Aplicação do *Random Survival Forest* na análise da sobrevida para câncer da mama

Application of Random Survival Forest in breast cancer survival analysis

Aplicación del *Random Survival Forest* en el análisis de la supervivencia del cáncer de mama

Daniela Schimitz de Carvalho[1],Thallys da Silva Nogueira[1], Priscila Vanessa Zabala Capriles Goliatt [1]

1 Programa de Pós-Graduação em Modelagem Computacional, Universidade Federal de Juiz de Fora – UFJF, Juiz de Fora (MG), Brasil.

Autor correspondente: Daniela Schimitz de Carvalho
E-mail: daniela.schimitz@estudante.ufjf.br

Resumo

Este trabalho tem por objetivo aplicar um método de aprendizado de máquina supervisionado a um conjunto de dados clínicos da Zona da Mata Mineira, para se avaliar o desempenho da precisão da predição de sobrevida para câncer de mama. O banco de dados utilizado passou por pré-processamento fornecendo as variáveis a serem empregadas no *Random Survival Forest*. Os resultados apresentam as métricas de desempenho satisfatória para métodos de predição da sobrevida. Sendo concluído, que os métodos de aprendizagem de máquina são promissores na assistência e orientação na prática clínica.

Descritores: Câncer de Mama; Aprendizado de Máquina; Análise de Sobrevida

Abstract

This paper aims to apply a supervised machine learning method to a clinical dataset from Zona da Mata Mineira, to evaluate the performance of survival prediction accuracy for breast cancer. The database utilized went through pre-processing providing the variables used in the Random Survival Forest. The results show satisfactory

performance metrics for survival prediction methods. Concluding that, the machine learning methods are promising assisting and guiding clinical practice.
Keywords: Breast Neoplasms; Machine Learning; Survival Analysis

Resumen

Este trabajo tiene como objetivo aplicar un método de aprendizaje automático supervisado a un conjunto de datos clínicos de la Zona da Mata Mineira, para evaluar el rendimiento de la precisión de la predicción de la supervivencia para el cáncer de mama. La base de datos utilizada pasó por un preprocesamiento que proporcionó las variables que se emplearían en el *Random Survival Forest*. Los resultados presentan métricas de rendimiento satisfactorias para los métodos de predicción de la supervivencia. Concluyendo que los métodos de aprendizaje automático son prometedores en la asistencia y orientación en la práctica clínica.
Descriptores: Neoplasias de mama; Machine Learning; Análisis de Supervivencia

Introdução

O câncer de mama (CM) é um dos principais problemas no âmbito da saúde pública mundial, sendo o mais diagnosticado entre todos os cânceres em ambos os sexos.[1,2] Segundo, estimativas de 2020 da Organização Mundial de Saúde (OMS) esta patologia é mais prevalente entre as mulheres, com 2,26 milhões de novos casos e 685 mil óbitos.[2] Já no Brasil, 66 mil novos casos e 18 mil óbitos, segundo estimativas do Instituto Nacional do Câncer (INCA) de 2020.[3]

No que tange a saúde da mulher, o CM continua com elevadas taxas de incidência, prevalência e mortalidade. Estas altas taxas são decorrentes dos fatores de risco, fatores prognósticos e diagnósticos tardios, que podem evoluir para a forma mais agressiva da doença, a metástase sistêmica e consequentemente a morte.[1,4,6]

Entre os fatores de risco destaca-se os relacionados à vida reprodutiva da mulher: menopausa tardia, menarca precoce, não ter filhos, idade avançada no nascimento do primeiro filho e menos filhos, uso de anticoncepcionais orais e reposição hormonal.[5,7,8] E os relacionados à contemporaneidade como: estilo de vida,

envelhecimento, sedentarismo, obesidade, tabagismo e etilismo. Por fim, enfatiza-se a relevância dos fatores protetores como amamentação e atividade física.[4,7]

Já entre os fatores prognósticos, vale ressaltar os que interferem na sobrevida. O termo sobrevida indica o tempo específico em que pacientes sobreviveram a uma determinada patologia após o diagnóstico.[5,10] No caso do CM alguns fatores como: idade ao diagnóstico, tamanho do tumor, comprometimento linfonodal, estadiamento, tipo e grau histológico, marcadores moleculares e imuno-histoquímicos; influenciam a sobrevida. Na prática clínica, também se avalia os status dos receptores de estrogênio, progesterona e fator de crescimento humano epidérmico receptor-2 (HER2) para a subclassificação do CM e indicações clínico-terapêuticas específicas.[5,8,9]

