



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
CURSO DE BIOTECNOLOGIA

MARIA JÚLIA PEREIRA DAVI

**UTILIZANDO MÉTODO DE APRENDIZADO DE MÁQUINA PARA ANÁLISES DE
SOBREVIDA DE PACIENTES ONCOLÓGICOS**

MOSSORÓ

2019

MARIA JÚLIA PEREIRA DAVI

**UTILIZANDO MÉTODO DE APRENDIZADO DE MÁQUINA PARA ANÁLISES DE
SOBREVIDA DE PACIENTES ONCOLÓGICOS**

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Biotecnologia.

Orientador: Prof. Msc. Taffarel Melo Torres

MOSSORÓ

2019

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

P249u Pereira Davi, Maria Júlia.
Utilizando método de aprendizado de máquina
para análises de sobrevida de pacientes
oncológicos / Maria Júlia Pereira Davi. - 2020.
35 f. : il.

Orientador: Taffarel Melo Torres.
Monografia (graduação) - Universidade Federal
Rural do Semi-árido, Curso de Biotecnologia,
2020.

1. Bioestatística. 2. Bioinformática. 3. Câncer.
4. Random Forest. I. Melo Torres, Taffarel,
orient. II. Título.

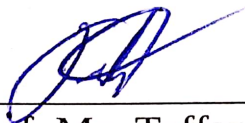
MARIA JÚLIA PEREIRA DAVI

**UTILIZANDO MÉTODO DE APRENDIZADO DE MÁQUINA PARA ANÁLISES DE
SOBREVIDA DE PACIENTES ONCOLÓGICOS**

Monografia apresentada a Universidade Federal
Rural do Semi-Árido como requisito para
obtenção do título de Bacharel em Biotecnologia.

Defendida em: 07 / 02 / 2 020.

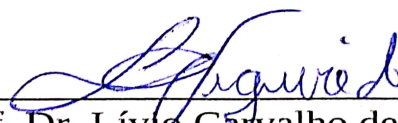
BANCA EXAMINADORA



Prof. Me. Taffarel Melo Torres
Orientador



Profa. Dra. Ana Carla Diógenes Suassuna Bezerra
Examinadora



Prof. Dr. Lívio Carvalho de Figueirêdo
Examinador

AGRADECIMENTOS

Agradeço à minha família, pelo apoio e confiança.

Ao meu orientador, que de fato contribuiu não só para este trabalho, como também para a minha formação desde o início do curso, sempre disposto a ensinar da melhor forma possível e se preocupando não só comigo, mas com todos os seus orientandos.

Aos amigos que fiz durante o curso, que tornaram todo esse processo muito mais divertido e leve, e que me farão lembrar desses quatro anos com muito mais amor (talvez até saudade).

Às minhas amigas de sempre: Fernanda, Luana e Marília, por simplesmente se manterem sempre presentes e sempre sendo vocês.

Aos membros da banca, por contribuírem com seus conhecimentos para a melhoria deste trabalho.

*Only want to say that I gave it all I had
That I felt afraid and I didn't step back
Whether right or wrong
I did everything with love*

Love too much - Keane

RESUMO

O câncer é uma doença multifatorial e complexa, caracterizada pelo crescimento desordenado de células dos mais diversos tecidos, provocado por alteração genética. Os fatores que contribuirão para o quadro de desenvolvimento da doença variam em cada indivíduo e em cada tipo de câncer, sendo a análise de sobrevida uma das formas de se avaliar a probabilidade de sobrevivência de acordo com a característica de determinado fator. Dessa forma, ao se utilizar de técnicas de aprendizado de máquina, é possível prever a probabilidade baseando-se em todos os fatores associados a cada paciente, de maneira que se identifique o padrão de risco de todo um conjunto de pacientes de forma otimizada. Tendo em vista o cenário local, o estudo objetivou analisar o padrão de sobrevida dos pacientes oncológicos atendidos pelo Hospital da Liga Mossoroense de Estudos e Combate ao Câncer durante o período de 2006 a 2016, utilizando método de aprendizado de máquina e mineração de dados para realização de análise de sobrevida. No presente estudo, foi utilizada a biblioteca randomForestSRC, desenvolvida na linguagem R. O algoritmo foi utilizado para identificação das variáveis mais relevantes para a predição da sobrevida, além de ter sido avaliada a taxa de erro e performance do método. Para os 11 grupos analisados, 7 apresentaram taxa de erro abaixo de 20% e 4 apresentaram Brier Score abaixo de 0.1.

Palavras-chave: Câncer. Bioestatística. Mineração de dados.

ABSTRACT

The cancer is a complex and multifactorial disease, described as disordered growth of cells from different types of tissues, caused by genetic alteration. Factors that will contribute to the development of the disease vary in each individual and in each type of cancer, and survival analysis is one of the ways to assess the probability of survival according to the characteristic of a given factor. Thus, by using machine learning techniques, it is possible to predict the probability based on all factors associated to each patient, so that the risk pattern of an entire patient set can be identified optimally. Considering the local scenario, this study aimed to analyze the risk pattern of the oncological patients attended by the Hospital Mossoroense de Estudos e Combate ao Câncer during the period of 2006 to 2016, using machine learning e data mining methods for survival analysis. In present study, was used randomForestSRC library, developed in R language. The algorithm was used to identify the most relevant variables for survival prediction, and also the error rate and performance of the method were evaluated. Considering the 11 groups analyzed, 7 presented error rate below 20% and 4 presented Brier Score below 0.1.

Keywords: Cancer. Biostatistics. Data mining.

LISTA DE FIGURAS

Figura 1 – Taxa de mortalidade entre 2008 e 2017.....	12
Figura 2 – Gráfico de sobrevida para o câncer de pele.....	19
Figura 3 – Classificação das variáveis para o câncer de pele.....	22
Figura 4 – Classificação das variáveis para o câncer de mama.....	23
Figura 5 – Gráfico de sobrevida para o câncer de mama.....	24
Figura 6 – Classificação das variáveis para o câncer de próstata.....	25
Figura 7 – Gráfico de sobrevida para o câncer de próstata.....	26
Figura 8 – Gráficos de sobrevida.....	27

SUMÁRIO

1 INTRODUÇÃO.....	10
2 FUNDAMENTAÇÃO TEÓRICA.....	11
2.1 CÂNCER: EPIDEMIOLOGIA.....	11
2.1.1 Câncer: análises de sobrevida.....	12
2.2 DATA MINING, MODELOS DE PREDIÇÃO E APRENDIZADO DE MÁQUINA.....	13
2.2.1 Algoritmo <i>Random Forest</i>	14
3 OBJETIVOS.....	16
3.1 OBJETIVO GERAL.....	16
3.2 OBJETIVOS ESPECÍFICOS.....	16
4 MATERIAIS E MÉTODOS.....	17
4.1 FORMATAÇÃO E PADRONIZAÇÃO DOS DADOS.....	17
4.2 ANÁLISE DE SOBREVIDA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA.....	17
5 RESULTADOS E DISCUSSÃO.....	19
5.1 TAXA DE ERRO DO MODELO DE PREDIÇÃO.....	19
5.2 ANÁLISES DE SOBREVIDA E CLASSIFICAÇÃO DAS VARIÁVEIS.....	20
6 CONSIDERAÇÕES FINAIS.....	29
REFERÊNCIAS.....	30

1 INTRODUÇÃO

A sociedade contemporânea produz dados rotineiramente na grande maioria de suas atividades, sejam científicas ou em rotinas de trabalho, principalmente na área da saúde, ocasionando uma grande quantidade de dados digitalizados que podem ser convertidos em informação. Esses dados são gerados todos os dias, apresentando grande potencial para descobertas científicas e desenvolvimento de novas tecnologias. O termo *big data* foi utilizado pelos cientistas, para caracterizar essa grande quantidade de dados disponível eletronicamente (MURDOCH *et al*, 2013). O ser humano não é capaz de interpretar e classificar essa quantidade de dados utilizando o gerenciamento convencional de banco de dados (AMORIM, 2006). Para que esta interpretação seja efetiva são necessárias ferramentas de análise sofisticadas, como a tecnologia de mineração de dados. Esta tecnologia surgiu da união entre as áreas de conhecimento da estatística clássica, inteligência artificial e aprendizado de máquina (SFERRA *et al*, 2003).

Uma alternativa que apresenta grande potencial para a medicina no controle do câncer e diversas doenças é a aplicação de técnicas de mineração de dados na área da saúde. A maioria dos registros hospitalares digitais contêm dados qualitativos e quantitativos que podem ser uma fonte de informação útil (MURDOCH *et al*, 2013). No período de 2003 a 2015, segundo Jothi *et al* (2015), cerca de 3840 artigos relacionados ao tema de *data mining* (DM) e assistência médica foram encontrados nos bancos de dados de publicações.

O câncer é uma doença multifatorial, podendo tomar diferentes formas dependendo das células de origem, apresentando diferenças genéticas para indivíduo. As variações genéticas e diversos fatores clínicos influenciam no sucesso do tratamento (National Cancer Institute, 2019). Informações importantes podem ser extraídas a partir da análise integrada dos dados de pacientes oncológicos uma vez que perfis genômicos e dados clínicos são disponibilizados em bancos de dados públicos. No Brasil, estima-se 600 mil novos casos de câncer para o biênio 2018-2019 (INCA, 2019). Uma fonte de informações clínicas organizadas de forma sistemática e disponibilizada abertamente são os Registros Hospitalares de Câncer (RHC), que representam um potencial alvo para mineração de dados e descoberta de padrões para a doença no Brasil.

