caso, o RSF demonstrou maior separação entre pacientes de alto risco em comparação com os de baixo risco. Este estudo permitiu a comparação dos dois algoritmos, mostrando que o método adotado pelo MSR-RF apresentou melhor resultado em classificar o risco dos pacientes, mas o RSF mostra mais eficiência na precisão da estimativa da probabilidade de sobrevivência ao longo do tempo. Além disso, a criação da biblioteca *SurvivalLib*, utilizada neste trabalho, poderá facilitar novas pesquisas na área de análise de sobrevida.

Palavras-chave: Aprendizado de máquina, Random Survival Forests, Glioblastoma, Análise de sobrevida.

ABSTRACT

OLIVEIRA, D. B.. **Estimativa de Sobrevida de Pacientes com Glioblastoma por meio de Algoritmos Baseados em Random Forests**. 2020. 53 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Glioblastoma is one of the most aggressive brain cancer, showing a high mortality rate. The life expectancy after diagnosis is only 15 months. At the same time, new algorithms focused on survival analysis were described in the last years, which can handle the main problems encountered on medical databases: high number of columns with a low number of samples and censored data. This study aims to evaluate the performance of Random Survival Forests (RSF) and its recent modification: Maximally Selected Rank Statistics Random Forests (MSR-RF), applied to a database of glioblastoma patients containing genetic data (mRNA), considering two aspects: ability to separate patients risk (C-index) and survival functions estimations (Brier Score). Both can be assorted as modifications of the well known machine learning algorithm Random Forests, while trying to maximize survival difference at node splits. A new tool was implemented to wrap survival analysis functions, on top of common operations like variable importance processing and feature selection. The MSR-RF showed a better score for the C-index metric, with 0.869, while RSF got only 0.727. On the other hand, for the Brier Score metric, both performed alike, with a 0.123 score for the RSF and 0.128 for MSR-RF. The survival function was plotted for all cases, on all available time-frames, for all patients on the test set. In this case, the RSF demonstrated a better separation between high and low risk cases. This work allowed for a comparison of the two algorithms, indicating a better performance of MSR-RF on ranking patients risk, but RSF was more precise on estimating the survival function. The development of the *SurvivalLib* will help new research on survival analysis field.

Key-words: Machine Learning, Random Survival Forests, Glioblastoma, Survival Analysis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Motivação e Contextualização

1.1.1 Câncer

O câncer ainda é um dos maiores desafios que a humanidade enfrenta. Todo ano são contabilizados milhões de novos casos, que infelizmente levam à milhares de mortes, devido à alta taxa de mortalidade. Segundo a *GLOBOCAN* (*Global Cancer Observatory*), em 2018, foi estimado um total de 18,1 milhões de casos, com 9,6 milhões de mortes (BRAY *et al.*, 2018). Além disso, a doença ocorre em diversas partes do corpo humano, cada uma com suas particularidades. O tipo mais comum em diagnósticos depende bastante do estilo de vida da população, variando entre países. Câncer de pulmão, em termos globais, soma o maior número de casos, com cerca de 2 milhões em 2018, levando à 1,7 milhões de mortes (BRAY *et al.*, 2018). Outro exemplo que se destaca é o câncer cerebral e de sistema nervoso central, que está entre os mais agressivos. A maioria desses casos concentram-se na categoria Glioblastoma, um tipo de tumor que apresenta alta resistência aos tratamentos e apresenta uma taxa média de sobrevivência de 15 meses (STUPP *et al.*, 2005).

O tratamento padrão para o Glioblastoma é severo. A primeira opção é a cirurgia de remoção do tumor, seguida de radioterapia (ou radioncologia). Porém, nos últimos anos, pesquisas estão sendo realizadas para analisar o efeito da interação da *temozolomida* (TMZ), uma droga bastante usada contra o Glioblastoma, no tratamento dos pacientes, em conjunto com as técnicas já conhecidas. Este método apresenta bons resultados e conclusões estatísticas de sua eficácia (STUPP *et al.*, 2005). Não só, recentes estudos visam entender também a interação genética na prevenção e novos alvos terapêuticos para a doença (BLEEKER; MOLENAAR; LEENSTRA, 2012).

Uma abordagem para o estudo dessas interações de tratamento, bem como o impacto no tempo de sobrevida após o diagnóstico é pela informação genética do indivíduo (LOPEZ *et al.*, 2018). Essa análise é uma tarefa bastante complexa, por envolver milhares de agentes simultaneamente. Além isso, usar esse conhecimento para obter discernimento à respeito de estratégias de tratamento é ainda mais difícil, já que o estado de saúde de uma pessoa envolve seu estilo de vida, alimentação, frequência de exercícios físicos e mais inúmeros outros fatores. Dessa forma, métodos muito eficazes e robustos são necessários para analisar grandes quantidades de

dados, a fim de extrair informações úteis.

1.1.2 Análise de Sobrevida

O estudo de novos métodos de análise de sobrevida, campo que foca no estudo do tempo decorrido até um evento de interesse (KLEIN, 2003), é amplamente aplicado em diversas áreas do conhecimento, mas principalmente medicina e engenharia. Na primeira, o foco é voltado para análise da influência de novos tratamentos e drogas em pacientes, e na segunda, a importância de agentes internos e externos na durabilidade de equipamentos, como na prevenção de falhas em equipamentos (ALI *et al.*, 2015). Dessa forma, esses estudos são de extrema importância na evolução e aprimoramentos das técnicas e métodos utilizados dentro de cada área. Levando em conta o campo médico, observamos uma grande quantidade de trabalhos de aquisição de dados, em que, durante um período de tempo, pesquisadores coletam medidas clínicas e genéticas de pacientes, que são compiladas e publicadas em grandes veículos, como a plataforma cBio (CERAMI *et al.*, 2012) e (GAO *et al.*, 2013). Assim, é clara a facilidade de obtenção de bases de dados para realização de estudos de sobrevida de pacientes e, assim, contribuir com o progresso de soluções nessa área.

O aprendizado de máquina evoluiu consideravelmente nos últimos anos, nos quais diversas novas técnicas são apresentadas regularmente. Com algoritmos cada vez mais poderosos, apresentando boa acurácia em diversas aplicações, o poder de predição tem potencial de se tornar preciso, quando aplicada uma metodologia consistente e iterativa, observando, nos dados, as características que contribuem e atrapalham o aprendizado. Historicamente, foram desenvolvidos diversos métodos estatísticos que abordam o problema, separados em três grupos: métodos não-paramétricos, semi-paramétricos e paramétricos (WANG; LI; REDDY, 2019).

Apesar de muito eficientes em alguns casos, principalmente quando são observadas distribuições definidas nas variáveis analisadas, as técnicas de *Machine Learning* podem apresentar vantagens significativas em comparação aos métodos estatísticos mais tradicionais, como Cox, quando exploradas suas vantagens e controladas as desvantagens (DELEN; WALKER; KADAM, 2005). Dessa maneira, são publicados trabalhos com o objetivo de avaliar o desempenhos dessas novas técnicas e algoritmos em dados médicos (DATEMA *et al.*, 2011) e (KOUROU *et al.*, 2015), visando entender os fatores que mais influenciam, positivo quanto negativamente, no tratamento de pacientes com câncer.

Além de analisar as influências, é possível estimar a probalidade de sobrevivência de um paciente até um determinado tempo t, com uma abordagem de regressão (OMURLU; TURE; TOKATLI, 2009). Contudo, os algoritmos clássicos de regressão não se demonstram tão eficientes quando os dados apresentam *censoring* (censura), que ocorre quando não há uma conclusão nos dados de um paciente específico, como por exemplo sua desistência do estudo. Esse problema se deve ao fato de ser necessário excluir as amostras que apresentam censura, fato que será detalhado na Revisão Bibliográfica.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral avaliar o desempenho de algoritmos recentes de aprendizado de máquina, aplicados à base de dados com informações de pacientes portadores de câncer cerebral do tipo Glioblastoma, considerando a situação problema da análise de sobrevida, considerando o problema de *censoring*.

1.2.2 Objetivos específicos

Dentro dos objetivos podem ser elencados os seguintes tópicos:

- Revisão da literatura recente envolvendo análise de sobrevida e aprendizado de máquina.
- Elaboração de uma metodologia abordando principais conceitos revisados para tratamento de bases de dados com poucas amostras e muitas colunas.
- Treinamento dos modelos e validação das predições das curvas de sobrevivência.

1.3 Organização

Este trabalho está organizado em 4 capítulos, dos quais este é o primeiro. Em seguida, no CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA, será apresentada a fundamentação teórica dos tópicos trabalhados nesta pesquisa. Depois, no CAPÍTULO 3: DESENVOLVIMENTO, é apresentada a metodologia e a implementação do projeto é detalhada, além dos resultados encontrados. Por fim, no CAPÍTULO 4: CONCLUSÃO, são apresentadas as conclusões e são elencadas propostas para trabalhos futuros, além de considerações sobre o curso ao qual o autor está matriculado.

Capítulo 2

REVISÃO BIBLIOGRÁFICA

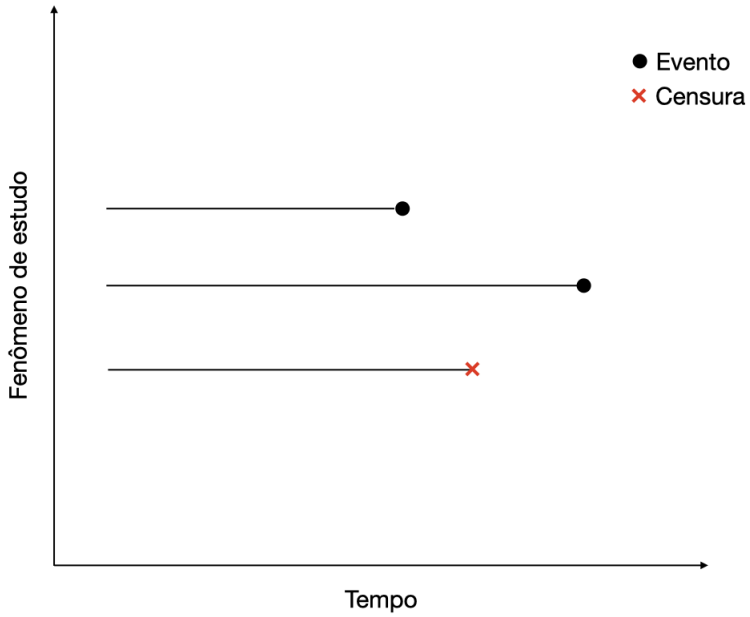
2.1 Considerações iniciais

Neste capítulo são apresentados os principais conceitos e terminologia trabalhados na literatura relacionada à análise de sobrevivência, com foco em soluções de aprendizado de máquina. São discutidos de forma mais profunda os métodos baseados em Árvore de Decisão (AD), como *Random Forests* (RF) e *Random Survival Forest* (RSF), que são a base desse trabalho.

