# Triagem automatizada de pacientes com risco de Câncer de Mama*

Beatriz Assumpção Tergolino[1 [0000−0001−9294−8883]], Fiama Jimenez Ochoa[1 [0000−0002−1793−7516]], Michele Garcia Muraro[2], and Sandro da Silva Camargo[1 [0000−0001−8871−3950]]

[1] Programa de Pós-Graduação em Computação Aplicada, Universidade Federal do Pampa, Bagé, RS, Brasil
http://cursos.unipampa.edu.br/cursos/ppgcap/
{sandrocamargo,beatriztergolino.aluno,fiamaochoa.aluno}@unipampa.edu.br
[2] Santa Casa de Porto Alegre, Porto Alegre, RS, Brasil
michelemuraro.mm@gmail.com

**Abstract.** O câncer de mama é a segunda causa mais frequente de mortes por câncer entre as mulheres, atrás do câncer de pele. Pra reduzir a mortalidade, podem ser aplicados testes de triagem, em pacientes assintomáticos, a fim de permitir um diagnóstico precoce e antecipação do início do tratamento. Este trabalho visa construir modelos preditivos baseados em dados de exames de sangue de rotina para fazer triagem de pacientes com indicativo de câncer de mama. Os modelos aqui apresentados atingiram uma acurácia de 77%, com uma sensibilidade, para predizer casos positivos, de 83%.

**Keywords:** Mineração de Dados · Classificação · Exame de Sangue · Oncologia.

## 1 Introdução

Segundo estatísticas recentes, para o ano de 2020, o Instituto Nacional de Câncer estimou 625 mil novos casos de câncer no Brasil, sendo em torno 309 mil em homens e 316 mil em mulheres. Entre os tipos de neoplasias mais frequentes, o câncer de pele não melanoma ocupou a primeira posição com 177 mil novos casos, seguido pelos câncer de mama feminina e de próstata, ambos com estimativa de 66 mil novos casos[1].

Testes de câncer em pessoas que não apresentam sintomas é um processo chamado triagem. Estratégias de triagem têm provado sua capacidade de detectar a presença de câncer em seus estágios iniciais, permitindo um diagnóstico precoce e o início antecipado do tratamento, sendo um fator crítico para reduzir a taxa de mortalidade [2]. Dados oficiais indicam que é evidente a diferença entre as magnitudes de incidência e mortalidade, demonstrando a importância do

* Este trabalho foi desenvolvido com apoio financeiro da CAPES/FAPERGS (PDPG), Edital nº 18/2020.

acesso aos serviços de saúde para detecção precoce, diagnóstico e tratamento oportuno, em consonância com as diretrizes para a detecção precoce do câncer de mama no Brasil[1].

Neste contexto, a grande quantidade de fatores envolvidos e a complexidade inerente aos processos biológicos dificulta a plena compreensão dos mecanismos envolvidos no câncer. Por outro lado, nos dias atuais, um dos grandes impulsos para o desenvolvimento da computação é o projeto de modelos inteligentes com a finalidade de tratar problemas de análise de dados incrementalmente complexos[3]. Como consequência deste desenvolvimento, a literatura tem relatado muitos casos de sucesso através da modelagem de inteligência biológica e natural, com base em dados, nas mais variadas áreas de conhecimento, resultando no que se convencionou chamar de "Aprendizado de Máquina"[4]. Assim, sistemas inteligentes que possam aprender com base em dados coletados em consultas de rotina e exames de sangue podem fornecer uma importante contribuição na forma de novos recursos de triagem[5].

Neste contexto, o objetivo deste trabalho é construir um sistema de triagem inteligente, criado a partir da aplciação de técnicas de aprendizado de máquina sobre dados de exames de sangue de rotina, coletados em um trabalho anterior[5]. O resultado é um modelo de árvores de decisão para triagem, além da avaliação da capacidade preditiva deste modelo.

Este artigo está organizado da seguinte forma: na Seção 2, é exposta a metodologia proposta para o desenvolvimento do trabalho. Na Seção 3, são apresentados e discutidos os resultados obtidos no presente estudo. Finalmente, na Seção 4, são apresentadas as conclusões e trabalhos futuros.

## 2 Material e Métodos

### 2.1 Base de Dados

Para este estudo, foram coletados dados de 116 pacientes, catalogados em um estudo anterior [5]. As mulheres diagnosticadas com Cancer de Mama foram recrutadas pelo Departamento de Ginecologia do Centro Hospitalar e Universitário de Coimbra (CHUC) entre 2009 e 2013. Dentre os dados coletados, estão 10 preditores quantitativos e uma variável dependente binária, que indica a presença ou ausência de Cancer de Mama. Estes preditores são dados antropométricos e parâmetros que podem ser coletados em exames de sangue de rotina [6]. As amostras do estudo incluíam 64 mulheres com cancer de mama e 52 voluntários saudáveis. Para cada amostra, foram coletados dados de idade, Índice de Massa Corporal (IMC), Glicose, Insulina, modelo de Avaliação da HOMeostase da resistência à insulina (HOMA), Leptina, Adiponectina, Resistina e Proteína Quimiotática de Monócitos 1 (MCP-1).

### 2.2 Inteligência Computacional e Árvores de Decisão

A Inteligência Computacional (IC) é um ramo da computação que trata da automatização do comportamento inteligente. Atualmente, a área de IC engloba

uma ampla variedade de sub-campos, dentre eles, uma das mais férteis áreas de pesquisa é a que se preocupa com a construção de sistemas de alto desempenho capazes de aprender através da experiência e obter conhecimento a partir de dados [4].