Por fim, o diagnóstico precoce está diretamente relacionado ao bom prognóstico e consequentemente ao início do tratamento, evitando-se a evolução do CM para estágios mais avançados e assim melhorar a sobrevida e diminuir o sofrimento do paciente.[4,11,12] Este diagnóstico é efetuado pelo rastreamento mamográfico e exame físico (presença de nódulo), logo após, são avaliados a história pregressa, os fatores de risco e prognósticos, os resultados dos marcadores tumorais e estadiamento para assim se estabelecer as estratégias de tratamento.[5,8,13]

Diante disto, o conhecimento destes fatores são de extrema importância para o planejamento terapêutico e avaliação do curso clínico-terapêutico.[7,11,13] Atualmente, este processo é realizado de forma holística, subsequente a ponderação dos fatores significativos envolvidos no processo evolutivo desta doença.[7,8,14]

Neste sentido, os modelos computacionais fornecem informações preditivas importantes nas tomadas de decisões clínico-terapêuticas, além de preencherem lacunas da prática médica guiadas atualmente por variáveis clínicas observáveis.[9,11,14] Visto que, vive-se o desdobramento do emprego da inteligência computacional na prestação de entendimento e elucidação de problemas nas diversas áreas.[10,15]

Nesta perspectiva, os métodos de aprendizado de máquina (MAM) vêm ganhando espaço e aplicabilidade na área oncológica, fornecendo informações significativas, como, por exemplo, para o diagnóstico, manejo e prognóstico dos pacientes.[11,12,16] Consequentemente, com o avanço da ciência de dados e

inovações das tecnologias surgem novos modelos prognósticos para CM, como ferramentas iminentes de alto potencial e aplicabilidade na prática médica.[9,12,16,17]

Como, por exemplo, o MAM supervisionado *Random Survival Forest* (RSF)[18], amplamente aplicado na predição de sobrevida.[9,12,16,17] O RSF se fundamenta nos princípios do método original, *Random Forest* (RF)[19], em que se cultiva árvores usando dados de *bootstrap*, os nós se dividem usando uma seleção aleatória de recursos, e geram uma predição constituída pela média dos preditores de cada árvore.[17,18,20]

Recentemente, vários métodos de predição prognóstica foram propostos, porém, apresentam como limitação a não validação com dados locais, resultando em uma capacidade preditiva abaixo do ideal.[9,12,16,18] Deste modo, esta pesquisa tem por objetivo aplicar o RSF a um banco de dados clínicos de pacientes diagnosticadas com CM, tratadas e acompanhadas em centros de referência oncológica da Zona da Mata Mineira, a fim de avaliar o desempenho da predição de sobrevida para o CM.

Material e Métodos

Neste trabalho, aplicou-se um modelo computacional a um banco de dados clínico de pacientes diagnosticadas com CM. O MAM supervisionado selecionado RSF, implementado na linguagem Python por meio da biblioteca Scikit-survival[21]. Esta biblioteca, permite a realização de análises de sobrevida através da correlação entre as covariáveis e o tempo de evento, e também métricas de desempenho específicas.[21,22,23]

O tratamento e formatação do banco de dados foram realizados nas etapas de pré-processamento e análise descritiva dos dados. Após estas etapas, os dados resultantes foram utilizados na implementação do método para a predição da sobrevida para CM, e validação do RSF pelas métricas de desempenho.[22,24,25]

Banco de dados

Os dados são provenientes de uma coorte de base hospitalar, coletados manualmente nos prontuários dos registros hospitalares dos centros de referência de oncologia da região da Zona da Mata Mineira. O banco de dados refere-se a uma

população de mulheres diagnosticadas com CM entre janeiro de 2003 e janeiro de 2005, submetidas à terapêutica local e/ou sistêmica. Vale enfatizar que estes dados são oriundos de pesquisas já concluídas[5], que não utilizaram desta metodologia. Após a submissão e aprovação pelo Comitê de Ética na pesquisa com seres humanos da Universidade Federal de Juiz de Fora, CAAE:58758822.6.0000.51.47.