2.2 Análise de Sobrevivência

Os estudos que visam analisar e modelar um determinado tempo **T**, em que espera-se que um evento de interesse ocorra, são classificados dentro do campo da estatística, como análise de sobrevivência (WANG; LI; REDDY, 2019). Em termos gerais, os métodos se distribuem em dois principais grupos: estatísticos ou aprendizado de máquina. Ambos tem o mesmo objetivo de estimar as curvas de sobrevivência para o fenômeno analisado, mas o primeiro foca no estudo das distribuições e parâmetros que o caracterizam, e o segundo foca na predição do evento de interesse. As análises, à primeira vista, assemelham-se à uma regressão comum, onde é pretendido obter um modelo capaz de prever um resultado numérico, à partir das variáveis de entrada. Porém, observa-se que em diversas áreas de estudo, a compilação de uma base de dados completa para análise posterior é uma atividade complexa, e muitas vezes, não é possível obter a informação do tempo total do fenômeno. Amostras são coletadas quando conveniente, durante a existência do fenômeno, mas nem sempre corresponde ao tempo de ocorrência do evento de interesse. Esse problema é denominado *censoring* (KLEIN, 2003). Esse problema é ilustrado na Figura 1 amostras de um fenômeno em estudo são coletadas ao longo do tempo, e em dois casos sabe-se o tempo exato que o evento de interesse ocorreu. Já no caso marcado em vermelho, a amostragem foi interrompida e não se sabe a duração total do fenômeno.

Figura 1 – Exemplo de ocorrência de censura nos dados.



Fonte: Elaborada pelo autor.

Segundo (LEE, 2003), existem três tipos de *censoring* :

1. *Right censoring*: Ocorre quando há informação sobre o início do fenômeno, mas não se sabe quando o evento de interesse ocorre.

2. *Left censoring*: Ocorre quando há informação sobre o evento, mas não se sabe o início do período de observação.

3. *Interval censoring*: Não se sabe o tempo exato do evento, somente que este ocorreu durante um intervalo.

De acordo com os objetivos desse trabalho, o foco se dá em métodos que solucionam o problema de *right censoring*, levando em conta que ao decorrer de um estudo com muitos pacientes, observamos que, por diversas razões, não se obteve o tempo de sobrevida para alguns indivíduos. Assim, ao utilizar um algoritmo de aprendizado de máquina supervisionado, seria necessária a exclusão de boa parte das bases de dados, na tentativa de possuir tuplas suficientes para treinamento e teste dos modelos. Essa é uma vantagem do uso de algoritmos que incorporam esse dado censurado no modelo, pois ainda que não há a informação conclusiva do tempo total de sobrevida, essa amostra pode contribuir com o aprendizado do modelo (DELEN; WALKER; KADAM, 2005).

2.3 Aprendizado de máquina

Dentro da grande área de aprendizado de máquina existem dois principais tipos de algoritmos utilizados na análise de dados:

- Aprendizado supervisionado: a variável dependente (alvo) está presente nos dados, e o modelo é treinado para estimar uma função que mapeia as variáveis de entrada para a variável alvo. Alguns exemplos: classificação, regressão e árvores de decisão.
- Aprendizado não supervisionado: a variável alvo não está presente nos dados, e a principal ideia nesse caso é a construção de agrupamentos que façam segmentação de classes diferentes dos dados. Alguns exemplos: Clusterização, K-NN.

A análise de sobrevivência caracteriza-se como **aprendizado supervisionado**, já que é utilizado um conjunto de dados de treino, composto por variáveis independentes (entrada) e uma variável dependente (saída), o qual o algoritmo escolhido usa para aproximar a função que mapeia essa relação (RUSSELL, 2010).

Portanto, como o objetivo desse trabalho é a elaboração de modelos preditivos, para estimar a sobrevida dos pacientes após o diagnóstico de câncer, é importante a utilização de uma base de dados com quantidade significativa de amostras contendo a informação do tempo de sobrevida. Entretanto, nos casos de pacientes que apresentam *censoring*, essa informação está ausente, prejudicando a quantidade de amostras úteis para treinamento e teste do modelo (WANG; LI; REDDY, 2019), ao passo que não é possível executar um modelo de aprendizado de máquina com dados faltantes, sem adicionar ruído ao sistema. A solução trivial de simplesmente retirar as amostras censuradas da base resulta em um modelo não ótimo (DELEN; WALKER; KADAM, 2005).

Em outro contexto, se as bases de dados com estudos de câncer estivessem disponíveis com grande quantidade de pacientes sem *censoring*, seria possível construir e testar modelos com algoritmos clássicos de regressão, removendo da base as amostras sem o evento de interesse. Porém, ao realizar essa operação, bastante informação é retirada sem contribuição para o entendimento do fenômeno.

Estudos de regressão para prever o tempo de evento em análise de sobrevivência podem ser feitos com diversas técnicas. Entretanto, quando o assunto do estudo se relaciona com a área médica, muitos algoritmos clássicos do aprendizado de máquina podem perder eficácia, principalmente pelos dois seguintes problemas: *Curse of dimentionality* (DONOHO, 2000) e Dados censurados. A primeira está relacionada com a natureza dos estudos médicos, com focos nos que possuem dados genéticos. A dificuldade (operacional e financeira) de realizar um estudo que envolva muitos pacientes em estados crítico de saúde é grande, além da complicação de tempo do estudo, que precisa se prolongar por anos. Assim, o comum das bases de dados que

disponibilizam dados genéticos é possuir poucas amostras. Ao passo que muitos algoritmos apoiam-se na premissa de um mínimo de exemplos para garantir aprendizado, ou seja, as distribuições das covariáveis abrangem um amplo espectro, apresentando ao algoritmo uma relação com a variável independente (se é que existe).

Além disso, o problema se torna ainda maior quando é considerado a dimensão da informação genética disponível. Existem diversas formas de se obter ciência da atuação de determinado gene. Uma delas é a observação da expressão de proteínas, que indicam a ativação genética produtora dessa molécula. A variedade da expressão proteica é muito grande, o que resulta em tabelas com milhares de colunas. Este é um problema bastante complexo na área de aprendizado de máquina, como descreve Mirza *et al.* (2019). Com os dois pontos apresentados, por fim, trabalha-se com um banco de dados de muitas covariáveis e poucas amostras (p grande e n pequeno), o que constitui o problema de *curse of dimentionality* (DONOHO, 2000).

O segundo problema é o dos dados censurados, nos quais não há informação de quando o evento ocorreu. A amostra foi recolhida enquanto o paciente estava em tratamento, e não houve uma conclusão, por qualquer razão. Como não existe a medida do tempo total de sobrevida, do diagnóstico até o falecimento, a amostra não poderia ser utilizada nos algoritmos clássicos de aprendizado supervisionado por não possuir rótulo. Por consequência, a censura dos dados agravaria ainda mais o problema de dimensionalidade descrito acima.

O tratamento de bases de dados desse tipo é uma tarefa complexa, quando se considera todos os aspectos mencionados acima. É comum a presença de covariáveis com amostras ausentes, mas que não têm importância significativa para prever a variável alvo. Nessa situação, é importante avaliar os impactos da adição ou remoção dessa *feature* no desempenho do modelo final. Neste caso, a decisão envolve, por um lado, incorporar no modelo a informação do problema que a variável entrega quando computada, mas por outro lado, perder as amostras que serão retiradas do modelo nos casos em que há valores faltantes.

A tecnologia utilizada na leitura da informação genética empregada neste trabalho é chamada de RNA-Seq (GOLDMAN; DOMSCHKE, 2014). Desenvolvida recentemente, apresenta grandes vantagens como grande precisão nas leituras de pares transcritos, boa qualidade da quantificação das amostras, quando comparada à técnica muito utilizada anteriormente, *Microarray* (WANG; GERSTEIN; SNYDER, 2009). Dessa forma, a RNA-Seq fornece um ótimo método para incorporar informação genética dos pacientes nos modelos preditivos.

2.4 Random Forests

Um antigo método de aprendizado de máquina é a construção de árvores de decisão (BREIMAN, 1993). Neste algoritmo, o objetivo é criar uma estrutura de decisão (árvore binária) que executa uma tarefa, como separar um banco de dados em duas classes diferentes. A decisão envolve a escolha de uma variável para separar os dados, a partir de um certo valor. A escolha

da melhor variável para fazer essa separação (*split*) é feita medindo a qualidade dos dois nós filhos, utilizando uma métrica pré-definida. A mais comum é a pureza dos nós filhos, calculada por exemplo pelo método Gini.

O *Random Forests* consiste em um *ensemble* (conjunto) de árvores de decisão. O primeiro passo é a mecânica de *splits*, na qual o algoritmo seleciona uma variável para segmentar a base de maneira que a separação promove similaridade entre os blocos resultantes. Esta divisão tem por objetivo aumentar a homogeneidade dos nós filhos, em comparação com o nó pai. A maximização da homogeneidade, originalmente proposta por (BREIMAN, 2001), refere-se à pureza dos nós filhos, que pode ser calculada, por exemplo, pela quantidade de amostras da mesma classe dentro do mesmo nó. Assim, nós com classes semelhantes apresentam maior pureza.