2 FUNDAMENTAÇÃO TEÓRICA

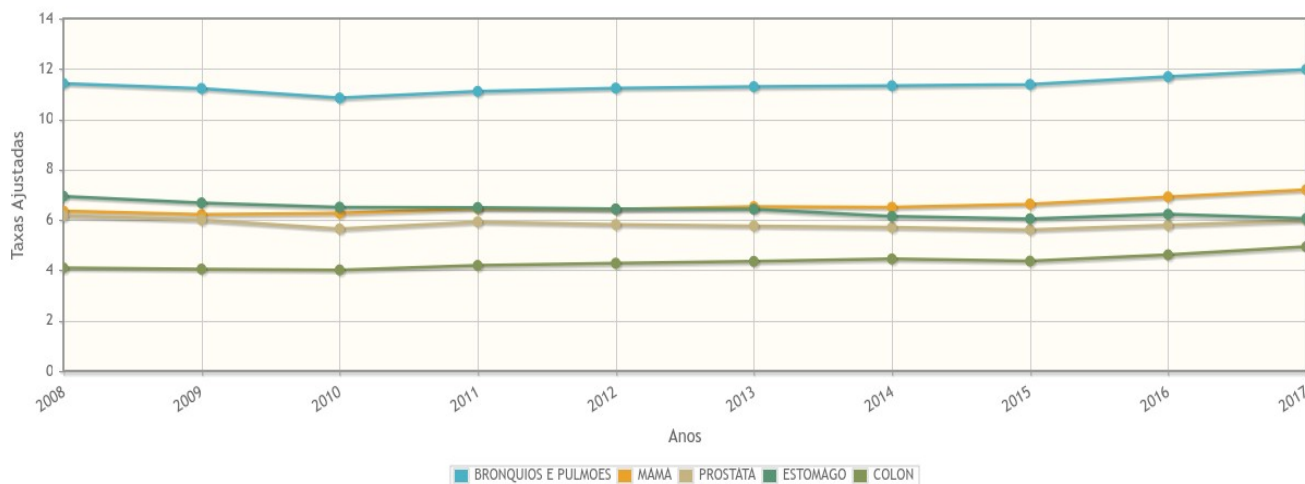
2.1 CÂNCER: EPIDEMIOLOGIA

Por definição, o câncer é uma doença que corresponde a um conjunto de mais de 100 tipos de enfermidades, que possuem em comum o crescimento desordenado de células, que invadem tecidos e órgãos, formando tumores de diferentes tipos celulares (INCA, 2019). Mundialmente, é considerado como uma ameaça para a saúde pública, especialmente em países de médio e baixo Índice de Desenvolvimento Humano, uma vez que para estes as estimativas apontam aumento de 81% e 100% na taxa de incidência entre os anos de 2008 e 2030 (FIDLER *et al*, 2017). No Brasil, o número de novos casos para o ano de 2018 foi de 300.140 e 282.450 para homens e mulheres, respectivamente. A neoplasia mais frequente para o sexo masculino correspondeu a de próstata, e para as mulheres a da mama (INCA, 2017).

Em países mais desenvolvidos, como os Estados Unidos da América, a taxa de mortalidade do câncer vem decrescendo desde 1991 até 2016, e as incidências para homens e mulheres na última década apresentaram um declínio anual de 1.8% e 1.4%, respectivamente. Um dos principais fatores associados as taxas de incidência e mortalidade é o desenvolvimento socioeconômico do país, onde se observa uma taxa de mortalidade maior em países mais pobres, até mesmo para os casos que podem ser prevenidos (SIEGEL *et al*, 2019).

Apesar dos avanços nos tratamentos disponíveis, segundo os dados mais recentes consolidados pelo INCA, a taxa de mortalidade para o ano de 2017 foi de 115.057 para os homens e 103.583 para as mulheres (INCA, 2017). Segundo Atlas de Mortalidade por Câncer (MINISTÉRIO DA SAÚDE, 2014), ao longo de 10 anos, as taxas de mortalidade se mantiveram relativamente constantes para ambos os sexos, como é possível visualizar na Figura 1, onde são apresentados os 5 tipos de câncer mais frequentes no país.

Figura 1- Mortalidade entre 2008 e 2017: Taxas de mortalidade para as 5 localizações primárias mais frequentes, no período de 2008-2017, ajustadas por idade, pela população mundial, por 100.000, para ambos os sexos.



Fontes: MS/SVS/DASIS/CGIAE/Sistema de Informação sobre Mortalidade - SIM; MP/Fundação Instituto Brasileiro de Geografia e Estatística - IBGE; MS/INCA/Conprev/Divisão de Vigilância.

O controle epidemiológico do câncer é importante para que os padrões específicos de cada região sejam apontados e estratégias sejam desenvolvidas para solucionar os problemas de cada localidade ou para que as estimativas indiquem quais ações preventivas devem ser tomadas (FERLAY *et al*, 2018).

2.1.1 Câncer: análises de sobrevida

As análises de sobrevida consistem na observação de eventos em função de um tempo determinado, sendo este evento normalmente o óbito. Para que sejam realizadas, é utilizado o método estatístico de Kaplan-Meier, desenvolvido inicialmente para medir a frequência ou número de pacientes que sobreviviam a um determinado tratamento médico (ETIKAN *et al*, 2017). Dessa forma, é calculada a probabilidade de sobrevivência ao longo do tempo, onde se comparam os efeitos de diferentes fatores na sobrevivência dos pacientes.

Na oncologia, a prática frequente é avaliar o risco por meio de análises de sobrevida global ou recorrência/progressão da doença. A predição do risco de cada paciente de forma individualizada é um dos elementos mais importantes na medicina personalizada (HU, 2018).

Não só para o câncer, como também para outras doenças, o processo de análise por abordagens epidemiológicas convencionais requer bastante tempo do pesquisador para selecionar variáveis mais importantes, limpar valores ausentes, transformar variáveis, entre outros processos necessários (STEELE *et al*, 2018).

As variáveis relacionadas ao TNM são amplamente utilizadas em estudos de sobrevida, sendo consideradas determinantes no prognóstico dos pacientes. O sistema TNM se baseia na avaliação de características anatômicas da doença, considerando o tumor primário (T), os linfonodos (N) e a presença ou ausência de metástases a distância (M) (INCA, 2019). No entanto, a utilização desse sistema isoladamente não constitui um método de prognóstico altamente preciso, sendo inconclusivo para alguns tipos de câncer (MAHAR *et al*, 2018; NOMA *et al*, 2018).

São avaliadas também outras variáveis relacionadas ao câncer em estudos de sobrevida, como dados genéticos, tipo de tratamento aplicado, estadiamento, sexo, ocupação, variando de acordo com o conjunto de dados disponível para análise e objetivo específico do estudo (LÁNCISKY *et al*, 2016; ADAMU *et al*, 2019). Nos estudos citados, é realizada a análise para cada uma dessas variáveis isoladamente, comparando-se os os grupos formados dentro de um mesmo conjunto de pacientes.

2.2 DATA MINING, MODELOS DE PREDIÇÃO E APRENDIZADO DE MÁQUINA

Atualmente, um considerável volume de informações é coletado e armazenado por grandes instituições de saúde. Esses dados são provenientes de prontuários e registros hospitalares, sendo o Registro Hospitalar do Câncer uma base de integração dos dados de hospitais de todo o Brasil. A quantidade de informações disponibilizada por esse sistema requer uma análise mais robusta e adequada para que conhecimento útil e de qualidade seja gerado a partir desses dados (AMORIM, 2006; KAUR *et al*, 2018). A técnica de mineração de dados proporciona a obtenção de conhecimento que estava até então não visível na base de dados, e pode ser utilizada para prever tendências ou descrever características do passado (SFERRA *et al*, 2003; SELLAMY *et al*, 2018). Os métodos de DM podem incluir técnicas de classificação,

generalização, caracterização, agrupamento, associação, evolução, descoberta de padrões, visualização de dados e mineração guiada por regra (JOTHI *et al*, 2015).

O objetivo da aplicação de DM é o processo de descoberta de conhecimento em bancos de dados ou *Knowledge Discovery in Databases*¹ (KDD), cuja definição é dada por Fayyad *et al* (1996) como sendo “o processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e, no final das contas, compreensíveis em dados”. Atualmente, a aplicação de técnicas de *data mining* possibilitam a extração de informação útil a partir de dados biológicos complexos (LAN *et al*, 2018) e se relacionam diretamente com a área da Bioinformática, cujo objetivo principal é integrar e estudar qualquer dado biológico de forma estratégica, através da tecnologia da informação e ciência da computação (HASHEMI *et al*, 2017).

Em relação ao desenvolvimento de novas tecnologias provenientes desta área em ascensão, é possível observar em buscas recentes a priorização da otimização de processos como diagnóstico por imagem (BÜHEMANN *et al*, 2014; DO NASCIMENTO *et al*, 2018), classificação de pacientes em grupos de risco, predição de risco em transfusão (GURM *et al*, 2014), identificação de novos fatores de risco de doenças como câncer e diabetes (TOMCZAK *et al*, 2015; ESMAILY *et al*, 2018). Além dessas aplicações, os algoritmos também são bastante aplicados para descoberta de novas combinações potenciais de fármacos para tratamento de doenças como o câncer e mal de Parkinson (GAYVERT *et al*, 2017; XU *et al*, 2018).