A coleta dos dados constituída pelo: recrutamento (2009 até 2010) nos arquivos de registro de câncer de base hospitalar; seguida da avaliação das características clínicas, sociodemográficas e exclusão das pacientes que realizaram apenas um único procedimento nas instituições; e buscas (2011) por ligações telefônicas, consultas de CPF válidos e com mastologistas, e no sistema de informações de mortalidade.[5]

Pré-processamento dos dados

Nos dados brutos do banco original, em consequência da forma de coleta (manual) e do número de dados faltantes, foram realizados dois pré-processamentos. O primeiro pré-processamento, através de tratamento dos dados faltantes e padronização da formatação dos dados, como segue discriminado abaixo:

1. **Strings**: corrigido os erros de digitação; retirando os acentos e caracteres; formatando em minúsculas; inserindo a palavra 'ignorado' nos dados ausentes;
2. **Valores numéricos:** formatando os números inteiros e contínuos; para os dados faltantes de variáveis categorizadas e de contagem atribuindo, respectivamente, os valores 9 e -1;
3. **Datas**: verificando os valores ausentes da data de seguimento e laudo histopatológico (critério de exclusão), e formatadas as datas (ano, mês e dia).

O segundo pré-processamento, estabelecendo um novo tratamento, resultando na construção de 3 bancos de dados:

a) **Banco I:** constituído pelos dados do primeiro pré-processamento, com imputação dos termos: 'ignorado', 9 e -1 para os dados faltantes;
b) **Banco II**: os dados não coletados foram convertidos para a média ou moda das variáveis após análise das mesmas; e
c) **Banco III**: as linhas com os dados: 'ignorado', 9 e -1 foram excluídas, aplicando apenas ao modelo os valores coletados das variáveis.

MAM para predição de sobrevida

Os MAM para análise de sobrevida tem por objetivo estabelecer a relação entre as variáveis e o momento de ocorrência do evento, e assim, aprender com estes dados para fornecer a previsão do evento de interesse. A relação entre o tempo de sobrevivência e a ocorrência do evento se dá pelas funções de sobrevida (retorna a probabilidade de tempo de sobrevida) e de risco (retorna a probabilidade de ocorrência do evento). Onde, a previsão da sobrevida é decorrente da avaliação da medida de correlação entre o risco previsto e observado no teste.[21,22,23]

O RSF fornece a previsão do paciente sobreviver após um determinado tempo de seguimento, através das funções de sobrevida Kaplan-Meier e de risco Nelson-Aalen. No qual, o cultivo das árvores incorporado às informações de censura é realizado pela regra de divisão aleatória long-rank, que visa dividir os nós com sobrevidas diferentes, e assim maximizar a diferença de sobrevida entre nós.[9,16,18,20]

Portanto, o RSF tem por objetivo prever a sobrevida de cada paciente determinando o quão bem o modelo generaliza a predição do tempo de sobrevida. Para se alcançar esta finalidade, o método não aplica apenas dois conjuntos de valores de entrada e saída; mas sim três conjuntos: um constituído pelas variáveis independentes (prognósticas), os outros formados por um vetor com o indicador do evento (censurado ou não censurado) e dos tempos de sobrevida (dias).[21,22]

Atributos e hiperparâmetros do RSF

Na construção do RSF, as variáveis prognósticas selecionadas foram: idade, tamanho do tumor, status menopausal, receptores estrogênio e progesterona, hormonioterapia, linfonodos positivos e grau do tumor, como proposto pela biblioteca; e inclusão da variável HER2 justificada pela construção do banco de dados.[5,21,22] Já as variáveis de valor: status (0-censurado ou 1-evento) e tempo de sobrevida (dias).[21,22]

Como o banco de dados original tem por finalidade analisar as taxas de sobrevida de 5 anos e os principais fatores associados ao perfil imuno-histoquímico.[5] E também, como na prática clínica os resultados dos receptores estrogênio, progesterona e HER2 resultam na subclassificação do CM e direcionam a condução clínico-terapêutica.[5,8,9] Assim, optou-se pela inserção da variável prognóstica HER2.