Um ponto inovador do RF é a introdução de aleatoriedades por dois processos: *boostrap* (EFRON, 1994) e dentro de cada nó, na seleção da variável para *split*. O *boostrap* consiste em fazer uma amostragem dos dados ao construir cada árvore, de modo que cada amostra retirada é reposta na base. O processo é repetido para o mesmo número de amostras na base de dados. Denotando a probabilidade de uma amostra x_i ser escolhida, dentro de um conjunto de n amostras por $1/n$, a probabilidade de x_i não ser escolhido é:

$$\pi_i = 1 - \frac{1}{n} \tag{2.1}$$

Expandindo essa probabilidade $p_i(n)$ para a enésima amostra retirada no processo:

$$p_i(n) = \prod_{j=1}^{n} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\ldots\left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right)^n \tag{2.2}$$

Considerando n grande, é fácil observar que $p_i(n)$ tende para e^{-1}, que é aproximadamente $0,368$. Essa fração dos dados não é usada na construção de cada árvore, provendo assim uma parcela de amostras para teste. Esse conjunto de amostras é chamado de amostra *Out of Bag* (OOB).

O segundo processo de introdução de aleatoriedade envolve a seleção de um subconjunto de covariáveis para fazer o *split*, ao invés de testar todas as possibilidades. A vantagem é o desacoplamento entre as árvores, diminuindo a variância do *ensemble* construído. Esta estratégia é conhecida como *bagging* (BREIMAN, 1996), e pode aumentar consideravelmente a acurácia do algoritmo base que foi replicado.

O procedimento completo do RF é descrito nos seguintes passos (ISHWARAN *et al.*, 2008):

1. Realizar n amostragens nos dados originais com *boostrap*.

2. Para cada novo conjunto, construir uma árvore de decisão, realizando uma seleção aleatória de covariáveis em cada nó.
3. A escolha da covariável é feita maximizando a homogeneidade em cada nó.
4. Construir a árvore repetindo esse processo recursivamente, até que cada nó folha não tenha menos de m_0 ocorrências.
5. Agregar a contribuição de cada árvore, calculando a média entre os resultados, no caso de um *ensemble* regressivo.
6. Calcular o erro com as amostras OOB.

2.5 Random Survival Forests

Visando atacar o problema de *right censoring*, foram desenvolvidas diversas técnicas, como a adaptação do RF proposta por Ishwaran *et al.* (2008), o *Random Survival Forest* (RSF). Esse algoritmo apresenta uma mudança na maximização da homogeneidade no momento do *split*, medindo a eficácia não mais por pureza, mas sim por diferença de sobrevida. Dessa forma, situações diferentes são separadas em nós distintos. Esse processo é realizado recursivamente, populando a árvore durante o crescimento com casos similares de sobrevida.

Além de operar diretamente sobre o tempo de sobrevida, uma outra grande vantagem do RSF é a possibilidade de incorporação de variáveis censuradas. Este algoritmo não usa o clássico padrão de entrada X e saída y. Nele, são consideradas três variáveis: X, contendo as variáveis de entrada, T, contendo o tempo de sobrevida e E, variável *booleana* informando se a amostra é censurada ou não.

O uso da informação de censura é aplicado principalmente no momento do cálculo do *split*, cuja decisão da variável x e valor do *split* em x é dado pela função log-rank Equação 2.3 (ISHWARAN *et al.*, 2008):

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}} \tag{2.3}$$

As variáveis assumem os seguintes significados:

- j: Nó filho.
- Y_i: Número de amostras sem censura ou em risco, em t.
- $Y_{i,j}$: Número de amostras sem censura ou em risco, para o nó filho, em t.
- d_i: Número de eventos, em t.

- $d_{i,j}$: Número de eventos, para o nó filho, em t.

Dessa forma, é possível utilizar a informação censurada para construir aprendizado ao modelo, o que pode melhorar o desempenho do mesmo. É importante mencionar que a saída produzida com o RSF é a *Survival Funcion*, função que mostra a probabilidade do paciente sobreviver após um tempo t de teste.

Um outro método recente proposto por Wright, Dankowski e Ziegler (2016) envolve uma outra forma de realizar o processamento de bases com tempo de sobrevida. A estrutura do algoritmo, no geral, é mesma do RSF. Serão usadas as mesmas três variáveis para o treinamento, X, T e E. Porém, a construção do modelo é baseada em inferência estatística condicional, utilizando um teste de hipótese para selecionar as variáveis no momento do *split*. O objetivo é remover ou diminuir o viés ao escolher a variável, que ocorre ao utilizar o método anterior com a formulação de *log-rank*, pois este tende a favorecer variáveis com muitas pontos para o *split*. A formulação e construção do algoritmo pode ser encontrada com mais detalhes em Wright, Dankowski e Ziegler (2016). Esse método é denominado *maximally selected rank statistics random forests* (MSR-RF).

Outro quesito importante é a forma de cálculo do risco dos pacientes, que também pode ser gerada pelos dois modelos. A medida de risco é calculada segundo a equação Equação 2.4, em que H corresponde à *hazard function*, que é a função que mede a probabilidade do evento ocorrer logo depois do tempo T testado, e J é o total de pontos temporais usado pelo modelo:

$$r(x) = \sum_{j=1}^{J} H(t_j, x) \tag{2.4}$$

2.6 Considerações Finais

Os dois modelos apresentados, RSF e MSR-RF, apresentam aplicação direta para o problema de análise de sobrevida com dados de câncer, com o objetivo de estimar a função de sobrevivência para cada paciente, bem como analisar a situação de risco que estes se encontram. Estudos envolvendo aprendizado de máquina com informação genética e predição de sobrevida são encontrados na literatura, porém em baixíssima quantidade, além de não apresentar a comparação com o mais recente MSR-RF (OMURLU; TURE; TOKATLI, 2009), (LOPEZ *et al.*, 2018), (DELEN; WALKER; KADAM, 2005). Ao incorporar os dados censurados no aprendizado do modelo, é natural a ocasião para observar o comportamento desses algoritmos com bases de dados relacionadas ao Glioblastoma. O Próximo Capítulo detalhará a implementação e uso de ambos algoritmos para análise de sobrevida.

Capítulo 3

DESENVOLVIMENTO

3.1 Considerações Iniciais

O presente Capítulo discorrerá à respeito do desenvolvimento do projeto elaborado. Inicialmente, os passos da metodologia adotada serão descritos detalhadamente. Em seguida, a ferramenta concebida será apresentada, bem como todos os seus módulos e como estes atuam em conjunto para a solução do problema. Por fim, serão apresentados os resultados.

3.2 Metodologia

3.2.1 Aquisição de dados

O primeiro passo realizado no desenvolvimento do projeto foi a obtenção das bases de dados com informações clínicas e genéticas. A fonte escolhida foi a plataforma cBio (CERAMI *et al.*, 2012), que disponibiliza gratuitamente diversas pesquisas realizadas na área oncológica. Além disso, o sistema conta com uma ferramenta de consulta genética, na qual é possível buscar, em vários estudos ao mesmo tempo, diversos genes de interesse. Ainda, a plataforma disponibiliza várias formas de visualização de dados, com diversos artifícios gráficos para cada tipo de variável analisada. Usualmente, os estudos são divididos em diversas tabelas, com medições de diferentes propriedades clínicas dos pacientes. As duas principais que são utilizadas nesse trabalho são a tabela de dados clínicos, que contém informações gerais sobre o indivíduo, como idade no momento do diagnóstico e gênero. A segunda refere-se à informação genética do paciente, apresentando, para cara gene, um valor de mutação em relação à um *baseline*. Inúmeras métricas são oferecidas, mas a escolhida para o projeto foi a de *mRNA Z-scores* (escore padrão), que exibe a métrica estatística da quantidade de desvios padrões que a amostra está em comparação com a média das leituras por RNA-Seq. Dessa forma, os dados são normalizados pela média, prática que facilita a manipulação das bases e facilita a identificação de casos extremos (CHEADLE *et al.*, 2003).

A base de dados escolhida é a referente ao câncer do tipo Glioblastoma, que atua no cérebro. É considerado o mais comum, porém mais agressivo câncer cerebral, no qual o paciente tem uma expectativa mediana de vida de 15 meses (BLEEKER; MOLENAAR; LEENSTRA, 2012). A plataforma cBio disponibiliza seis estudos referentes ao Glioblastoma.

A base selecionada é a do programa de pesquisas TCGA (*The Cancer Genome Atlas*), de 2013 (BRENNAN, 2013). Este estudo é interessante pois, dentre as *features* disponíveis, está o tratamento com a droga *temozolomide*, ou *TMZ*, que consiste no tratamento mais recente e é considerada tratamento padrão para pacientes recém diagnosticados (BLEEKER; MOLENAAR; LEENSTRA, 2012). A base possui um total de 543 pacientes, dos quais 152 dispõem de informação genética.

3.2.2 Pré-processamento de dados

Após feito o *download* das bases de dados, é necessário fazer uma limpeza inicial, removendo covariáveis ruidosas, com muitas amostras faltantes. Este é o caso da tabela de dados clínicos, na qual pode ser observadas lacunas de tamanho significativo nas amostras. A fim de automatizar esse processo, é tomada a decisão de empregar a biblioteca *Pandas Profiling* [1], disponibilizada para as versões mais recentes de *Python*.

Esta ferramenta possibilita o processamento automatizado de bases de dados, oferencendo informações referentes à diversos pontos importantes dentro de cada variável, como:

1. Contagem de itens distintos: utilizado para identificar variáveis índice na base;
2. Porcentagem de amostras únicas: útil na observação da distribuição de variáveis categóricas;
3. Porcentagem de amostras faltantes: análise de variáveis problemáticas;
4. Distribuição em histograma de variáveis numéricas: comportamento de variáveis numéricas;
5. Avisos de variáveis problemáticas: identificação de variáveis com lacunas;
6. Correlação de variáveis: analisar possíveis exclusões;
7. Gráficos de dispersão de variáveis: identificar interações;

Dessa forma, é possível, rapidamente, observar o comportamento geral dos dados e o comportamento das variáveis, bem como suas interações. Tendo em vista os itens elencados acima, o mais interessante para essa fase é o item 5, possibilitando a rápida eliminação de variáveis que possuem lacunas maiores que 5%. É claro que, antes de eliminar, é vantajosa a analise do impacto da variável no modelo final. Porém, nesse caso, como a quantidade de amostras é bem reduzido (152 amostras possuem dados genéticos), perdas pequenas resultarão em alto impacto negativo no modelo, devido à *curse of dimentionality* (DONOHO, 2000).