2.2.1 Algoritmo *Random Forest*

O algoritmo *Random Forest* (RF) é um algoritmo criado por Leo Breiman (2001) que tem como objetivos principais solucionar dois tipos de problemas: construir uma regra de predição em um problema de aprendizado supervisionado e acessar e classificar variáveis de acordo com sua habilidade de prever determinada resposta (BOULESTEIX *et al*, 2012). A predição final é uma combinação das predições geradas em cada árvore de decisão obtida durante o processo, sendo cada árvore gerada ligeiramente diferente. Cada nó da árvore de decisão é resultado de uma série de perguntas simples com respostas “sim” ou “não” (TAYLOR, 2011).

Os modelos de predição baseado em *Random Forest*² já são utilizados em ferramentas que buscam prever risco de procedimentos como transfusão de sangue, onde o cálculo do risco é feito através de interface web gratuita (GURM *et al*, 2014). Seu papel também é presente em análises de imuno-histoquímica, onde foi possível classificar parâmetros obtidos na análise de padrões de modificação de histonas para prever recorrência de tumores (SELIGSON *et al*, 2005).

Em estudos recentes de Hamidi e colaboradores (2016), foi utilizado o *Random Survival Forest*³ (RSF), método baseado no RF, para identificar fatores de risco no processo de transplante renal, mostrando um melhor resultado para o RSF do que para os modelos convencionais. O uso do método RF comparado ao RSF é citado na literatura como sendo utilizado para prever sobrevida global, metástase a distância e recorrência local do câncer retal (HUANG *et al*, 2017).

Além de servir como base para implementação de outros modelos, a partir do método RF também foi implementado um modelo unificado para análise de sobrevivência, regressão e classificação (RF-SRC: *Random forests for Survival, Regression and Classification*), disponibilizado na linguagem R e criado pelos autores Udaya Kogalur e Hemant Ishwaran (2016). Para análises de sobrevivência, os nós terminais das árvores criadas dão respostas baseadas na estimativa de Kaplan-Meier (KM), a qual por sua vez prevê uma probabilidade de sobrevivência ao longo de um período de tempo determinado (ISHWARAN *et al*, 2016).

Dados de sobrevida são normalmente analisados a partir de métodos que utilizam hipóteses restritas que assumem riscos proporcionais a determinada suposição. O problema de tais métodos é a exigência de dados paramétricos, além de apresentar dificuldade para se analisar interação de múltiplas variáveis (ISHWARAN *et al*, 2007). O RF é um método estatístico totalmente não-paramétrico, o que dispensa a exigência de distribuição normal para análise dos dados, mas que ainda assim oferece eficiência na análise de diferentes conjuntos de dados (EHRLINGER *et al*, 2015). A utilização deste modelo pode otimizar o processo de análise da influência de cada variável na sobrevivência, tanto de forma isolada como associada com outras variáveis.

2 Tradução livre: floresta aleatória.

3 Tradução livre: floresta de sobrevivência aleatória.

3 OBJETIVOS

3.1 OBJETIVO GERAL

Utilizar método de aprendizado de máquina em análises de sobrevida a partir dos dados clínicos do Registro Hospitalar de Câncer (RHC) de pacientes atendidos pela Liga Mossoroense de Estudos e Combate ao Câncer (LMECC).

3.2 OBJETIVOS ESPECÍFICOS

- a) Implementar o uso de algoritmo de mineração de dados para análise dos dados do Registro Hospitalar de Câncer;
- b) Realizar análises de sobrevida para os tipos de câncer mais frequentes na região atendida;

4 MATERIAIS E MÉTODOS

4.1 FORMATAÇÃO E PADRONIZAÇÃO DOS DADOS

Todos os dados submetidos para análise foram organizados em planilhas, com variáveis padronizadas e sem remoção de campos com informações ausentes. Os tipos de câncer foram agrupados de acordo com a Classificação Internacional de Doenças (CID-10), e o período abrangido pela análise foi de 2006 a 2016. Foram utilizados somente os dados de pacientes atendidos pelo Hospital da Liga Mossoroense de Estudos e Combate ao Câncer.

Foram incluídos os seguintes campos do Registro Hospitalar de Câncer (RHC): idade, sexo, raça/cor, ocupação, tratamento, estadiamento do tumor (TNM), estadiamento patológico do tumor (pTNM), tipo histológico, histórico de consumo de álcool e tabaco e histórico de câncer na família e localização primária do tumor. Os pacientes foram agrupados por tipo de câncer, sendo descartados os grupos que apresentaram amostragem inferior a 50 pacientes, restando apenas os seguintes grupos: cabeça, face e pescoço; cólon e reto; colo do útero; estômago; fígado; linfonodos; mama; pâncreas; pele; próstata; pulmão. No total os grupos somaram 5.568.

Variáveis representadas: idade, primeiro tratamento (prim_tratamento), progressão do tratamento (prog_tratamento), classificação TNM patológica (PTNM), estadiamento da doença (estadiamento), classificação TNM de tumores malignos (TNM), ocupação do paciente (ocupa), histórico de tabagismo (TABAGISMO), sexo, raça ou cor, histologia do tumor primário (histologia), histórico de alcoolismo (ALCOOLISMO), lateralidade do tumor (LATERALIDADE), base utilizada para diagnóstico (base_diag), histórico familiar de câncer (hist_familia), metástase a distância (met_dist), escolaridade do paciente (escolaridade) e localização primária do tumor (loctupri).

4.2 ANÁLISE DE SOBREVIVÊNCIA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Foi utilizada a biblioteca randomForestSRC disponível na plataforma R Studio,

para análise do banco de dados do Registro Hospitalar de Câncer, com o objetivo de se classificar as variáveis em ordem de relevância para predição de sobrevida, além de avaliar performance do algoritmo no processo de predição em si (classificação dos indivíduos em diferentes níveis de risco). A classificação das variáveis foi quantificada pelo *vimp* (*variable importance*), valor numérico que corresponde ao valor que a remoção da variável iria adicionar a taxa de erro, caso seja classificada como de importância positiva, ou valor que a sua remoção iria diminuir na taxa de erro, caso a mesma seja classificada como de importância negativa.

Para todos os grupos, foi configurado um padrão de criação de 1000 árvores, utilizando divisão de log-rank (*log-rank splitting*), e incluindo os campos com informações ausentes por meio da imputação adaptativa da árvore. A taxa de erro foi avaliada utilizando o Brier Score (índice que varia de 0 a 1), onde 0 indica predição perfeitamente correta, além de taxa de erro de predição de acordo com o número de árvores. O período de tempo foi representado em dias nas análises de sobrevida, em um total de 365 dias.

4.3 ASPECTOS ÉTICOS

Os dados clínicos utilizados se encontram disponíveis no Integrador do Registro Hospitalar de Câncer (IntegradorRHC) e não possuem identificação dos pacientes. Os resultados apresentados também não possibilitam a identificação individual de nenhum paciente incluído no estudo.

5 RESULTADOS E DISCUSSÃO

5.1 TAXA DE ERRO DO MODELO DE PREDIÇÃO

Técnicas de aprendizado de máquina como Support Vector Machine (SVM), 1-Nearest Neighbor (1NN), Multilayer Perception (MLP) e Random Forest (RF) são amplamente utilizadas para desenvolvimento de ferramentas para diagnóstico ou tomada de decisões médicas (MONTAZERI *et al*, 2016; ALIČKOVIĆ *et al*, 2017). Em todos os casos, é importante que se avalie a precisão da técnica, onde com o mesmo conjunto de dados é possível que diferentes técnicas apresentem performances distintas.

Na Tabela 1 são comparadas as diferentes taxas de erro entre todos os grupos de pacientes presentes no banco de dados, onde a menor taxa foi de 7.11% para o grupo do câncer de pele, e a maior taxa de 65% para o grupo do câncer de fígado.

Tabela 1 – Taxa de erro de predição: Taxa de erro de predição e média do valor do Brier Score para cada tipo de câncer, considerando-se um total de 1000 árvores criadas.

Tipo de câncer	Taxa de erro	Brier Score	Número de indivíduos
Cabeça, face e pescoço	11.89%	0.1	855
Cólon e reto	18.69%	0.1	321
Colo do útero	15.31%	0.09	233
Estômago	26.72%	0.15	253
Fígado	65%	0.26	59
Linfonodos	16.48%	0.15	207
Mama	10.14%	0.09	981
Pâncreas	59.86%	0.25	53
Pele	7.11%	0.01	2.038
Próstata	10.84%	0.05	629
Pulmão	30.35%	0.19	248
			Total: 5.877

Fonte: autoria própria.

No caso do RF, é possível avaliar a taxa de erro na predição ao se comparar as previsões realizadas pelo algoritmo no grupo de treinamento com o conjunto de dados desconhecido pelo programa (grupo OOB), além de se avaliar também o valor do Brier Score.

O índice Brier é uma medida que varia de 0 a 1 e é utilizada para avaliar previsões de respostas binárias, onde 0 é o valor correspondente a uma previsão perfeita do resultado; quanto mais próximo de 1 menor a precisão do modelo avaliado (ASSEL *et al*, 2017). Dentre os valores observados, a maioria se encontrou próxima a 0, com destaque para os grupo do câncer de pele e colo do útero, onde o valor foi em média de 0.01 e 0.09, respectivamente, embora os dois grupos apresentem amostragem de tamanhos discrepantes.