Já os hiperparâmetros utilizados no RSF, foram, respectivamente, o número de árvores na floresta de 1000, o número mínimo de amostras necessárias para dividir um nó e em um único nó, respectivamente, 10 e 15. Os dados foram divididos 75% para treinamento e 25% para testes, onde os critérios de divisão aplicados para a construção de cada árvore se baseiam no teste de log-rank.[21]

Métrica de Desempenho

As medidas de desempenho aplicadas aos MAM para análise de sobrevida tem a funcionalidade de avaliar o quão bem o método prever os diferentes tempos de sobrevida. Porém, como estes tempos estão sujeitos à censura, não se utilizam métricas usuais como: erro quadrático médio ou correlação.[21,22,23] Logo, para se avaliar o RSF foram selecionadas as métricas: índice de concordância (C-index)[24] e *Brier Score*[25], mais robustas e específicas para captar o comportamento analisado.[12,21]

O C-index quantifica a capacidade em discriminar os escores de risco dos diferentes tempos previstos no método pelos observados nos dados (teste), por intermédio de uma medida de concordância entre os pares concordantes pelos não descartados. Os pares descartados tem como característica: o menor tempo do evento (morte); pares com o mesmo tempo, a menos que o evento ocorra em um ou em ambos; e por fim, a não ocorrência do evento em ambos os elementos do par.[9,21,24]

Em resumo, o C-index quantifica o poder do método em predizer e classificar os tempos de morte dos pacientes. O intervalo do desempenho do método, assume valores de 0,0 até 1,0, com a seguinte interpretação: 0,5 para desempenho médio, sem discriminação preditiva; e 1,0 para desempenho perfeito, referente a um modelo capaz de separar os pacientes com diferentes desfechos. [9,17,23]

O *Brier Score* corresponde a uma medida similar ao erro quadrático médio, através de testes entre a precisão das probabilidades previstas nas funções de sobrevida, com o status observado nos dados para os momentos (T) selecionados do teste, ou seja, a calibração do modelo. Desta forma, avalia a qualidade de predição do método através da probabilidade do paciente permanecer livre do evento (morte), em

que valores mais baixos sugerem melhores resultados, representando de forma satisfatória a previsão individual do método para cada paciente.[21,25]

Resultados e Discussão

Os resultados apresentados são decorrentes da aplicação do RSF a um banco de dados clínico composto por uma população de 563 mulheres diagnosticadas com CM, limitando-se o tempo do seguimento por 5 anos (1825 dias), em consequência do banco original ter como proposta a análise de 5 anos, e assim, identificar a capacidade de sobrevivência neste período. A contagem de tempo de sobrevida de 5 anos se inicia com a data do laudo histopatológico (diagnóstico) e se finaliza com a data do evento adverso (óbito por CM) ou da censura (último dia de acompanhamento, ou limitado ao tempo proposto de análise). Portanto, após o pré-processamento apenas uma paciente foi excluída por não possuir a data do laudo histopatológico, restando 562 pacientes.[5,10]

Análise descritiva dos dados

O banco de dados original coletado de forma manual (prontuários), justificando assim os tratamentos e descrições detalhadas dos dados.[5] Deste modo, o Quadro I especifica as características das variáveis utilizadas pelo RSF.[5,18]

Quadro 1 – Descrição das variáveis clínicas aplicadas ao RSF [5]

Variáveis clínicas do banco de dados	Características das variáveis
Idade	Idade da paciente na primeira consulta oncológica (suspeita de diagnóstico)
Tamanho do tumor	Medida do tamanho do tumor (em cm)
Linfonodos comprometidos	Número de linfonodos comprometidos (após a cirurgia)
Receptor de Estrogênio	Resultado do exame imuno-histoquímico (em cruzes)
Receptor de Progesterona	Resultado do exame imuno-histoquímico (em cruzes)
Receptor Her2	Resultado do exame imuno-histoquímico (em cruzes)

Tempo de sobrevida 5 anos	O tempo de sobrevivência dos pacientes em dias (limitado a 1825 dias)
Status menopausal	Estado menopausal (pós-menopausa-1 ou pré-menopausa-0)
Hormonioterapia	Administrada a hormonioterapia (sim-1 ou não-0)
Grau histopatológico	Grau histopatológico tumoral (bem diferenciado-1 ou moderadamente diferenciado-2 ou pouco diferenciado-3)
Seguimento	Status do seguimento, vivas ou óbito por CM (censurado-0 ou não censurado-1)

A Tabela 1 mostra a análise descritiva dos dados de cada banco (I, II e III), já definidas anteriormente no Quadro 1; sumarizando e explorando o comportamento de cada variável.