[1] <https://github.com/pandas-profiling/pandas-profiling>

Em posse das bases clínica e genética, ambas identificadas por uma variável única *sample_id*, é necessária a remoção das *features* explicativas indicadas pelo *pandas profiling*, de acordo com o *threshold* indicado acima. São elas:

1. G-CIMP_METHYLATION;
2. GENE_EXPRESSION_SUBTYPE;
3. IDH1_MUTATION;
4. METHYLATION_STATUS;
5. MGMT_STATUS;
6. FRACTION_GENOME_ALTERED;
7. MUTATION_COUNT;
8. DISEASE_FREE_(MONTHS);
9. DISEASE_FREE_STATUS;

Os itens 1-7 acima sofrem de lacunas excessivas e devem ser removidas da base de dados clínicos. Essa decisão deve-se ao fato de ser necessário possuir os dados completos ao treinar e testar o modelo. Já os itens 8 e 9 são removidos pois diretamente relacionado à variável alvo, o tempo de sobrevida dos pacientes, identificada por *OVERALL_SURVIVAL_(MONTHS)*. Desse modo, são removidas para não causar ruído desnecessário no modelo e prejudicar o efeito de outras *features*.

Podem haver casos em que um mesmo paciente foi analisado duas vezes, em tempos diferentes. Nesse caso, são expostas duas amostras com o mesmo identificador de paciente. Em particular, para a TCGA 2013, esse fato não ocorre.

A base de dados com as referências para o mRNA *Z-scores* precisa de mais alguns passos de pré-processamento, como o ajuste na nomenclatura dos genes (é apresentada tanto o padrão por *Hugo Symbols* quando por *Entrez Gene ID* (MAGLOTT *et al.*, 2010)), removendo espaços e alterando os caracteres para maiúscula. Todas as manipulações são feitas utilizando a biblioteca *Pandas*[2] em ambiente de desenvolvimento *Python*.

Um ponto importante que também precisa ser feito durante a fase de pré-processamento é aplicar um limite superior de tempo que será considerado para o estudo. A distribuição natural dessa base apresenta um número considerável de indivíduos com tempo de sobrevida acima do normal para o Glioblastoma. Esses casos, embora de excelente natureza para os pacientes e o campo da oncologia, distorcem o conjunto de dados. A metodologia empregada envolve a

[2] <https://pandas.pydata.org/>

observação da distribuição do tempo de sobrevida, comparando o desvio padrão com a média. Serão excluídos todas as amostras que exibirem tempo de sobrevida maior que um desvio padrão mais a média. Após aplicar esse filtro, a quantidade de amostras resultante é 132.

3.2.3 Random Survival Forests

Na abordagem de um problema de aprendizado de máquina envolvendo poucas amostras, como é o presente caso, a presença de dados com censura levaria à perda de diversas amostras, se adotado os algoritmos clássicos, como foi apontado no capítulo anterior. No caso comum de regressão do aprendizado supervisionado, todas as amostras precisam de um rótulo para haver aprendizado. Como as amostras censuradas carecem de rótulo, deveriam ser excluídas na fase de pré-processamento. Porém, ao analisar a quantidade de amostras classificadas com evento é 99, contra 53 sem rótulo. Em outras palavras, cerca de 35% das amostras seriam retiradas caso fosse necessária a exclusão por falta de rótulo.

Embora não seja conhecida a causa da censura para cada uma das amostras, ainda existe informação nas mesmas, principalmente pela presença do tempo. O período do diagnóstico até a realização da coleta de exames e posterior inserção na base é conhecido. Esse dado agrega valor ao modelo e pode ser usado para melhorar os resultados (WANG; LI; REDDY, 2019). Dessa forma, a construção do algoritmo *Random Survival Forests* é especialmente focada nesse problema, e consegue utilizar as amostras censuradas no treinamento, otimização e testes.

O conceito da adaptação do *Random Forests* para análise de sobrevivência, o *Random Survival Forests*, envolve o uso de não somente dois conjuntos de valores - entrada e saída - mas sim três: Conjunto de *features*, tempos de sobrevida e um vetor indicador do evento. O primeiro, usualmente chamado de variáveis independentes, não tem diferença quando comparado ao comum do aprendizado de máquina. O segundo pode ser comparado ao vetor rótulo num caso trivial de regressão: é o vetor com as medidas, esperando-se que a modelagem consiga mapeá-las com através das *features*. Já o terceiro corresponde ao aspecto menos usual: é o vetor que indica se a amostra indica o evento, ou seja, que o tempo de sobrevida é final e conclusivo. Nos casos em que o evento não é indicado, observa-se o caso de censura do dado (ISHWARAN *et al.*, 2008). A implementação utilizada nesse trabalho foi feita em *Python*, pela *PySurvival*, e é apresentada em (FOTSO *et al.*, 2019).

Levando em conta a necessidade da nova estrutura, a base de dados foi divida em três novos componentes: X, para indicar o conjunto de variáveis independentes, T, para a variável alvo, copiada de *OVERALL_SURVIVAL_(MONTHS)* e, por fim, E, que indica a presença de evento ou censura no dado, sendo atribuído o valor 1 para evento e 0 para censura. Essa informação é retirada da variável *OVERALL_SURVIVAL_STATUS*.

Dessa forma, agora o conjunto de dados que é utilizado para todas as funções de treinamento, otimização e teste, refere-se ao conjunto de dados formado pelas três novas variáveis X,

T e E.

3.2.4 Variable Importance

O próximo passo da metodologia é reduzir a dimencionalidade da base de dados. Usualmente os dados clínicos não apresentam mais de 100 *features*. Por outro lado, a base de *mRNA Z-scores* pode conter uma quantidade bem maior de variáveis independentes, como no caso da TCGA 2013, 19979. O custo computacional é muito alto para processar essa quantidade de *features*, além do fato que a contribuição de cada uma para o aprendizado do modelo diminui (DONOHO, 2000).

A metodologia escolhida para abordar esse problema foi reduzir o número de variáveis de acordo com um ranking de importância, calculado de acordo com o CSF ou o RSF. Por serem algoritmos baseados em *Random Forests*, ambos podem calcular a métrica VIMP (*Variable Importance*) ao calcular os efeitos no erro de predição, com a adição de ruído nas variáveis (BREIMAN, 2001). Esse método permite que seja processada uma tabela com as variáveis da base, ordenadas de forma descendente por ordem de importância. Assim, a primeira será a mais influente na previsão da variável alvo, e a última será a menos influente.

Dessa forma, a primeira tarefa é o processamento da base, com todas as variáveis independentes. É um processamento custoso, pois mesmo com poucas amostras, o algoritmo possui várias etapas que levam mais tempo para completar em função do número de colunas e da quantidade de árvores construídas. Além disso, para determinar a importância das variáveis, podemos usar a base inteira, já que não é um processo que será usado futuramente para predição. Assim, garantimos o maior uso possível da informação para determinar os melhores preditores da variável dependente. Ademais, o RF apresenta um ótimo resultado em situações problema com p grande e n pequeno, ao passo que impõe regularização das árvores, possibilitando uma inferência adaptativa mais robusta (CHEN; ISHWARAN, 2012).

É válido comentar que existem diversos métodos de cálculo da importância das variáveis em um algoritmo de RF. Considerando a implementação utilizada nesse projeto, foi escolhido o método descrito por (SANDRI; ZUCCOLOTTO, 2008), que pode ser selecionado colocando *impurity* como valor do parâmetro *importance_mode* presente no momento de treinar o modelo.

3.2.5 Feature Selection

Em posse da tabela com o ranking de importância das variáveis, é preciso estabelecer um método para selecionar um conjunto alvo pequeno com o menor número possível de preditores, mas que ainda mantenha um bom resultado. O racional dessa fase é iniciar com um pequeno grupo, com as melhores variáveis, testando o desempenho do modelo com uma validação cruzada de 5 *folds*. Feito o primeiro teste, o resultado é guardado e gradativamente o conjunto é diminuído de cerca de 20% das variáveis. Dessa forma, a cada nova iteração, é esperado um dos

dois cenários:

1. A capacidade preditiva do modelo é aumentada, com o diminuição de ruído.
2. A capacidade preditiva do modelo é diminuída, pela falta de informação preditiva.

A metodologia inicia a partir de computada a primeira tabela com a importância das variáveis. Em posse do *ranking* geral com todas os 19979 preditores, é empregada a técnica de seleção de variáveis descrita por (CHEN; ISHWARAN, 2012), com os seguintes passos:

1. Remover uma porção das variáveis menos influentes (cerca de 20%)
2. Processar novamente a nova base com menos variáveis e repetir o passo 1.
3. Continuar processo até obtenção do menor erro com o conjunto de amostras OOB.

Esse modo de operação é considerado um método guloso, mas como aqui o número de testes será baixo, não há um impacto significativo em desempenho. O algoritmo pode testar algumas centenas de possibilidades de conjuntos dentro de alguns minutos. Terminada essa fase, a dimensionalidade do problema será drasticamente reduzida, já que, no início, o conjunto contava com quase 20.000 variáveis, e, ao fim, é otimizado com menos de 50.

Nesta fase também foi considerada a possibilidade de utilizar um métodos de redução de dimensionalidade como *Principal Component Analysis* (PCA) (WOLD; ESBENSEN; GELADI, 1987). Esse método estima uma nova base ortogonal para o conjunto de dados, diminuindo a correlação. A nova base é chamada de componentes principais. Porém, um aspecto importante é a interpretabilidade do modelo, ou seja, quão simples é o entendimento das variáveis e de sua influência no resultado preditivo. Ao processar as variáveis com o PCA, não se trabalha mais com as variáveis originais, mas com seus componentes gerados. Não é trivial a interpretação do resultado do PCA, portanto o uso desse algoritmo não foi adotado.

3.2.6 Medidas de Desempenho

Levando em consideração o caso específico do aprendizado de máquina em que o presente problema se encontra, as clássicas métricas de avaliação de desempenho não funcionam, devido ao problema da censura dos dados. Numa amostra sem a informação de resultado não é possível checar a distância entra a previsão do modelo e o valor esperado (HEAGERTY; ZHENG, 2005). Por esse motivo, outras métricas foram desenvolvidas a fim de solucionar esse problema, possibilitando avaliar de modo mais robusto o comportamento dos modelos de análise de sobrevivência. Considerando tais fatos, foram selecionadas as seguintes métricas para a análise deste estudo:

1. C-index (UNO *et al.*, 2011).

2. Brier Score (GRAF *et al.*, 1999).

Chamado também de estatística C, o *C-Index* mede a capacidade do modelo em discernir, entre duas instâncias, qual tem maior risco. Na aplicação em análise de sobrevivência, esse conceito se encaixa muito bem, ao proporcionar uma forma de media a qualidade de segmentação do modelo. É muito interessante essa categorização de pacientes entre baixo e alto risco, podendo levar à decisões de tratamentos mais focados em cada um dos casos.