Os diferentes atributos de cada conjunto de pacientes contribuem para a diferença entre as taxas de erro, uma vez que a quantidade de dados ausentes nas variáveis de entrada é elevada em alguns grupos, como no de câncer de pâncreas, fígado, estômago e pulmão. Ao se utilizar variáveis com dados ausentes, o risco calculado para o paciente pode ser maior que o risco calculado com base em dados disponíveis, comprometendo a predição correta (STEELE *et al*, 2018).

A performance também é melhorada quando se inserem diferentes tipos de dados, além daqueles provenientes de registros clínicos, como dados referentes aos medicamentos utilizados, variáveis genéticas, imagens de testes laboratoriais, entre outras informações relevantes (ZHAO *et al*, 2015). Para alcançar resultados mais robustos, é necessária a estruturação de uma base de dados clínicos relacionados com as diversas variáveis associadas ao câncer, uma doença considerada multifatorial.

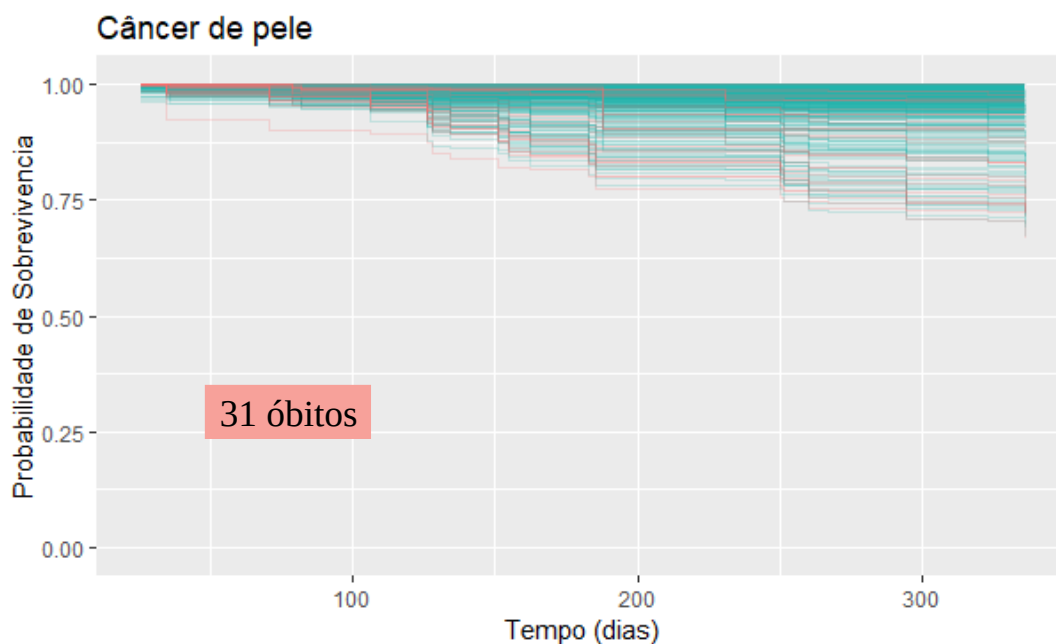
5.2 ANÁLISES DE SOBREVIDA E CLASSIFICAÇÃO DAS VARIÁVEIS

Os resultados das análises para cada grupo podem ser visualizados em gráficos de sobrevivência (Figura 2), onde se avalia a probabilidade de ocorrência do evento de óbito ao longo do tempo, e é possível visualizar em qual ocorrência os indivíduos foram classificados. Nos gráficos são representados apenas os indivíduos não incluídos no processo de treinamento do algoritmo, ou seja, aqueles cuja sobrevivência a máquina irá

prever. Nas figuras a seguir, são apresentados os grupos que apresentaram taxa de erro abaixo de 20%. Resumidamente, os grupos que apresentaram boa predição foram os de câncer de pele, mama, próstata, colo do útero, cólon e reto, sistema linfático e cabeça, face e pescoço.

Considerando o grupo do câncer de pele, se observa uma alta probabilidade de sobrevivência para todos os indivíduos (Figura 2), com poucos casos de óbito previstos (linhas vermelhas), sendo 31 o total de óbitos previstos em um grupo de 2.038 pacientes, e a probabilidade média total de sobrevivência acima de 95%. O grupo incluiu o câncer de pele dos tipos melanoma e não-melanoma.

Figura 2 – Gráfico de sobrevida para o câncer de pele: Probabilidade de sobrevivência ao longo do tempo, para o grupo de pacientes não-treinado (grupo *Out-of-bag*). Em vermelho previsão de ocorrências de óbito, em azul sobrevivência.

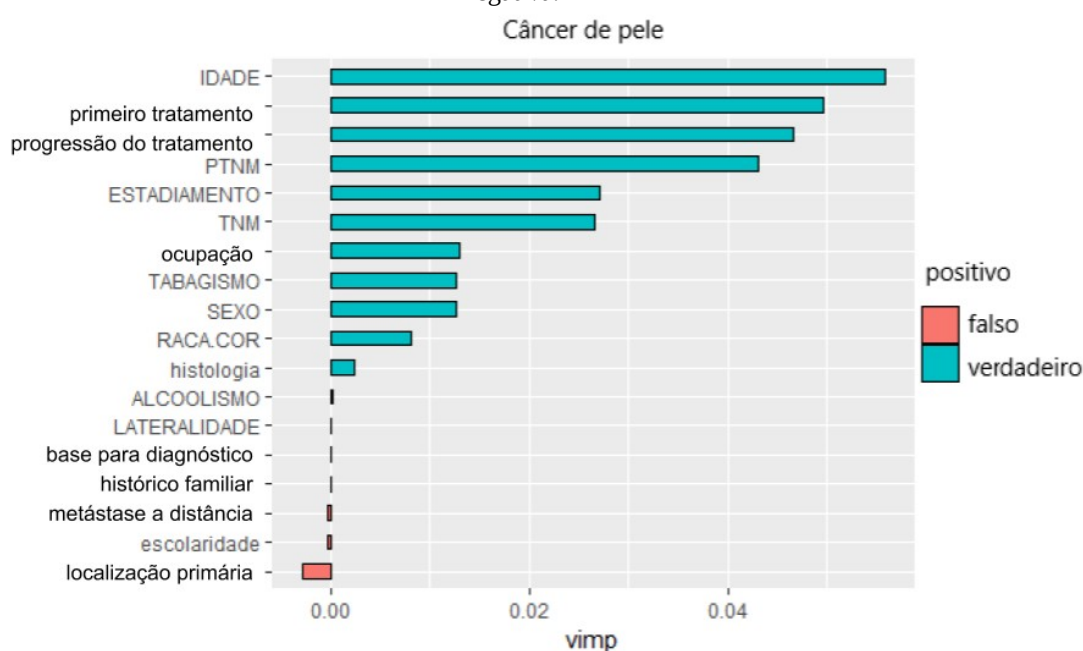


Fonte: autoria própria.

As variáveis classificadas como mais relevantes para predição do resultado citado acima (Figura 2) estão representadas na Figura 3, as quais possuem valores de importância atribuídos. Os valores indicam numericamente a mudança que a remoção da variável provocaria na taxa de erro, sendo a mais importante a que adicionaria um maior valor a essa taxa (AKRITAS *et al*, 2014). Variáveis com valores abaixo de zero indicam que a omissão das mesmas melhoram a performance da predição (LAAS *et al*, 2019). Com a classificação dada pelo algoritmo, são identificados os fatores clínicos

mais relevantes para o prognóstico da doença, considerando o padrão regional dos pacientes.

Figura 3 - Classificação de variáveis para o câncer de pele: Importância das variáveis para predição (*vimp*: *variable importance*). Barras em azul representam influência positiva, vermelho influência negativa.



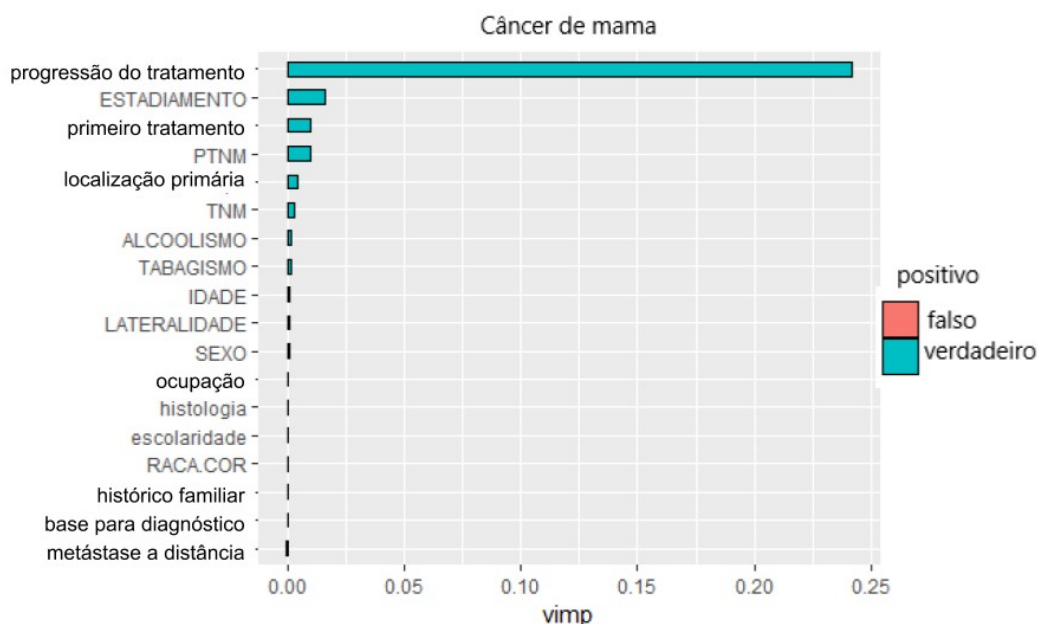
Fonte: autoria própria.