Tabela 1 – Análise descritiva das variáveis utilizadas no RSF.

Variáveis	**Banco I**	**Banco II**	**Banco III**
Idade * (mínimo - máximo)	58.14 ± 13.67 (26 - 91)	58.14 ± 13.67 (26 - 91)	56.57 ± 12.90 (26 - 91)
Tamanho do tumor * (mínimo - máximo)	2.99 ± 2.91 (-1.0 - 25.0)	3.22 ± 2.75 (0.0 - 25.0)	3.20 ± 2.51 (0.4 - 15.0)
Linfonodos comprometidos* (mínimo - máximo)	2.39 ± 4.74 (-1 - 37)	2.53 ± 4.69 (0 - 37)	2.90 ± 4.95 (0 - 37)
Receptor Estrogênio * (mínimo - máximo)	2.89 ± 2.52 (0 - 9)	2.35 ± 2.06 (0 - 5)	2.38 ± 2.02 (0 - 5)
Receptor Progesterona * (mínimo - máximo)	2.59 ± 2.47 (0 - 9)	2.05 ± 1.93 (0 - 5)	2.16 ± 1.89 (0 - 5)
Receptor Her2 * (mínimo - máximo)	1.82 ± 3.02 (0 - 9)	0.83 ± 1.38 (0 - 5)	0.76 ± 1.41 (0 - 5)
Tempo de sobrevida* (mínimo - máximo)	1557.72 ± 505.60 (9 -1825)	1557.72 ± 505.60 (9 -1825)	1621.85 ± 429.74 (149 -1825)
Status menopausal **			
Pós-menopausa	375	375	215
Pré-menopausa	187	187	123
Hormonioterapia **			
Uso	157	157	86
Não uso	405	405	252
Grau histopatológico **			
1 (bem diferenciado)	127	127	110
2 (moderadamente)	184	346	160

3 (pouco diferenciado)	89	89	68
Seguimento **			
Censurado (vivas)	433	433	269
Não censurado (óbito)	129	129	69
Total **			
(população de pacientes)	562	562	338

* Média e desvio padrão
** Frequência

As análises descritivas das variáveis são apresentadas na Tabela 1, contendo nas variáveis numéricas: as médias, desvios padrões, valores mínimos e máximos; e nas categóricas as frequências numéricas de cada categoria. Desta forma, nota-se como as diferenças formas de pré-processamento dos dados em cada banco (I, II e III) influenciaram respectivamente sua descrição. Para exemplificar, observa-se: no Banco I e II o mesmo valor amostral e de frequência próximos, mas valores de média, desvio padrão, mínimos e máximos bem diferentes (exceto na idade e tempo de sobrevida 5 anos que foram coletados todos os valores); já entre os bancos II e III com valores mínimos e máximos iguais, exceto para variável tempo de sobrevida 5 anos, mas com valores amostrais, de frequências, média e desvio padrão diferentes.

Análise do desempenho do RSF

O desempenho do RSF apresentado por meio da análise das métricas C-index e *Brier Score*, usualmente aplicadas a métodos de análise de sobrevida, como demonstrados na Tabela 2.[12,21,22] Onde, destaca-se a realização de dois testes: teste 1 sem a inclusão da variável HER2 e teste 2 com a inclusão da variável HER2; aplicados em cada banco I, II e III, caracterizados e analisados descritivamente, respectivamente no Quadro I e Tabela I.

Tabela 2– Representação do desempenho do RSF.

Medidas de desempenho	**Banco I [a]**	**Banco II [b]**	**Banco III [c]**
C-index (sem HER2)	0,784295	0,790064	0,746695
C-index (com HER2)	0,792308	**0,801603**	0,739736

Brier Score (sem HER2)	0,125033	0,120964	0,156970
Brier Score (com HER2)	0,123146	0,118892	0,156214

[a] População de 562 pacientes diagnosticadas com CM.
[b] População de 562 pacientes diagnosticadas com CM.
[c] População de 338 pacientes diagnosticadas com CM.