O algoritmo de cálculo do *C-Index* funciona da seguinte forma: são formados pares entre todas as amostras que serão testadas. Depois, o modelo gera as estimativas de risco para cada amostra. Por fim, as estimativas são comparadas entre os pares formados anteriormente. Porém, são deficidas regras para tratamento das censuras: uma amostra censurada só pode ser comparada com outra sem censura, com menor duração, pois como não se sabe o tempo total antes do evento para a amostra censurada, não há conclusões depois da censura (UNO *et al.*, 2011). Um modelo que obteve algum aprendizado apresenta um valor para a métrica *C-Index* maior que $0,5$.

Já a segunda métrica, *Brier Score*, mede, de forma similar ao erro quadrático médio, a distância entre a probabilidade de sobrevivência (saída do modelo) e o status atual (real) do paciente, para um tempo T de teste (GRAF *et al.*, 1999). É uma forma de analisar as curvas de sobrevivência que o modelo produz em relação a realidade, ou seja, comparar a qualidade da previsão individual do modelo para cada paciente. Para o *Brier Score*, um modelo útil tem um valor nessa métrica menor que $0,25$.

Ambas são usadas para comparar a eficácia dos modelos, em todos os cenários testados.

3.3 Atividades Realizadas

3.3.1 Survival Library

A partir desse ponto no desenvolvimento do projeto, todas as funções e processamentos necessários foram incorporados na *SurvivalLib*[3], uma biblioteca para *Python* que encapsula todas as funções necessárias da metodologia aqui apresentada. Fornecida uma base completa, com a informação genética e também dados clínicos, a ferramenta possibilita ao usuário experimentar e testar configurações diversas, treinar diferentes modelos e avaliar os resultados.

Elencados todos os fatores para filtragem da base, além da seleção das variáveis que serão utilizadas na modelagem, é o momento de fazer o treino propriamente dito. Primeiramente, ao instanciar um objetivo da classe *SurvivalLib*, são apresentadas algumas opções de argumento para o usuário. São elas:

[3] <https://github.com/danilobso/tcc_cancer_survival>

1. Base de dados que será utilizada (pandas DataFrame).
2. Coluna para ser utilizada como alvo do modelo (está e a variável com os tempos de sobrevida).
3. Coluna para gerar o vetor de censura. Deve informar se o evento ocorreu ou não.
4. Lista com as colunas que não serão utilizadas como *features*, como ID, coluna alvo.
5. Tabela com o ranking de variáveis importantes (opcional, se o processamento já foi realizado).
6. Melhor número de variáveis para uso no modelo (tamanho do conjunto).
7. Modelo que será usado: RSF para *Random Survival Forests*, ou MSR-RF para *Maximally Selected Rank Statistics Random Forest*.

Assim, que o objeto é criado, a biblioteca gera também a coluna de censura, a partir do item 3 acima. Dessa forma, será guardada uma variável contendo 1 ou 0, em que 1 indica ocorrência do evento e 0 indica censura. Além disso, na inicialização, já é computada uma sugestão de máximo de tempo para ser filtrado a coluna alvo. Esse limite é importante ao passo que muitas amostras *outliers* nesta variável pode prejudicar o *performance* do modelo. A sugestão é calculada de acordo com a metodologia apresentada previamente. Por fim, a inicialização atribui um valor para o percentual dos dados que serão usados como amostra de testes. Essa porção da base não será utilizada em nenhum momento para treinamento, somente validação.

Criado o objeto da biblioteca, o segundo passo é processar a limpeza da base. Aqui, também são aceitos novos argumentos, que informam os seguintes aspectos:

1. Mínimo de tempo para filtro do tempo de sobrevida. Útil se existem *outliers* no início do eixo temporal;
2. Máximo de tempo para filtro de tempo de sobrevida. É sugerido utilizar o resultado anterior;
3. Colunas que não serão utilizadas na análise. Aqui, é importante experimentar, pois algumas colunas não acionam informação suficiente para justificar a exclusão de algumas amostras.

Em posse dessas informações, as colunas passadas como parâmetro são excluídas, e é executado uma exclusão de amostras restantes que têm alguma entrada faltante. É também executado o filtro de tempo, tanto para o filtro mínimo quanto o máximo. Ambos são executados como *maior igual* ou *menor igual*. Por fim, é processada a divisão da base de dados em treinamento e teste. A separação é feita com o parâmetro de porcentagem dos dados totais usado para teste, guardado no momento da inicialização da biblioteca.

Antes da divisão em treinamento e testes, o algoritmo verifica a necessidade de realizar o *encoding* em variáveis categóricas, processo conhecido como a geração de *dummies*, no qual cada categoria diferente é transformada em uma coluna e as amostras pertencentes à cada uma recebem o número 1 para a categoria correta. Assim, é possível trabalhar com as variáveis categóricas nos algoritmos de aprendizado de máquina (RUSSELL, 2010). É importante fazer esse procedimento antes da separação dos dois conjunto, pois pode haver o problema das colunas ficarem diferentes caso haja discrepância entre as amostras de cada um.

O processamento do *ranking* de importância de variáveis é feito com a base inteira, pois nesse caso não há prejuízo para as métricas de avaliação, já que é um processo prévio ao treinamento. É executada a metodologia acima, em que é realizado um *fit* (treinamento) do algoritmo desejado, e removidas 20% das que têm menos influência no modelo, através da métrica de *Variable Importance* (VIMP) do próprio modelo, calculada inserindo ruído em cada uma das variáveis e analisando o resultado na variável dependente. É o processo mais custoso, ao passo que trabalha com quase 20.000 preditores ao mesmo tempo.

Depois desse passo, já é possível fazer o primeiro treinamento do modelo com possibilidade de teste. O padrão é a utilização de 3.000 árvores, número que apresenta um bom balanço entre qualidade do resultado e tempo de processamento.

Os dois modelos apresentados no Capítulo 2 estão disponíveis para teste, e funcionam do mesmo modo, não é necessário nenhuma modificação por parte do usuário, após escolher o modelo desejado ao inicializar a biblioteca.

Para otimizar o treinamento e evitar problemas de mínimo local da função de perda, é feito também o processamento de uma *Grid Search*, na qual os hiperparâmetros do modelo são arranjados de forma que todas as combinações são testadas. São eles:

Para o RSF:

1. *max_features*: Número mínimo de amostras em um nó folha.
2. *min_node_size*: Máximo de covariáveis testadas para fazer o *split* de um nó.
3. *sample_size_pct*: Porcentagem das amostras originais usadas em cada árvore.
4. *max_depth*: Profundidade máxima da árvore. Controla *overfitting* no algoritmo.
5. *num_trees*: Número de árvores que farão parte do *ensemble*.

No caso do MSR-RF, são adicionados dois novos parâmetros, além dos três anteriores:

1. *alpha*: Nível de significância mínimo para fazer o *split* dos nós.
2. *minprop*: Menor quantil que será considerado para o *split*.

Outro ponto que é calculado pelo modelo é uma pontuação de risco para cada paciente. Esse número é calculado com base na *hazard function*, a função que mede a probabilidade de ocorrência do evento dentro de $T + dt$, ou seja, imediatamente depois de **T**. O equacionamento dessa métrica pode ser conferida na Equação 2.4.

3.4 Resultados

Primeiramente, os resultados foram gerados utilizando a metodologia apresentada neste Capítulo, com a biblioteca *SurvivalLib*. Além disso, as duas métricas escolhidas: *C-Index* e *Brier Score* são computadas, para medir duas características diferentes dos modelos, que é a capacidade do modelo em discernir entre pacientes de alto e baixo risco, e o erro do modelo em estimar a curva de sobrevivência para cada paciente, respectivamente.

Dessa modo, foram organizados 4 experimentos, considerando os dois algoritmos e as duas métricas. Para cada modelo, foi gerada uma tabela de importância de variáveis, que foi usada para cálculo das duas medidas de desempenho. Os hiperparâmetros do modelo e o número ótimo de variáveis explicativas são computados individualmente para cada caso. Vale lembrar que o conjunto de dados utilizado em todos os casos é exatamente o mesmo.

A tabela Tabela 1 apresenta os resultados comentados acima:

Tabela 1 – Desempenho dos modelos por métrica.

Métrica	RSF	MSR-RF
C-Index	0,727	0,869
Brier Score	0,128	0,123

Fonte: Dados da pesquisa.

Aqui, pode-se observar que o MSR-RF apresentou um resultado consideravelmente melhor para o *C-index*, demonstrando um bom desempenho para escolher entre pacientes de baixo e alto risco. Como a principal diferença entre os dois modelos é o método de seleção de variáveis, a influência dessa operação com menor *bias* do MSR-RF parece contribuir com o aprendizado para o problema de previsão de sobrevivência para o Glioblastoma. Um ponto importante é também a observação das variáveis mais importantes. Nos dois casos, a terapia utilizada é determinante para o aprendizado do modelo e também para a extensão do tempo de sobrevida. Para o MSR-RF, a terapia com TMZ foi o preditor mais influente. Este resultado é relevante ao passo que, na literatura médica, esse é o tratamento de melhor eficácia para esse tipo de câncer (STUPP *et al.*, 2005).

Para o caso do *C-index*, um modelo útil precisa ter uma pontuação maior que $0,5$. Em ambos os casos, o resultado foi maior do que esse *threshold* esperado, e a par com outros estudos médicos como Omurlu, Ture e Tokatli (2009), em que foi feito um estudo de modelagem e predição para dados de câncer de mama. Na pesquisa, são apresentados os resultados para um

modelo de RSF, com o *C-index* próximo de $0,7$. Dessa forma, os resultados obtidos com a utilização da *SurvivalLib* estão de acordo com a literatura da área.

São apresentados os parâmetros após todas as etapas de otimização dos dois modelos, para o *C-index*, na Tabela 2:

Tabela 2 – Parâmetros otimizados dos modelos para métrica C-Index.