A idade é apontada como principal indicador dentro dos atributos dos pacientes, o que é esperado para o tipo de câncer em questão, uma vez que o mesmo é mais frequente em pessoas de idade mais avançada (QUEEN, 2017). As variáveis referentes ao tratamento também se posicionam no topo do gráfico, apontando a influência do primeiro tratamento utilizado e da progressão do mesmo, demonstrando que no grupo avaliado existe uma correlação positiva entre esses indicadores e a sobrevida do paciente. Em seguida, variáveis relativas a classificação patológica do tumor e estadiamento da doença são importantes por serem indicadores clínicos já consolidados (LIM *et al*, 2018; HUANG *et al*, 2017).

Na figura 4 é apresentado o mesmo gráfico de classificação de variáveis para o grupo do câncer de mama, onde se observa que no conjunto de dados em questão a progressão do tratamento é apontado como maior indicador para previsão de sobrevivência. Pode-se considerar que o padrão do grupo é determinado majoritariamente pelo resultado do tratamento. Os avanços nos tratamentos para o câncer nos últimos anos têm demonstrado um aumento na sobrevida de grupos de

diferentes idades (VAN ABEMMA *et al*, 2019). No Brasil, estudos avaliando a eficiência dos tratamentos por meio de análises de sobrevida mostram um padrão de aumento dessa taxa nas regiões mais desenvolvidas no país (Sul e Sudeste), onde se destaca a importância do diagnóstico e tratamento efetivos (RESENDE *et al*, 2017).

Figura 4 – Classificação das variáveis para o câncer de mama: Classificação das variáveis de acordo com importância para predição (vimp). Barras em azul indicam influência positiva, em vermelho influência negativa.

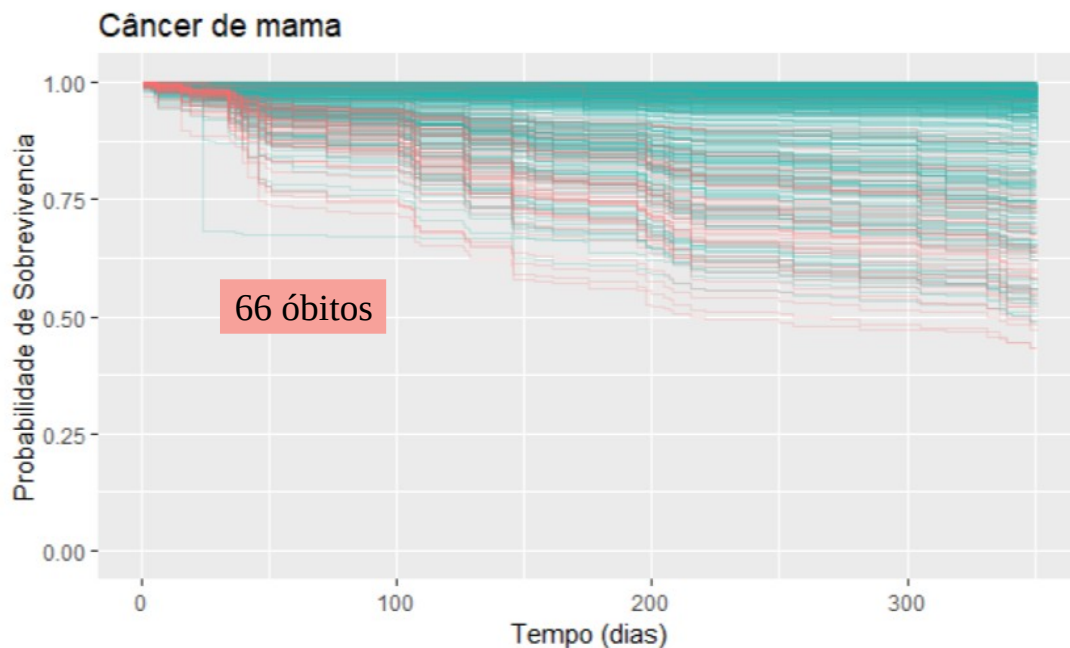


Fonte: autoria própria.

Dentre os outros fatores, o estadiamento e pTNM são variáveis consideradas como importantes para o prognóstico do câncer de mama em análise de sobrevida na região Sudeste do Brasil, considerando um período de dez anos, utilizando modelo de riscos proporcionais de Cox, onde também foi classificada a variável relacionada com o tamanho do tumor (FAYER *et al*, 2016).

Em relação a probabilidade de sobrevivência para o câncer de mama (Figura 5), foram previstos 66 óbitos, com uma taxa de sobrevida acima de 40% para todas as pacientes, sendo a média total acima de 90%.

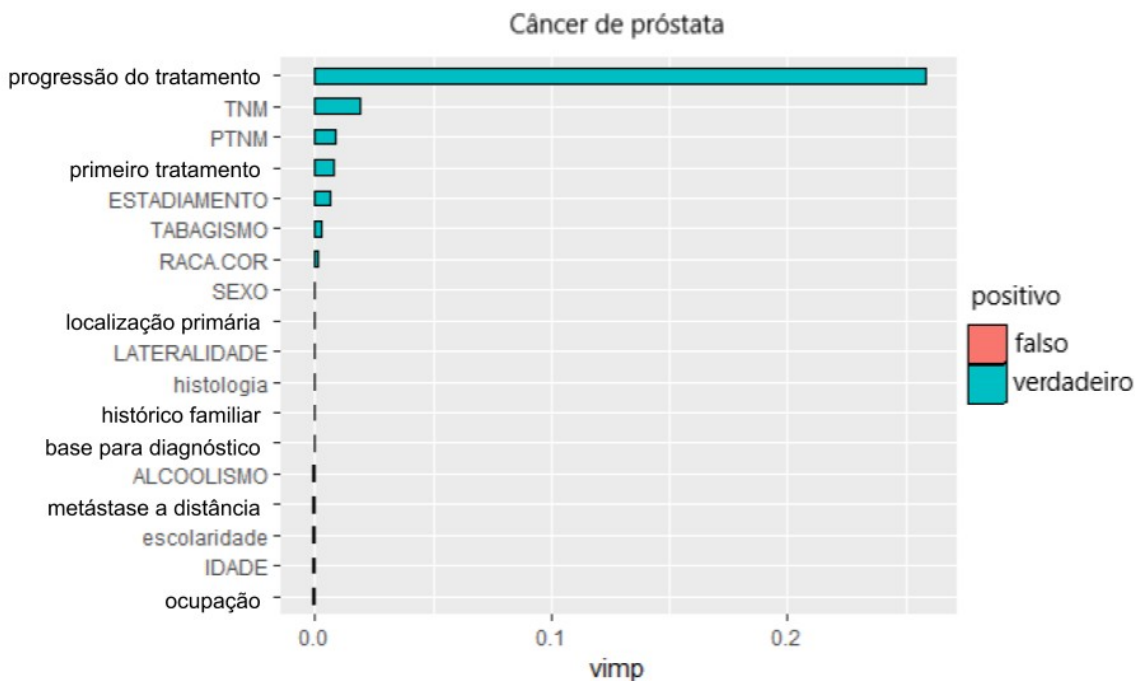
Figura 5 – Sobrevivência para o câncer de mama: Probabilidade de sobrevivência ao longo do tempo, para o grupo de pacientes não-treinado (grupo *Out-of-bag*). Em linhas vermelhas previsão de ocorrências de óbito, em azul sobrevivência.



Fonte: autoria própria.

Em relação aos fatores prognósticos para o câncer de próstata (Figura 6), as variáveis classificadas como mais importantes foram progressão do tratamento, TNM e pTNM. Estratégias de diagnóstico (TNM e pTNM) e de tratamento (progressão ou eficiência) são relacionadas com a sobrevida quando se deseja avaliar a eficiência de serviços de saúde aplicados em uma população, levando a uma otimização no processo de tomada de decisões para cada tipo de câncer (LITWIN *et al*, 2017).

Figura 6 – Classificação das variáveis para o câncer de próstata: Classificação das variáveis de acordo com importância para predição (vimp). Barras em azul indicam influência positiva, em vermelho influência negativa.