De maneira geral, as medidas de desempenho do C-index visam verificar a precisão da previsão da sobrevida.[13] Os resultados obtidos nos bancos I, II e III são superiores aos alcançados de 0,68 pelo método proposto RSF por Ishwaran [17]. Agora, ao se comparar o valor de desempenho do C-index entre os testes, ou seja, a inclusão ou não da variável HER2 e entre os bancos I, II e III, observa-se:

- Nos bancos I e II, a inclusão da variável HER2 melhora-se sutilmente a eficácia de predição do RSF, com um aumento da precisão (C_{dist} = C_{indexH} - C_{index}): no banco I, em *0,792308 – 0,784295 = 0,008013*; e no banco II, em *0,801603 – 0,790064 = 0,011539*, esta diferença pode ser atribuída ao pré-processamento.
- No banco III se inverte esta classificação, onde a inclusão obteve uma pequena diminuição da precisão, em *0,739736 - 0,746695 = -0,006959*, não melhorando o desempenho do RSF, que pode ter sido influenciada pela redução da quantidade amostral.

Já as medidas de desempenho do *Brier Score* apontam para uma boa calibração e predição do RSF, com métrica inferior a 0,25 em todos os bancos e testes.[12,25] Os resultados desta métrica no banco III são um pouco maiores que dos outros bancos, e oposto entre os testes (com ou sem HER2), mas ainda na faixa de efetividade do desempenho do RSF que pode ter sido influenciada pelo número amostral reduzido.

As medidas de desempenho do C-index obtidas em todos os testes e bancos deste trabalho, são superiores quando comparados aos resultados de outros trabalhos da literatura que também aplicaram o RSF a um banco de dados de mulheres diagnosticadas com CM.[9,16,17] Porém, no estudo de Aivaliotis et al.[16] o valor de C-index foi de aproximadamente 0,57, entretanto o objetivo não era associar a sobrevida, mas sim a incidência do CM ao Índice de Massa Corporal, comparando o clássico modelo *Cox Proportional Hazards* (CPH) ao RSF.

O estudo de Moncada-Torres et al.[9], usou dados de pacientes com CM não metastático para comparar o desempenho do RSF a outros métodos como CPH, *Survival Support Vector Machines* (SSVM) e *Extreme Gradient Boosting* (XGB), demonstrando que os métodos RSF e SSVM apresentam um C-index de 0,63 enquanto o XGB 0,73, valores estes menores que os obtidos neste trabalho (Tabela 2).

O trabalho de Pinheiro et al.[17] também verifica os resultados dos modelos CPH e RSF aplicados a um conjunto de dados de CM, resultando em um C-index de 0,68, menor que os alcançados neste trabalho (Tabela 2). Vale destacar, que os dados das pacientes do presente trabalho constitui-se: todos os tipos histológicos de CM e também as metastáticas.

Já os resultados do trabalho Xiao et al.[12] também comparando-se os modelos CPH, SSVM e RSF, verificando-se que o RSF superou outros modelos em capacidade discriminativa com um C-index de 0,85 e um Brier Score de 0,045, resultados estes melhores que demonstrados na Tabela 2. Vale enfatizar que o estudo[12], constituído por uma população de mais 22 mil pacientes e 11 atributos de dados prognósticos aplicados ao RSF, reflete a importância de bancos robustos de alta qualidade para serem aplicados aos MAM.[9,10,13] Representando uma das limitações encontradas na aplicação destes métodos, cujo objetivo é obter uma previsão satisfatória do tempo de sobrevida para CM.[10,11,12,13]

Conclusão

Os MAM tornaram-se uma metodologia inovadora para previsão da sobrevida, mesmo que existam ainda melhorias e potencial para se agregar modelos computacionais adicionais. Os resultados provenientes deste trabalho, mostraram que o RSF possibilita análises de sobrevida promissoras para o CM, principalmente quando validadas por um banco de dados clínico robusto.

Afinal, nota-se que as pesquisas multidisciplinares, com construção de bancos de dados robustos e de alta qualidade, baseadas em descobertas anteriores e orientadas pelos especialistas da área médica, podem resultar em métodos eficazes para serem aplicados na prática clínica.

Agradecimentos

Agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro por bolsa concedida; ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora pelo aprendizado; ao professor Dr. Maximiliano Ribeiro Guerra e à mastologista Dra. Jane Rocha Duarte Cintra pela colaboração e incentivo disponibilizando o banco de dados clínicos; e aos Hospitais 9 de Julho e Instituto Oncológico apoio nesta pesquisa.