Parâmetro	RSF	MSR-RF
max_features	sqrt	sqrt
min_node_size	7	10
max_depth	5	6
min_survival_months	0	0
max_survival_months	21	21
sample_size_pct	0,63	0,63
alpha	-	0,5
minprop	-	0,12
num_trees	3000	3000
num_features	8	41

Fonte: Dados da pesquisa.

Tabela 3 – Parâmetros otimizados dos modelos para métrica Brier Score.

Parâmetro	RSF	MSR-RF
max_features	sqrt	sqrt
min_node_size	7	8
max_depth	5	5
min_survival_months	0	0
max_survival_months	21	21
sample_size_pct	0,63	0,63
alpha	-	0,5
minprop	-	0,08
num_trees	5000	5000
num_features	11	28

Fonte: Dados da pesquisa.

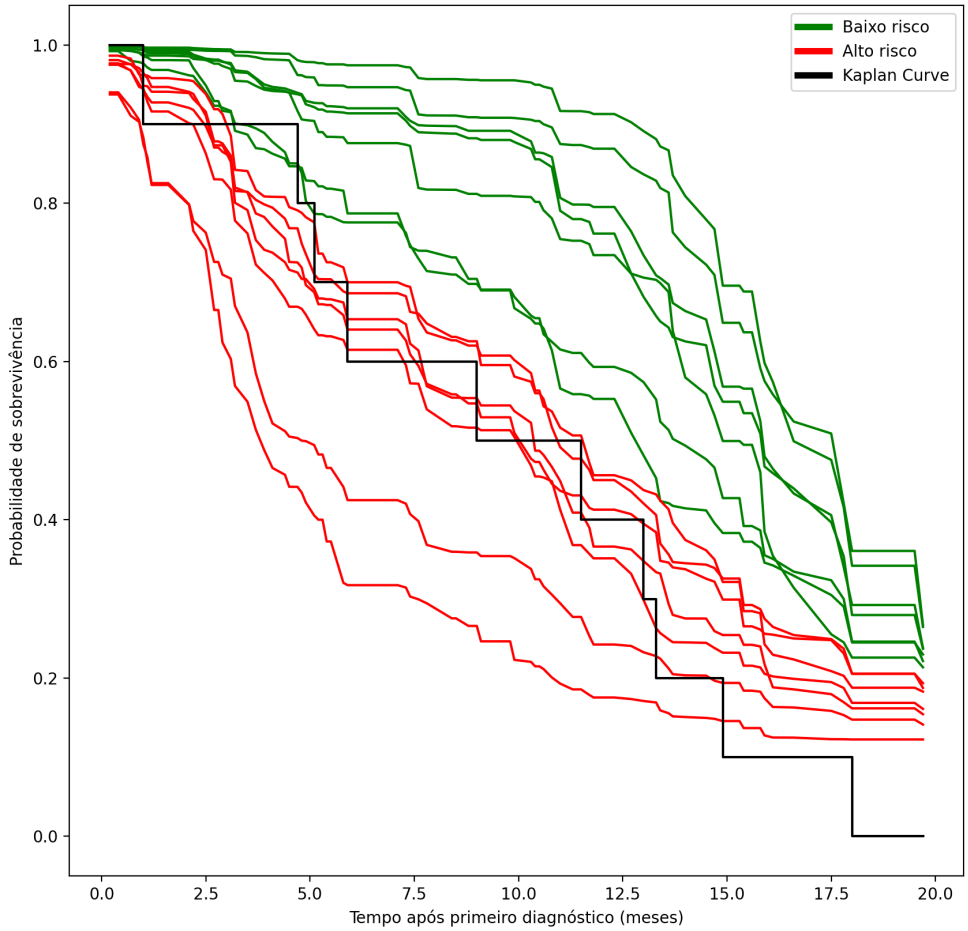
Destaca-se o fato do número reduzido de variáveis que a biblioteca otimizou para o RSF, que parece ser mais vulnerável ao ruído na adição de novas variáveis, para essa aplicação, já que as otimizações chegaram em um número menor de preditores.

Além da visualização dos resultados e parâmetros utilizados, é interessante a visualização da previsão em si. Trata-se da curva de probabilidade de sobrevivência dos pacientes, testado em cada ponto temporal guardado pelo modelo. Os gráficos apresentados a seguir mostram as curvas para cada caso otimizado apresentado acima: dois modelos e duas métricas. Cada curva corresponde à um paciente do grupo de teste, sendo as de cor verde os pacientes considerados de baixo risco e as vermelhas os de alto risco. A medida do risco também é calculado pelo modelo,

resulta em um número único que classifica o paciente. A separação dos grupos foi feita pela mediana dos riscos observados nas amostras de teste. Além disso, a curva Kaplan-Meier, com a porcentagem de pacientes sobreviventes ao longo do tempo, também é colocada no gráfico, a fim de comparação com o dado real dessa base.

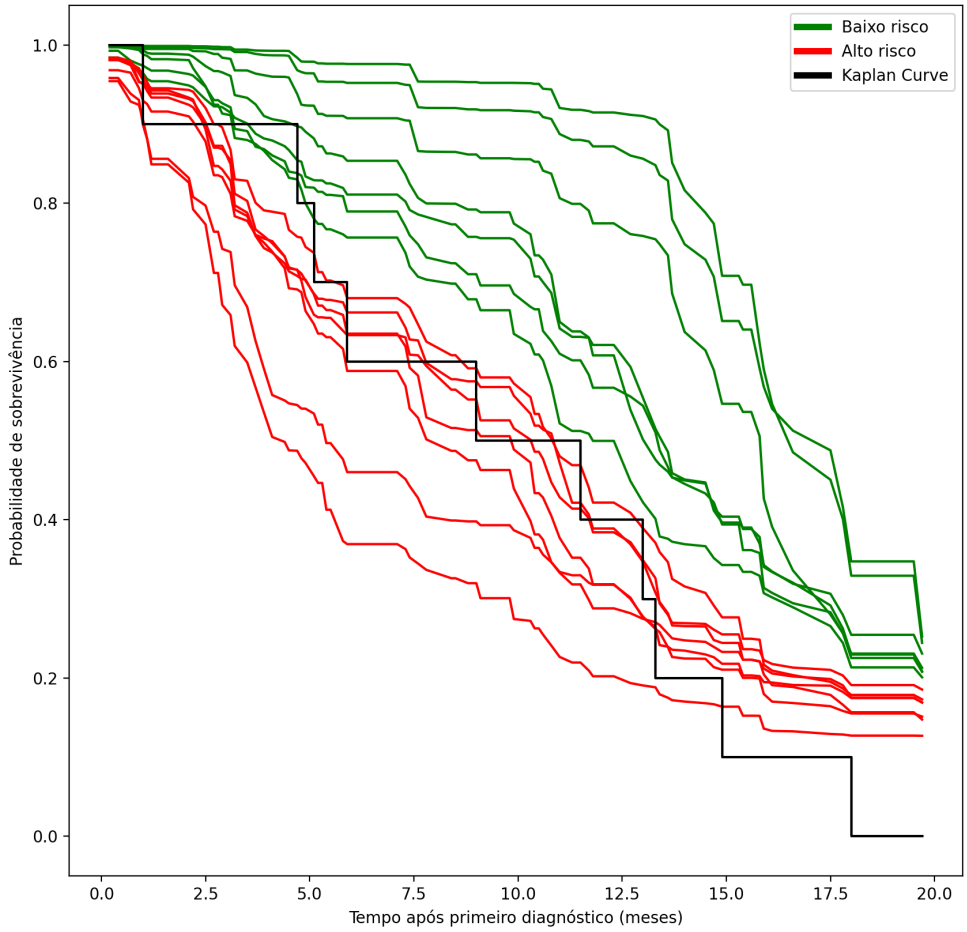
É interessante observar a diferença entre as curvas verdes e vermelhas, e a separação dessas da curva central estimada pelo método Kaplan-Meier. Isso mostra que o modelo consegue, a partir da separação da pontuação de riscos de cada paciente, prever a curva de probabilidade de sobrevivência ajustada para os diferentes riscos. Para pacientes que têm alto risco, é esperado que a curva decresça rapidamente, e para pacientes de baixo risco, é esperado que a curva decresça lentamente.

Figura 2 – Curvas de sobrevivência estimadas pelo RSF para pacientes do conjunto de teste, com parâmetros otimizados para métrica C-Index.



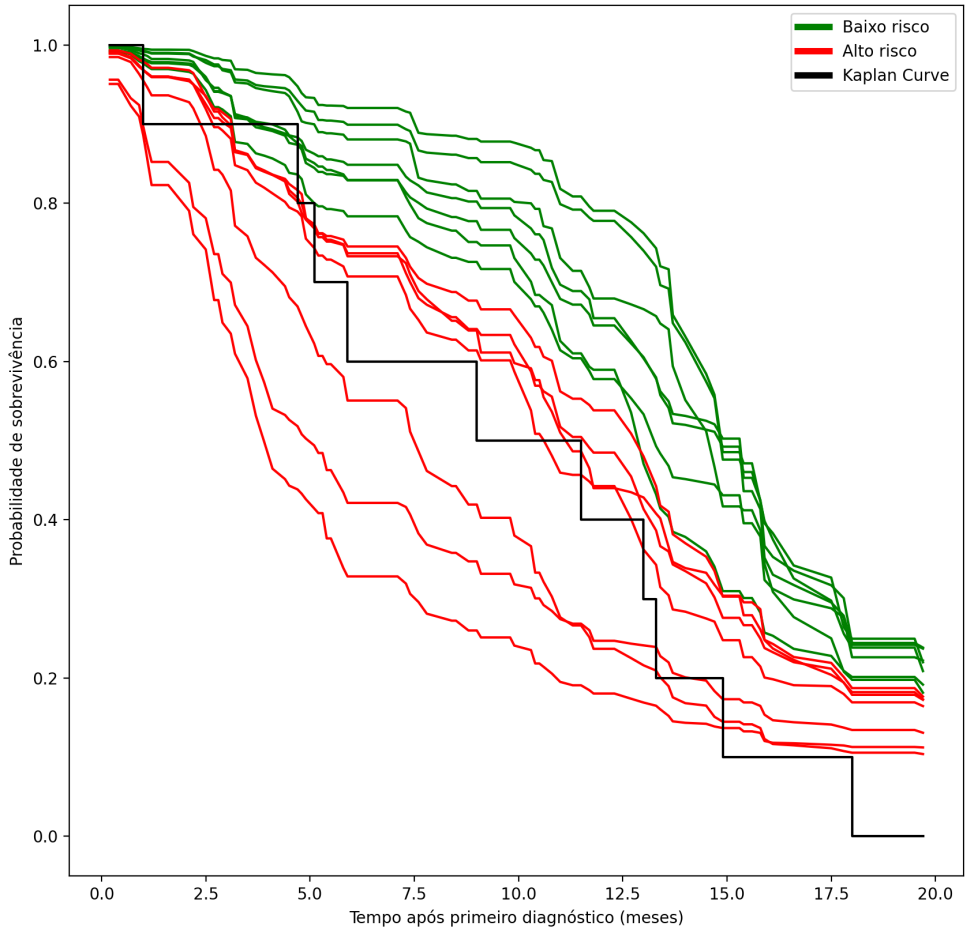
Fonte: Elaborada pelo autor.