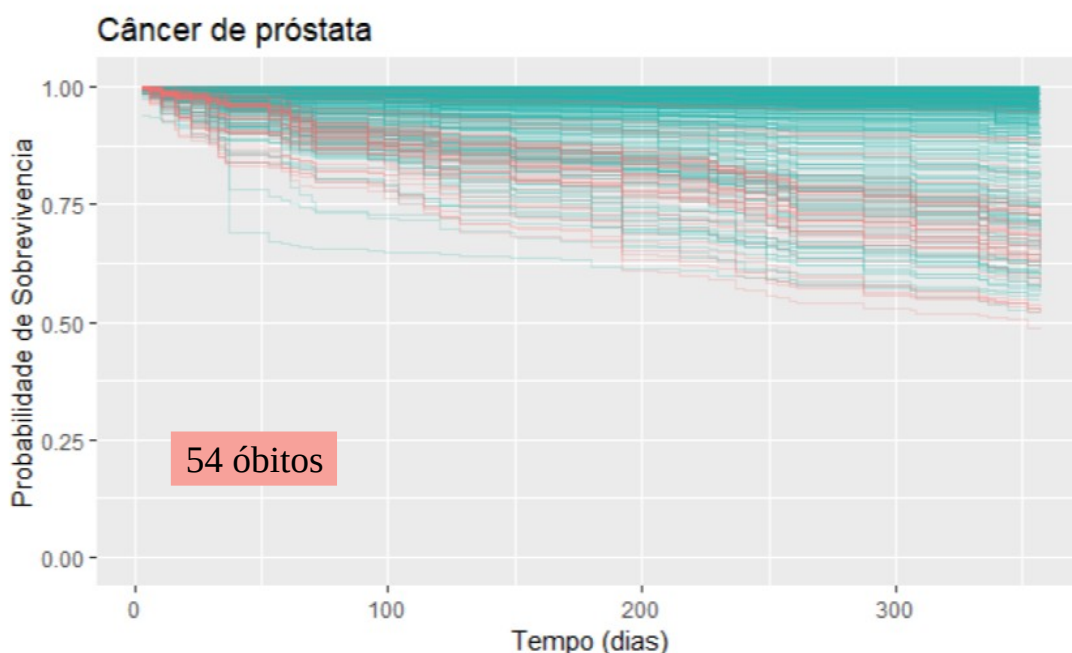


Fonte: autoria própria.

Considerando as predições acima, é possível verificar que o estadiamento do tumor é uma variável de importância moderada, sendo o sistema de classificação TNM dos tumores malignos amplamente aplicado para realizar esta classificação.

Para o grupo do câncer de próstata (Figura 7), a probabilidade de sobrevivência foi acima de 45% para todos os pacientes, sendo a média total acima de 90%, com observação de 54 óbitos segundo a previsão.

Figura 7 – Sobrevivência para o câncer de próstata: Probabilidade de sobrevivência ao longo do tempo, para o grupo de pacientes não-treinado (grupo *Out-of-bag*). Em linhas vermelhas previsão de ocorrências de óbito, em azul sobrevivência.

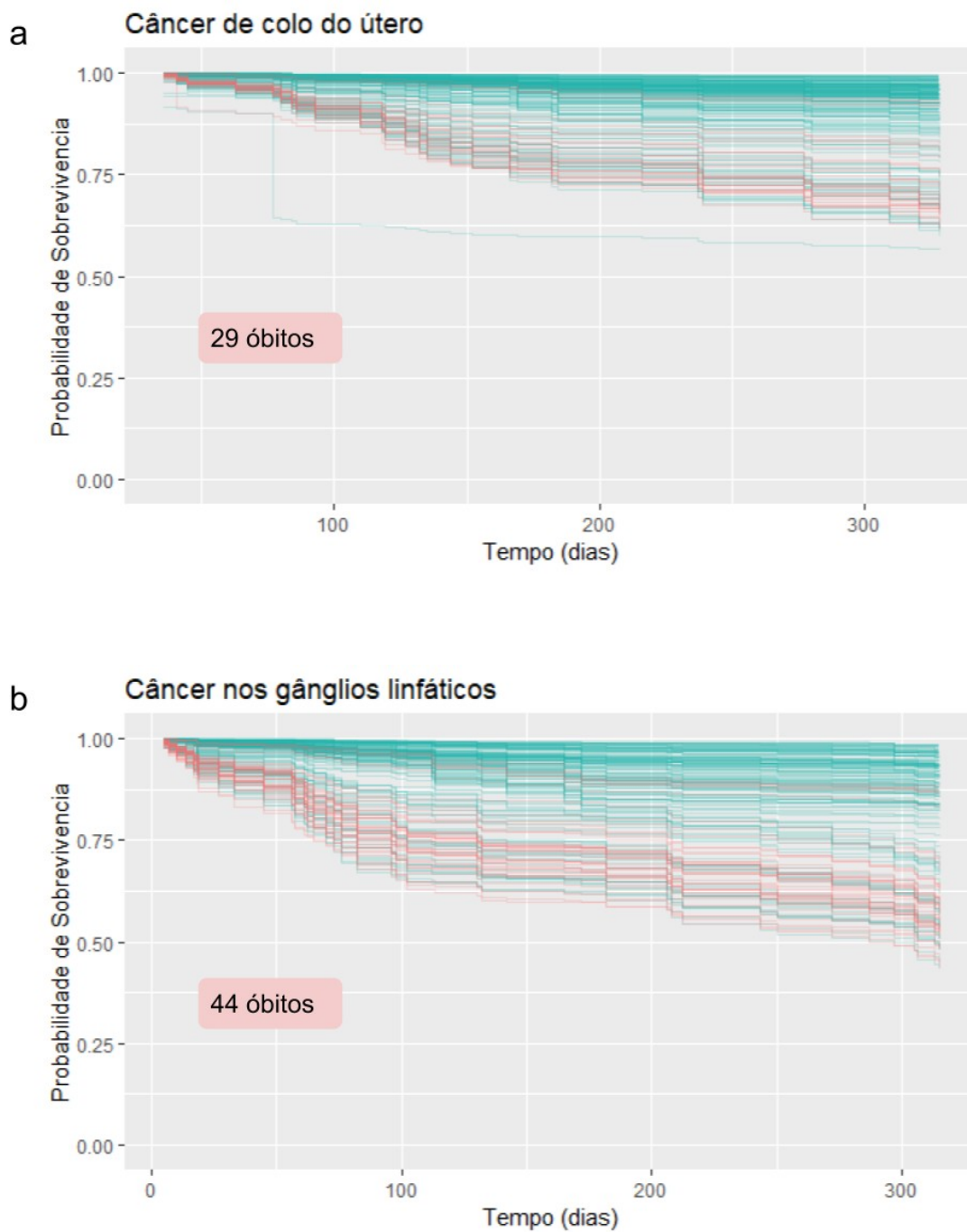


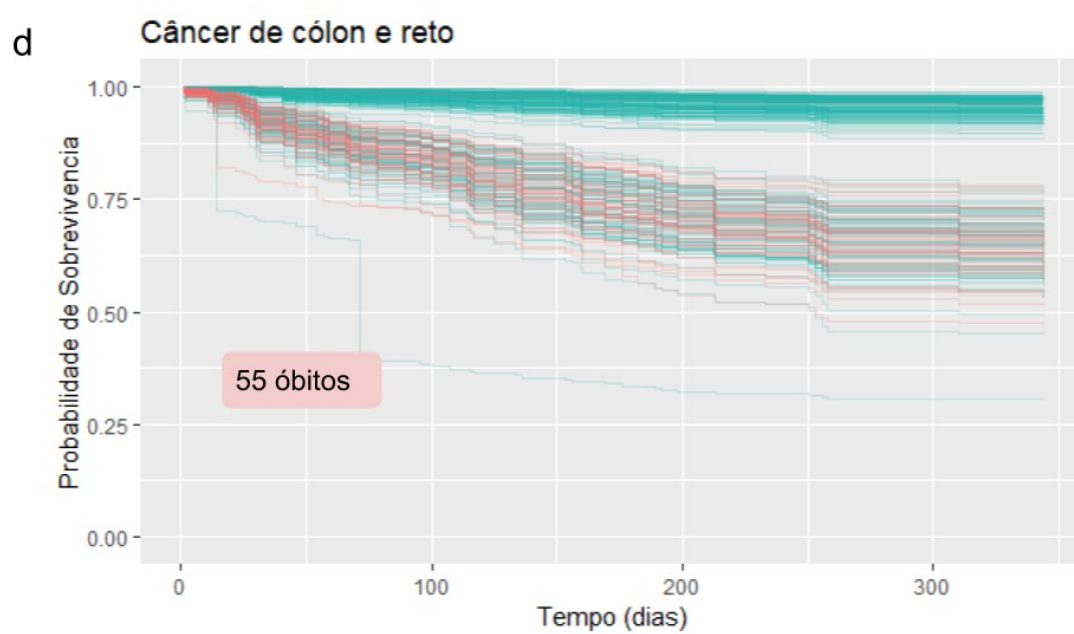
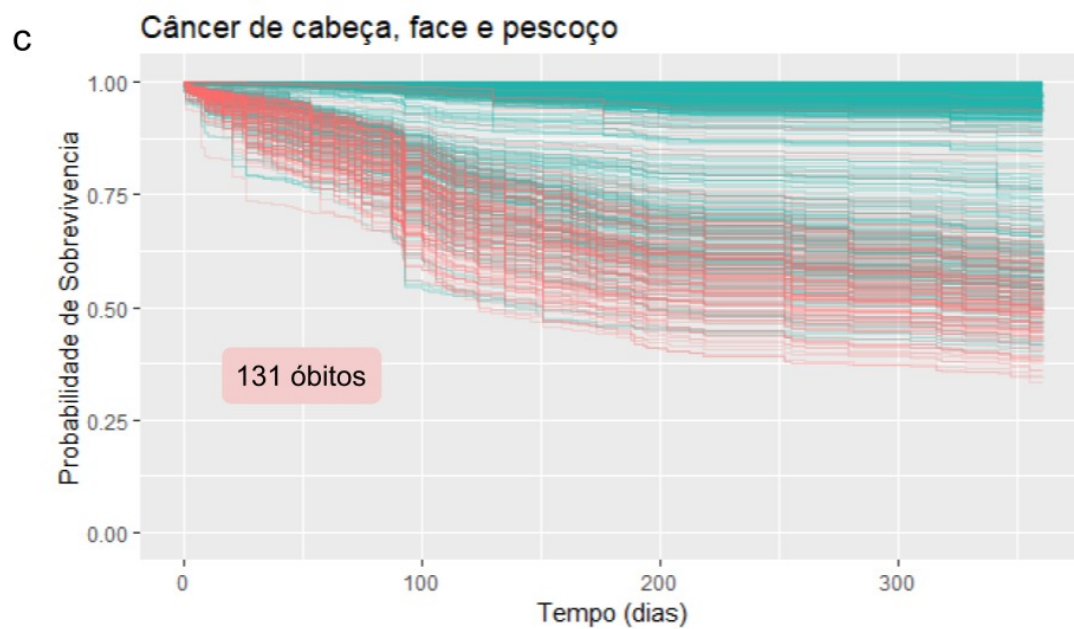
Fonte: autoria própria.