Referências

1 Ferlay J et al. Cancer statistics for the year 2020: An overview. International Journal of Cancer. 2021;149(4),p.778-89.

2 World Health Organization. Cancer [Internet]; c2022 [cited 2022 Set 12]. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer

3 Instituto Nacional de Câncer. Estatísticas de câncer [Internet]; c2022 [cited 2022 Set 10]. Available from: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/

4 Carvalho DS, Guerra MR, Barra LP, Queiroz RA. Aspectos gerais epidemiológicos da mortalidade por câncer de mama feminino no brasil e no mundo. Anais Simpósio de Enfermagem [Internet]. 2019 [cited 2022 Out 27];3:[about 1 p.]. Available from: http://pensaracademico.facig.edu.br/index.php/simposioenfermagem/article/view/1116

5 Cintra JR. Sobrevida e fatores associados em pacientes com câncer de mama, com diagnóstico entre 2003 e 2005 no município de Juiz de Fora – MG. [dissertation]. Juiz de Fora (JF): Universidade Federal de Juiz de Fora, 2012.

6 Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. Cancer Epidemiol Biomarkers Prev. 2017;26(4):444-457.

7 Carvalho DS, Guerra MR, Barra LP, Queiroz RA. Modelagem computacional do crescimento tumoral mamário. Anais Seminário Científico UNIFACIG [Internet]. 2017 [cited 2022 Out 27];3:[about 1 p.]. Available from: http://pensaracademico.facig.edu.br/index.php/semiariocientifico/article/view/438

8 Ministério da Saúde (BR). Secretaria de Atenção à Saúde. Protocolos clínicos e diretrizes terapêuticas em Oncologia/Ministério da Saúde, Secretaria de Atenção à Saúde – Brasília : Ministério da Saúde, 2014.

9 Moncada-Torres A et al. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports. 2021;11(1):p.1-13.

10 Li J et al. Predicting breast cancer 5-year survival using machine learning: a systematic review. PloS one [Internet]. 2021 [cited 2022 Out 27];16(4):[about 1 p.]. Available from: https://doi.org/10.1371/journal.pone.0250370

11 Tapak L et al. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health. 2019;7(3):p.293-9.

12 Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, He Y, Zheng Y. The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study. JMIR medical informatics [Internet]. 2022[cited 2022 Out 27]; 10(2):[about 1 p.]. Available from: https://medinform.jmir.org/2022/2/e33440

13 Hueman MT et al. Creating prognostic systems for cancer patients: A demonstration using breast cancer. Cancer medicine. 2018;7(8):p.3611-21.

14 Lai X et al. Toward Personalized Computer Simulation of Breast Cancer Treatment: A Multiscale Pharmacokinetic and Pharmacodynamic Model Informed by Multitype Patient Data. Cancer research. 2019;79(16):p.4293-304.

15 Nave O. Adding features from the mathematical model of breast cancer to predict the tumour size. International Journal of Computer Mathematics: Computer Systems Theory. 2020;5(3):p.159-174.

16 Aivaliotis G et al. A comparison of time to event analysis methods, using weight status and breast cancer as a case study. Scientific reports. 2021;11(1):p. 1-9.

17 Pinheiro TS et al. Machine Learning e Análise Multivariada aplicados à Sobrevida do Câncer Mama. J Health Inform [Internet]. 2022[cited 2022 Out 27];(14).[about 1 p.]. Available from: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/971

18 Ishwaran H et al. Random survival forests. The annals of applied statistics. 2008;2(3):p.841-60.

19 Breiman L. Random forests. Machine Learning, Springer Science and Business Media LLC. 2001;45(1):p.5–32.

20 Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Statistics in medicine. 2019;38(4):p.558-82.

21 Understanding Predictions in Survival Analysis. [Internet];[cited 2022 Set 12]. Available from: https://scikit-survival.readthedocs.io/en/stable/user_guide/

22 Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. J. Mach. Learn. Res. 2020;21(212):p.1-6.

23 Fast Unified Random Forests with random. [Internet];[cited 2022 Set 12]. Available from: https://www.randomforestsrc.org/articles/survival.html

24 Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On The C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics In medicine. 2011;30(10):1105-17.

25 Graf E, Schmoor C, Sauerbrei W, Schumacher M.Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999;18(17-18):2529-45.