Figura 3 – Curvas de sobrevivência estimadas pelo RSF para pacientes do conjunto de teste, com parâmetros otimizados para métrica Brier Score.



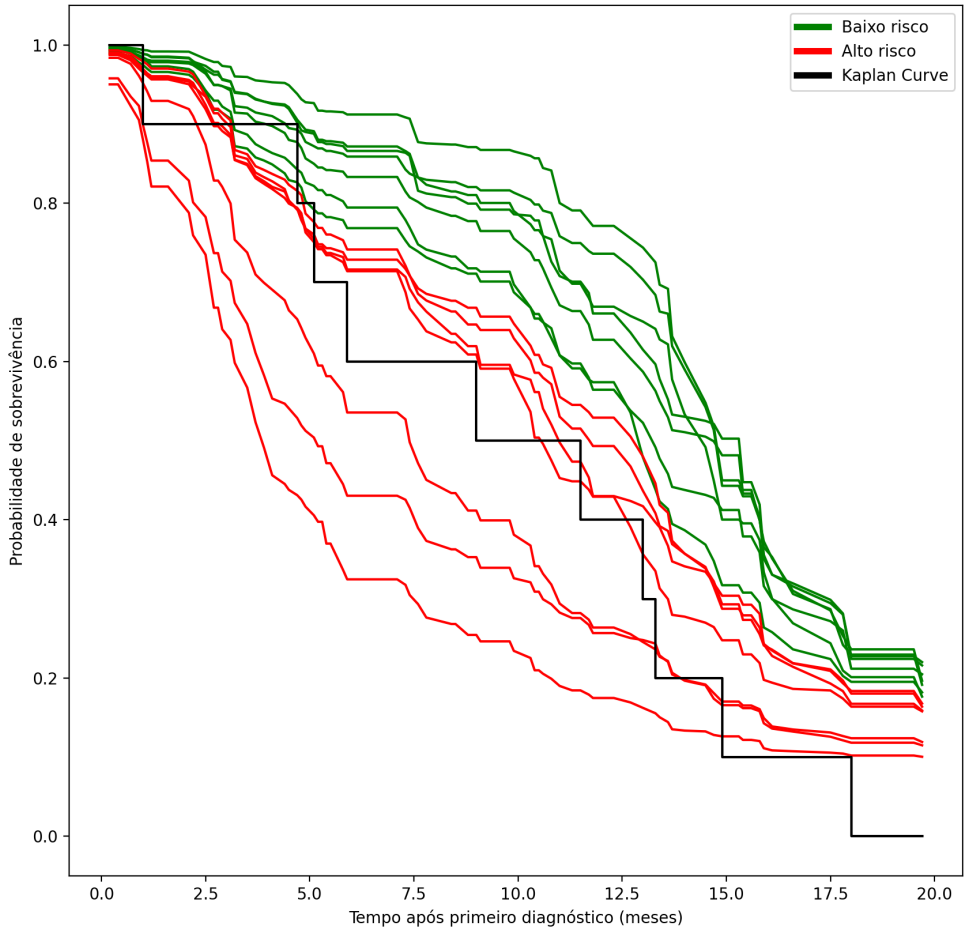
Fonte: Elaborada pelo autor.

Figura 4 – Curvas de sobrevivência estimadas pelo MSR-RF para pacientes do conjunto de teste, com parâmetros otimizados para métrica Brier Score.



Fonte: Elaborada pelo autor.

Figura 5 – Curvas de sobrevivência estimadas pelo MSR-RF para pacientes do conjunto de teste, com parâmetros otimizados para métrica C-index.



Fonte: Elaborada pelo autor.

Analisando os gráficos, é observável que o RSF demonstra uma maior separação das curvas de sobrevivência entre os pacientes, em comparação com o MSR-RF, apesar da diferença não ser de grande natureza. Esse resultado é interessante pois apesar do MSR-RF apresentar uma pontuação consideravelmente maior para *C-index*, ou seja, ele consegue discernir entre dois casos o de maior risco, as curvas de sobrevivência estimadas não se mostram com uma separação tão grande quanto as geradas pelo RSF. Isso significa que um paciente de baixo risco não terá uma probabilidade de sobrevivência muito diferente do que um paciente em alto risco, para o mesmo tempo T de teste.

3.5 Considerações Finais

Os resultados obtidos se mostraram consideravelmente diferentes entre o algoritmo RSF e sua adaptação MSR-RF. Apesar da análise das métricas ser importante na avaliação da performance dos modelos, a visualização das probabilidades ao longo do tempo promove um outro ângulo para entendimento do aprendizado de cada um. O próximo capítulo discorrerá sobre os desafios e aprendizados ao desenvolver esse trabalho, além da importância de estudos para a evolução do entendimento do Glioblastoma.

Capítulo 4

CONCLUSÃO

4.1 Contribuições

O aprendizado ao desenvolver esse trabalho foi muito grande. Desde novas leituras no campo médico, até leituras aprofundadas nas mais recentes pesquisas em aprendizado de máquina. O processo de desenvolvimento da pesquisa é muito enriquecedor, ao trabalhar com a revisão da literatura, proposição da hipótese e análise de um experimento. A disponibilidade de bases de dados de pesquisas tão importantes no ramo da oncologia é muito interessante para a inserção de profissionais da área técnica de ciência de dados e aprendizado de máquina em pesquisas multidisciplinares, pois os métodos trabalhados facilitam drasticamente o exame minucioso de uma grande quantidade de pacientes, feito que teria um custo de tempo muito alto se feito manualmente.

É claro que essas bases apresentam o dado de forma bruta, e precisam de diversos tratamentos para serem utilizadas com modelos de aprendizado de máquina. Caso não haja uma concentração de informação suficiente nas variáveis dependentes, é realmente muito difícil construir um modelo que aprenda. Dessa forma, é fundamental uma metodologia que facilite esse procedimento, principalmente para profissionais que desejam usufruir das vantagens de realizar uma análise automática, mas carecem de conhecimento técnico estatístico e de manuseio de bibliotecas recheadas de modelos para testar.

Primeiramente, é de suma importância o entendimento dos dados. O filtro de tempo máximo adotado nesse trabalho foi fundamental para ditar um bom desempenho dos modelos, pois grande parte das amostras após certo tempo podem ser consideradas *outliers*, e introduzem ruído no aprendizado da função de sobrevivência que se deseja estimar. Decerto, esse tipo de modelo preditivo nunca terá uma taxa de acerto perfeita, mas é interessante poder prever as probabilidades ao longo do tempo da *maioria* das amostras.

Diversas técnicas descritas por inúmeros autores foram unidas na construção da *SurvivalLib*, que funciona como um *wrapper* para efetuar análises de sobrevida. Ao juntar tantas fases dos procedimentos necessários para uma boa solução de aprendizado de máquina, como pré-processamento, *variable importance*, *feature selection*, treinamento e validação dos modelos, com métricas sugeridas na literatura para essa aplicação específica, a biblioteca funciona como um possibilitador para pesquisas futuras na área de análise de sobrevida, como previsão de falhas

em máquina, na área de engenharia, mas principalmente na pesquisa oncológica e médica que foca em predição a partir de informação genética.

O desempenho dos modelo em prever a função de sobrevivência foi, de certa forma, surpreendente, ao passo que a base de dados continha uma pequena quantidade de amostras. Houve uma significativa diferença entre os pacientes, em comparação com a curva de Kaplan-Meier, que funciona como um *baseline* nesse tipo de análise. O MSR-RF apresentou um ótimo resultado para o *C-index*, indicando que conseguiu aprender o que distingue os pacientes quanto ao risco. Isso abre oportunidade para pesquisas futuras que tenham foco na predição do risco do paciente, pois esse algoritmo, ao diminuir o viés na seleção das variáveis para *split*, parece estimar de forma mais precisa os indivíduos que se encontram em estado mais grave quando comparados com os que estão em condições mais amenas.

Por outro lado, o RSF apresentou uma melhor separação dos pacientes ao estimar a função de sobrevida ao longo do tempo. Esse resultado é interessante ao passo que, mesmo com a métrica *C-index* ligeiramente menor em comparação com o MSR-RF, apresenta maior precisão da previsão da probabilidade de sobrevida. Neste caso, pesquisas futuras seriam atraentes para entender o comportamento desse resultado em outras bases de dados, como por exemplo em outros casos de câncer. Esse estudo pode ser facilmente conduzido com uso da *SurvivalLib*.

Outro ponto que podem ser levado como tema para pesquisas futuras é a forma como a seleção das variáveis é feita. Apesar de bons resultados empíricos da técnica utilizada, descrita por Chen e Ishwaran (2012), uma outra opção promissora é descrita por Ishwaran *et al.* (2011), que também foi o idealizador do RSF. A proposta envolve a utilização de características intrínsecas às árvores de decisão para fazer a seleção de variáveis, como a frequência de utilização em *splits*. Além disso, outra alternativa para a etapa do *feature selection* está na identificação de profundidade mínima de sub-árvores, que é empregada em casos de alta dimensionalidade (ISHWARAN *et al.*, 2010).

Considerando as métricas de avaliação de desempenho, é também sugerida a análise da utilização da adaptação do conhecido R^2, com extensões de sensitividade e especificidade, propostas por Heagerty e Zheng (2005). O uso de dependência no tempo e interação com risco pode ter efeitos interessantes nas bases estudadas.