A classificação das variáveis para o restante dos grupos com taxa de erro abaixo de 20% seguiu um padrão onde a progressão do tratamento foi apresentada como variável mais importante, existindo variação apenas no valor do *vimp* em cada grupo. Em relação a taxa de sobrevivência, o grupo do câncer de colo do útero (Figura 8) apresentou probabilidade acima de 50% para todos os pacientes, sendo a média total acima de 80%, com 29 óbitos previstos.

Já para o câncer de cabeça (Figura 8), face e pescoço a probabilidade para todos os pacientes foi acima de 25%, sendo a média total acima de 75%, com 131 óbitos previstos. Para o câncer nos gânglios linfáticos (Figura 8), a probabilidade para todos os pacientes foi acima de 45% e média total acima de 75%, com 44 óbitos previstos. No grupo do câncer de cólon e reto (Figura 8), a probabilidade para todos os pacientes foi acima de 25% e média total acima de 80%, com 55 óbitos previstos.

Figura 8 - Gráficos de sobrevida: Probabilidade de sobrevivência ao longo do tempo, para o grupo de pacientes não-treinado (grupo *Out-of-bag*), representando de cima para baixo: (a) câncer de colo do útero; (b) câncer nos gânglios linfáticos; (c) câncer de cabeça, face e pescoço; (d) câncer de cólon e reto. Em vermelho previsão de ocorrências de óbito, em azul sobrevivência.





Fonte: autoria própria.

6 CONSIDERAÇÕES FINAIS

A partir dos dados clínicos provenientes dos Registros Hospitalares de Câncer, foi possível realizar análises de sobrevida considerando todas as variáveis do prontuário dos pacientes, com uma precisão relativamente alta para 4 dos 11 grupos avaliados. Com a classificação das variáveis mais importantes para a predição dos eventos, se observou o padrão característico da base de dados dos pacientes atendidos pelo hospital da Liga Mossoroense de Estudos e Combate ao Câncer (LMECC) durante o período de 2006 a 2016. O padrão correspondeu a uma resposta relacionada principalmente com a progressão do tratamento dos pacientes de todos os grupos, com exceção do grupo do câncer de pele, onde a variável mais importante correspondeu a idade.

As técnicas de aprendizado de máquina permitem uma avaliação global dos dados disponíveis, mostrando resultados que podem ser interpretados como informação importante para estudos epidemiológicos, mesmo em casos onde a amostragem é pequena e há ausência de informação em diversas variáveis.

REFERÊNCIAS

- ADAMU, P.I; ADAMU, M.O; OKAGBUE, H.I; OPOOLA, L; BISHOP, S.A. Survival Analysis of Cancer Patients in North Eastern Nigeria from 2004 – 2017 – A Kaplan - Meier Method. **Open Access Macedonian Journal of Medical Sciences**, v. 7, p. 643-650. 2019.
- ALIČKOVIĆ, Emina; SUBASI, Abdulhamit. Breast cancer diagnosis using GA feature selection and Rotation Forest. **Neural Computing and Applications**, v. 28, n. 4, p. 753–763, 2017.
- AKRITAS, Michael G.; LAHIRI, S. N.; POLITIS, Dimitris N. **Topics in Nonparametric Statistics: Proceedings of the First Conference of the International Society for Nonparametric Statistics**. p. 207-215. [s.l.]: Springer, 2014.
- AMERICAN CANCER SOCIETY. **Cancer Facts & Figures 2018**. p. 76, 1930.
- AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. Dissertação em Ciência da Computação - Universidade Federal de Pernambuco, Centro de Informática, Recife, 2006.
- ASSEL, Melissa; SJOBERG, Daniel D.; VICKERS, Andrew J. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. **Diagnostic and Prognostic Research**, v. 1, n. 1, p. 19, 2017.
- Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Sistema de Informações de Mortalidade [banco de dados na Internet]**. Brasília: Datasus; c2008. [cited 2011 Jun 1]. Available from: <http://www.datasus.gov.br>
- BRAY, Freddie; FERLAY, Jacques; SOERJOMATARAM, Isabelle; SIEGEL, Rebecca L.; TORRE, Lindsey A.; JEMAL, Ahmedin. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 68, n. 6, p. 394–424, 2018.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- BÜHNEMANN, Claudia; LI, Simon; YU, Haiyue; *et al.* Quantification of the Heterogeneity of Prognostic Cellular Biomarkers in Ewing Sarcoma Using Automated Image and Random Survival Forest Analysis. **PLoS ONE**, v. 9, n. 9, 2014. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4171480/>>. Acesso em: 13 fev. 2020.
- Cancer Research UK. **Lung cancer risk**. Disponível em: <<https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/risk-factors>>. Acesso em: 13 fev. 2020.

DO NASCIMENTO, Marcelo Zanchetta; MARTINS, Alessandro Santana; AZEVEDO TOSTA, Thaína Aparecida; NEVES, Leandro Alves. Lymphoma images analysis using morphological and non-morphological descriptors for classification. **Computer Methods and Programs in Biomedicine**, v. 163, p. 65–77, 2018.

EHRLINGER, J.; BLACKSTONE, E. H. ggRandomForests : Survival with Random Forests. In: [s.l.: s.n.], 2015. Disponível em: <https://pdfs.semanticscholar.org/386f/d1e44fae06e0ab73e8cc24f7b75fd6935ac3.pdf>> Acesso em: 13 fev. 2020.

ESMAILY, Habibollah; TAYEFI, Maryam; DOOSTI, Hassan; *et al.* A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. **Journal of Research in Health Sciences**, v. 18, n. 2, p. e00412, 2018.

ETIKAN, İ; ABUBAKAR, S; ALKASSIM, R. The Kaplan Meier Estimate in Survival Analysis. **Biometris & Biostatistics International Journal**, v. 5, p. 1.5. 2017.

FAYER, Vívian Assis; GUERRA, Maximiliano Ribeiro; CINTRA, Jane Rocha Duarte; *et al.* Sobrevida de dez anos e fatores prognósticos para o câncer de mama na região Sudeste do Brasil. **Revista Brasileira de Epidemiologia**, v. 19, n. 4, p. 766–778, 2016.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI MAGAZINE**, v. 17, n. 3, p. 37-54, 1996.

FERLAY, J; COLOMBET, M; SOERJOMATARAM, I; DYBA, T; RANDI, G; BETTIO, M; GAVIN, A; VISSER, O; BRAY, F. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. **European Journal of Cancer**, v. 103, p. 356-387, 2018.

FIDLER, M. M.; BRAY, F.; SOERJOMATARAM, I. The global cancer burden and human development: A review. **Scandinavian Journal of Public Health**, v. 46, p. 27–36. 2018.

GAYVERT, Kaitlyn M.; ALY, Omar; PLATT, James; *et al.* A Computational Approach for Identifying Synergistic Drug Combinations. **PLOS Computational Biology**, v. 13, n. 1, p. 1-11, 2017.

GIACOMELLI, Iron Pedro; STEIDLE, Leila John Marques; GIACOMELLI, Irai Luis; NESI, William Mazzuco; MORAL, Joanita A. G. Del; PINCELLI, Mariangela Pimentel. CÂNCER DE PULMÃO: DADOS DE TRÊS ANOS DO REGISTRO HOSPITALAR DE CÂNCER DE UM HOSPITAL DO SUL DO BRASIL. **Arquivos Catarinenses de Medicina**, v. 46, n. 3, p. 129–146, 2017.

GURM, H.S.; KOOIMAN J.; LALONDE T.; GRINES C.; SHARE D. *et al.* A Random Forest Based Risk Model for Reliable and Accurate Prediction of Receipt of Transfusion in Patients Undergoing Percutaneous Coronary Intervention. **PLOS ONE**. 2014. Disponível em:

<<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0096385>>. Acesso em: 13 fev. 2020.