4.2 Relação entre o Projeto e o Curso de Engenharia de Computação

O curso de graduação em Engenharia de Computação intensificou meu interesse em tecnologia no geral, mas com uma visão muito mais técnica, ao abranger tanto o aspecto da elétrica e eletrônica, quanto do desenvolvimento de software. No início, com as disciplinas de *Introdução à Ciência de Computação*, minha capacidade de projeto de software foi instigada e

promovida, ao serem propostos trabalhos que foram desenvolvidos ao longo do semestre. Esse tipo de projeto é muito interessante ao passo que promove uma linha de crescimento do software por vários meses, desde a concepção e até o relatório final.

Acredito que grande parte do foco do curso é em microeletrônica, o que é muito interessante por fazer parte do seleto grupo de universidades brasileiras que oferecem esse tipo de conhecimento. A construção do aprendizado é gradual e lógica, o que julgo importante para o sucesso do aluno. Porém, ao longo do curso me vi em diferentes áreas e, por fim, me encontrei na ciência de dados, área que não recebe muito foco na Engenharia de Computação. Apesar disso, a possibilidade de construir a grade horária com certa liberdade me permitiu cursar disciplinas da área, sendo a principal *Introdução à Ciência de Dados* foi de suma importância na minha formação.

Este trabalho permitiu a aplicação dos conhecimentos construídos ao longo de anos, bem como adquirir novas habilidades, aprendizados e superação desafios. A experiência da elaboração de uma pesquisa científica é muito importante para o Engenheiro, e claro muito enriquecedora para mim.

REFERÊNCIAS

ALI, J. B.; CHEBEL-MORELLO, B.; SAIDI, L.; MALINOWSKI, S.; FNAIECH, F. Accurate bearing remaining useful life prediction based on weibull distribution and artificial neural network. **Mechanical Systems and Signal Processing**, Elsevier BV, v. 56-57, p. 150–172, may 2015. Citado na página 20.

BLEEKER, F. E.; MOLENAAR, R. J.; LEENSTRA, S. Recent advances in the molecular understanding of glioblastoma. **Journal of Neuro-Oncology**, Springer Science and Business Media LLC, v. 108, n. 1, p. 11–27, jan 2012. Citado 3 vezes nas páginas 19, 31 e 32.

BRAY, F.; FERLAY, J.; SOERJOMATARAM, I.; SIEGEL, R. L.; TORRE, L. A.; JEMAL, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, Wiley, v. 68, n. 6, p. 394–424, sep 2018. Citado na página 19.

BREIMAN, L. **Classification and regression trees**. New York: Chapman & Hall, 1993. ISBN 9780412048418. Citado na página 26.

____. Bagging predictors. **Machine Learning**, Springer Science and Business Media LLC, v. 24, n. 2, p. 123–140, 1996. Citado na página 27.

____. Random forests. **Machine Learning**, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 27 e 35.

BRENNAN, C. W. The somatic genomic landscape of glioblastoma. **Cell**, Elsevier BV, v. 155, n. 2, p. 462–477, oct 2013. Citado na página 32.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E.; ANTIPIN, Y.; REVA, B.; GOLDBERG, A. P.; SANDER, C.; SCHULTZ, N. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data: Figure 1. **Cancer Discovery**, American Association for Cancer Research (AACR), v. 2, n. 5, p. 401–404, may 2012. Citado 2 vezes nas páginas 20 e 31.

CHEADLE, C.; VAWTER, M. P.; FREED, W. J.; BECKER, K. G. Analysis of microarray data using z score transformation. **The Journal of Molecular Diagnostics**, Elsevier BV, v. 5, n. 2, p. 73–81, may 2003. Citado na página 31.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, Elsevier BV, v. 99, n. 6, p. 323–329, jun 2012. Citado 3 vezes nas páginas 35, 36 e 48.

DATEMA, F. R.; MOYA, A.; KRAUSE, P.; BäCK, T.; WILLMES, L.; LANGEVELD, T.; JONG, R. J. B. de; BLOM, H. M. Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. **Head & Neck**, Wiley, v. 34, n. 1, p. 50–58, feb 2011. Citado na página 20.

DELEN, D.; WALKER, G.; KADAM, A. Predicting breast cancer survivability: a comparison of three data mining methods. **Artificial Intelligence in Medicine**, Elsevier BV, v. 34, n. 2, p. 113–127, jun 2005. Citado 4 vezes nas páginas 20, 24, 25 e 29.

DONOHO, D. L. High-dimensional data analysis: The curses and blessings of dimensionality. 2000. Citado 4 vezes nas páginas 25, 26, 32 e 35.

EFRON, B. **An introduction to the bootstrap**. New York: Chapman & Hall, 1994. ISBN 0412042312. Citado na página 27.

FOTSO, S. *et al.* **PySurvival: Open source package for Survival Analysis modeling**. 2019. Disponível em: <https://www.pysurvival.io/>. Citado na página 34.

GAO, J.; AKSOY, B. A.; DOGRUSOZ, U.; DRESDNER, G.; GROSS, B.; SUMER, S. O.; SUN, Y.; JACOBSEN, A.; SINHA, R.; LARSSON, E.; CERAMI, E.; SANDER, C.; SCHULTZ, N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. **Science Signaling**, American Association for the Advancement of Science (AAAS), v. 6, n. 269, p. pl1–pl1, mar 2013. Citado na página 20.

GOLDMAN, D.; DOMSCHKE, K. Making sense of deep sequencing. **The International Journal of Neuropsychopharmacology**, Oxford University Press (OUP), v. 17, n. 10, p. 1717–1725, jun 2014. Citado na página 26.

GRAF, E.; SCHMOOR, C.; SAUERBREI, W.; SCHUMACHER, M. Assessment and comparison of prognostic classification schemes for survival data. **Statistics in Medicine**, Wiley, v. 18, n. 17-18, p. 2529–2545, sep 1999. Citado na página 37.

HEAGERTY, P. J.; ZHENG, Y. Survival model predictive accuracy and ROC curves. **Biometrics**, Wiley, v. 61, n. 1, p. 92–105, mar 2005. Citado 2 vezes nas páginas 36 e 48.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. 2008. Citado 3 vezes nas páginas 27, 28 e 34.

ISHWARAN, H.; KOGALUR, U. B.; CHEN, X.; MINN, A. J. Random survival forests for high-dimensional data. **Statistical Analysis and Data Mining**, Wiley, v. 4, n. 1, p. 115–132, jan 2011. Citado na página 48.

ISHWARAN, H.; KOGALUR, U. B.; GORODESKI, E. Z.; MINN, A. J.; LAUER, M. S. High-dimensional variable selection for survival data. **Journal of the American Statistical Association**, Informa UK Limited, v. 105, n. 489, p. 205–217, mar 2010. Citado na página 48.

KLEIN, J. **Survival analysis : techniques for censored and truncated data**. New York: Springer, 2003. ISBN 9780387216454. Citado 2 vezes nas páginas 20 e 23.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. **Computational and Structural Biotechnology Journal**, Elsevier BV, v. 13, p. 8–17, 2015. Citado na página 20.

LEE, E. **Statistical methods for survival data analysis**. Hoboken, N.J: Wiley, 2003. ISBN 9780471458555. Citado na página 24.

LOPEZ, Y. O. N.; VICTORIA, B.; GOLUSINSKI, P.; GOLUSINSKI, W.; MASTERNAK, M. M. Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes. **Reports of Practical Oncology & Radiotherapy**, Elsevier BV, v. 23, n. 1, p. 6–20, jan 2018. Citado 2 vezes nas páginas 19 e 29.

MAGLOTT, D.; OSTELL, J.; PRUITT, K. D.; TATUSOVA, T. Entrez gene: gene-centered information at NCBI. **Nucleic Acids Research**, Oxford University Press (OUP), v. 39, n. Database, p. D52–D57, nov 2010. Citado na página 33.

MIRZA, B.; WANG, W.; WANG, J.; CHOI, H.; CHUNG, N. C.; PING, P. Machine learning and integrative analysis of biomedical big data. **Genes**, MDPI AG, v. 10, n. 2, p. 87, jan 2019. Citado na página 26.

OMURLU, I. K.; TURE, M.; TOKATLI, F. The comparisons of random survival forests and cox regression analysis with simulation and an application related to breast cancer. **Expert Systems with Applications**, Elsevier BV, v. 36, n. 4, p. 8582–8588, may 2009. Citado 3 vezes nas páginas 20, 29 e 40.

RUSSELL, S. **Artificial intelligence : a modern approach**. Upper Saddle River, New Jersey: Prentice Hall, 2010. ISBN 9780136042594. Citado 2 vezes nas páginas 25 e 39.

SANDRI, M.; ZUCCOLOTTO, P. A bias correction algorithm for the gini variable importance measure in classification trees. **Journal of Computational and Graphical Statistics**, Informa UK Limited, v. 17, n. 3, p. 611–628, sep 2008. Citado na página 35.

STUPP, R.; MASON, W. P.; BENT, M. J. van den; WELLER, M.; FISHER, B.; TAPHOORN, M. J.; BELANGER, K.; BRANDES, A. A.; MAROSI, C.; BOGDAHN, U.; CURSCHMANN, J.; JANZER, R. C.; LUDWIN, S. K.; GORLIA, T.; ALLGEIER, A.; LACOMBE, D.; CAIRNCROSS, J. G.; EISENHAUER, E.; MIRIMANOFF, R. O. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. **New England Journal of Medicine**, Massachusetts Medical Society, v. 352, n. 10, p. 987–996, mar 2005. Citado 2 vezes nas páginas 19 e 40.

UNO, H.; CAI, T.; PENCINA, M. J.; D'AGOSTINO, R. B.; WEI, L. J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. **Statistics in Medicine**, Wiley, p. n/a–n/a, 2011. Citado na página 37.

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 51, n. 6, p. 1–36, feb 2019. Citado 4 vezes nas páginas 20, 23, 25 e 34.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, Springer Science and Business Media LLC, v. 10, n. 1, p. 57–63, jan 2009. Citado na página 26.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, Elsevier BV, v. 2, n. 1-3, p. 37–52, aug 1987. Citado na página 36.

WRIGHT, M. N.; DANKOWSKI, T.; ZIEGLER, A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. 2016. Citado na página 29.