HASHEMI, Farahnaz Sadat Golestan; ISMAIL, Mohd Razi; YUSOP, Mohd Rafii; HASHEMI, Mahboobe S. G; SHAHRAKI, Mohammad H. N.; RASTEGARI, Hamid; MIAH, Gous; ASLANI, Farzad. Intelligent mining of large-scale bio-data: Bioinformatics applications: Bioinformatics applications. **Biotechnology & Biotechnological Equipment**, v. 32, n. 1. Disponível em: <<https://www.tandfonline.com/doi/full/10.1080/13102818.2017.1364977>>. Acesso em: 13 fev. 2020.

HU, Chen; STEINGRIMSSON, Jon Arni. Personalized Risk Prediction in Clinical Oncology Research: Applications and Practical Issues Using Survival Trees and Random Forests. **Journal of Biopharmaceutical Statistics**, v. 28, n. 2, p. 333–349, 2018.

HUANG, Mi; VAN SOEST, J; ZHONG, H; BEN-JOSEF, E; GENG, Huaizhi; WANG, J; ZHANG, Zheng; MOHIUDDIN, M; MEROPOL, N.J.; GAROFALO, M; ANNE, P.R.; LEE, R.J.; ERICKSON, B.A.; AUGSPURGER, M; SIGURDSON, E.R.; LASIO, Giovanni; DEKKER, André; XIAO, Ying. Rectal Cancer Outcome Study Based on Multiple NRG Oncology Clinical Trials with Random Survival Forests. **International Journal of Radiation Oncology**, v. 99, p. 168, 2017.

HUANG, S.H.; O’SULLIVAN, B. Overview of the 8th Edition TNM Classification for Head and Neck Cancer. **Curr. Treat. Options in Oncol**, v. 18, n. 40, 2017.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA (INCA). O que é câncer? **Instituto Nacional de Câncer José Alencar Gomes da Silva**. Disponível em: <<https://www.inca.gov.br/o-que-e-cancer>> Acesso em: 13 fev. 2020.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA (INCA). **Estadramento**. INCA - Instituto Nacional de Câncer. Disponível em: <<https://www.inca.gov.br/estadiamento>>. Acesso em: 13 fev. 2020.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA (INCA). Estimativa 2018: incidência de câncer no Brasil. **Instituto Nacional de Câncer José Alencar Gomes da Silva**. Coordenação de Prevenção e Vigilância. Rio de Janeiro: INCA, 2017.

ISHWARAN, H; KOGALUR, U. B. Random Survival Forests for R. **R News**, v. 7/2, p. 25. 2007.

ISHWARAN, H.; KOGALUR, U. randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC). **ResearchGate**. 2016.

JOTHI, Neesha; RASHID, Nur’Aini Abdul; HUSAIN, Wahidah. Data Mining in Healthcare – A Review. **Procedia Computer Science**, v. 72, p. 306–313, 2015.

KAUR, Chandee; OMISAKIN, Olufemi Muibi. Data Mining Methods to Improve Clinical Trials in Diabetic Patients. **Annals of Clinical and Laboratory Research**, v. 06, n. 04, 2018. Disponível em: <<http://www.aclr.com.es/clinical-research/data-mining-methods-to-improve-clinical-trials-in-diabetic-patients.php?aid=23709>>. Acesso em: 13 fev. 2020.

LAAS, Enora; HAMY, Anne-Sophie; MICHEL, Anne-Sophie; *et al.* Impact of time to local recurrence on the occurrence of metastasis in breast cancer patients treated with neoadjuvant chemotherapy: A random forest survival approach. **PLOS ONE**, v. 14, n. 1, p. 1-14, 2019.

LAN, Kun; WANG, Dan-tong; FONG, Simon; LIU, Lian-Sheng; DEY, Nilanjan. A Survey of Data Mining and Deep Learning in Bioinformatics. **Journal of Medical Systems**, v. 42, 2018.

LÁNCZKY, A; NAGY, A; BOTTAL, G; MUNKÁCSY, G; SZABÓ, A; SANTARPIA, L; GYÓRFFY, B. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. **Breast Cancer Research and Treatment**, v. 160, p. 439-446, 2016.

LIM, Wanyin; RIDGE, Carole A.; NICHOLSON, Andrew G.; MIRSADRAEE, Saeed. The 8th lung cancer TNM classification and clinical staging system: review of the changes and clinical implications. **Quantitative imaging in medicine and surgery**, v. 8, p. 709-718, 2018

LITWIN, Mark S.; TAN, Hung-Jui. The Diagnosis and Treatment of Prostate Cancer: A Review. **JAMA**, v. 317, n. 24, p. 2532–2542, 2017.

MAHAR, Alyson L.; ZAGORSKI, Brandon; KAGEDAN, Daniel; *et al.* Evaluating TNM stage prognostic ability in a population-based cohort of gastric adenocarcinoma patients in a low-incidence country. **Canadian Journal of Public Health**, v. 109, n. 4, p. 480–488, 2018.

MINISTÉRIO DA SAÚDE. **Atlas On-line de Mortalidade**. 1996-2014. Disponível em: <<https://mortalidade.inca.gov.br/MortalidadeWeb/>>. Acesso em: 13 fev. 2020.

MONTAZERI, Mitra; MONTAZERI, Mohadeseh; MONTAZERI, Mahdieh; *et al.* Machine learning models in breast cancer survival prediction. **Technology and Health Care**, v. 24, n. 1, p. 31–42, 2016.

MURDOCH, Travis B.; DETSKY, Allan S. The Inevitable Application of Big Data to Health Care. **JAMA**, v. 309, n. 13, p. 1351–1352, 2013.

NATIONAL CANCER INSTITUTE. **What Is Cancer?** Disponível em: <<https://www.cancer.gov/about-cancer/understanding/what-is-cancer>>. Acesso em: 13 fev. 2020.

NOMA, Daisuke; INAMURA, Kentaro; MATSUURA, Yosuke; *et al.* Prognostic Effect of Lymphovascular Invasion on TNM Staging in Stage I Non–Small-cell Lung Cancer. **Clinical Lung Cancer**, v. 19, n. 1, p. 109–122, 2018.

QUEEN, Lauren. Skin Cancer: Causes, Prevention, and Treatment. **Senior Honors Theses**, 2017. Disponível em: <<https://digitalcommons.liberty.edu/honors/648>>. Acesso em: 13. fev. 2020.

RESENDE, L. T.; PIMENTEL, M. S. L.; MENEZES, M. O.; GONÇALVES, L. L. C. Câncer de mama: avaliação de sobrevida. **International Nursing Congress: Good practices of nursing representations in the construction of society**, v. 1, n.1, 2017.

SELIGSON, DB.; HORVATH, S.; SHI, T.; YU, H.; TZE, S. *et al.* Global histone modification patterns predict risk of prostate cancer recurrence. **Nature**, v.435, p. 1262–1266, 2005.

SELLAMY, K.; FAKHRI, Y.; BOULAKNADEL, S.; *et al.* Web mining techniques and applications: Literature review and a proposal approach to improve performance of employment for young graduate in Morocco. **2018 International Conference on Intelligent Systems and Computer Vision (ISCV)**. [s.l.: s.n.], p. 1-5, 2018.

SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2019. **CA: A Cancer Journal for Clinicians**, v. 69, n.1, p. 7-34, 2019.

SFERRA, H. H.; CORRÊA, A. M. C. J. Data Mining Concepts and Applications. **Revista de Ciência e Tecnologia**, v. 11, p. 16, 2003.

SILVA FILHO, A. S. Inferência em amostras pequenas: métodos bootstrap. **Revista de Ciências Exatas e Tecnologia**, v. 5, n. 5, p. 115-126. 2010.

SOUZA, Mirian Carvalho de; CRUZ, Oswaldo Gonçalves; VASCONCELOS, Ana Glória Godoi; *et al.* Factors associated with disease-specific survival of patients with non-small cell lung cancer. **Jornal Brasileiro de Pneumologia**, v. 42, n. 5, p. 317–325, 2016.

STEELE, Andrew J.; DENAXAS, Spiros C.; SHAH, Anoop D.; *et al.* Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. **PLOS ONE**, v. 13, n. 8, p. 1-20, 2018.

TAYLOR, Jeremy G. Random Survival Forests. **Journal of Thoracic Oncology**, v.6, p. 1974-1975. 2011.

TOMCZAK, Katarzyna; CZERWIŃSKA, Patrycja; WIZNEROWICZ, Maciej. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. **Contemporary Oncology**, v. 19, n. 1A, p. A68–A77, 2015.

VAN ABBEMA, D.; VISSERS, P.; DE VOS-GEELEN, J.; LEMMENS, V.; JANSSEN-HEIJNEN, M.; TJAN-HEIJNEN, V. Trends in Overall Survival and Treatment Patterns in Two Large Population-Based Cohorts of Patients with Breast and Colorectal Cancer. **Cancers**, v.11, p. 1-15, 2019.

XU, Chuan; CHEN, Jiajun; XU, Xia; ZHANG, Yingyu. Potential Therapeutic Drugs for Parkinson's Disease Based on Data Mining and Bioinformatics Analysis. **Research Gate**. Disponível em:

<https://www.researchgate.net/publication/328034127_Potential_Therapeutic_Drugs_for_Parkinson's_Disease_Based_on_Data_Mining_and_Bioinformatics_Analysis>.

Acesso em: 13 fev. 2020.

ZHAO, Jing; HENRIKSSON, Aron; ASKER, Lars; *et al.* Predictive modeling of structured electronic health records for adverse drug event detection. **BMC Medical Informatics and Decision Making**, v. 15, n. 4, p. S1, 2015.