Há quatro diferentes classes de aprendizado de máquina: aprendizado supervisionado, aprendizado não supervisionado, aprendizado por reforço e aprendizado por programação em lógica indutiva [7]. No presente trabalho, foram aplicadas as técnicas de aprendizado supervisionado, que são uma classe de algoritmos que visam aprender uma função arbitrária que associa os dados de entrada e os dados de saída previamente conhecidos. Esta associação geralmente descreve uma função $f_o(x)$ presente de forma implícita em um conjunto de treinamento $D = [x(i), y(i)] \in \Re x \Re, i = 1, ..., l$ consistindo de $l$ pares $(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)$. As entradas $x$ consistem em um vetor n-dimensional onde $x \in \Re$, e as saídas $y$ consistem em um vetor 1-dimensional onde $y \in \Re$. Dependendo do valor a ser predito o aprendizado supervisionado pode ser de dois tipos: regressão quando os valores de saída são contínuos, e classificação quando os valores de saída são discretos [8]. Durante o processo de treinamento, as amostras são sucessivamente submetidas ao algoritmo de aprendizado. Para cada amostra de entrada, o algoritmo tenta predizer a saída. A saída predita pelo algoritmo é comparada com a saída real, a diferença entre elas é utilizada para reajustar os parâmetros do modelo. Desta forma, o algoritmo iterativamente ajusta seus parâmetros para criar um modelo que faça um mapeamento das entradas para a saída. Após o processo de aprendizado supervisionado, é criado um modelo que pode ser utilizado para simular o conhecimento do especialista do domínio.

Neste trabalho, foram aplicadas técnicas de classificação e foi priorizada a construção de modelos do tipo caixa branca, onde se enquadram os algoritmos de geração de regras ou árvores de decisão. Apesar destes modelos serem lineares e, tipicamente, terem uma capacidade preditiva menor que modelos não lineares, sua maior vantagem é a geração de modelos interpretáveis [4]. Na fase de avaliação dos modelos foi utilizada a técnica *holdout* para permitir a correta mensuração da capacidade preditiva dos modelos.

O algoritmo de geração de árvores utilizado foi o *J48*, que classifica instâncias ordenando-as da raiz da árvore em direção a suas folhas [9]. Os preditores mais relevantes posicionam-se mais perto da raiz da árvore. À medida que a relevância do preditor diminui, ele é posicionado mais longe da raiz e mais próximo às folhas.

### 2.3 Ferramentas

Para realização dos experimentos, foi utilizado um computador com processador Intel®Core™i5-2520M, de 2.5Ghz, com 5.7Gb de memória RAM, e disco de 500.1 Gb. Foram utilizados o ambiente *R-Studio* versão 1.2.5042, e a linguagem R versão 3.6.3 sobre o sistema operacional *linux* em plataforma x86 de 64 bits, com distribuição *Ubuntu* 20.04.2 LTS e ambiente gráfico GNOME versão 3.36.8. Foram também utilizados os pacotes *Classification and Regression Training* (Caret) versão 6.0-86, *Rpart.plot* versão 3.0.9.

## 3 Resultados e Discussão

Os resultados aqui apresentados são discutidos em duas perspectivas: uma análise estatística preliminar seguida pela criação dos modelos de árvores de decisão.

### 3.1 Estatística Descritiva

A Tabela 1 apresenta as estatísticas descritivas dos dados utilizados neste estudo. Esta tabela apresenta, para cada preditor, o valor mínimo (Min), primeiro quartil ($Q_1$), mediana, média, terceiro quartil ($Q_3$), valor máximo (Max), desvio padrão (DP), coeficiente de variação (CV), p-valor do teste de normalidade de Shapiro Wilk (Norm) e o p-valor do teste não paramétrico de médias de Wilcoxon (Dif) para diferenças estatisticamente significativas entre pacientes dos dois grupos: com presença ou com ausência de câncer de mama.

**Table 1.** Estatística descritiva dos preditores quantitativos.

| Preditor | Min | $Q_1$ | Mediana | Média | $Q_3$ | Max | DP | CV | Norm | Dif |
|---|---|---|---|---|---|---|---|---|---|---|
| Idade | 24 | 45 | 56 | 57.302 | 71 | 89 | 16.113 | 0.281 | 0.009 | 0.479 |
| IMC | 18.37 | 22.973 | 27.662 | 27.582 | 31.241 | 38.579 | 5.02 | 0.182 | 0.008 | 0.202 |
| Glicose | 60 | 85.75 | 92 | 97.793 | 102 | 201 | 22.525 | 0.23 | 0 | **0** |
| Insulina | 2.432 | 4.359 | 5.925 | 10.012 | 11.189 | 58.46 | 10.068 | 1.006 | 0 | **0.027** |
| HOMA | 0.467 | 0.918 | 1.381 | 2.695 | 2.858 | 25.05 | 3.642 | 1.351 | 0 | **0.003** |
| Leptina | 4.311 | 12.314 | 20.271 | 26.615 | 37.378 | 90.28 | 19.183 | 0.721 | 0 | 0.949 |
| Adipon. | 1.656 | 5.474 | 8.353 | 10.181 | 11.816 | 38.04 | 6.843 | 0.672 | 0 | 0.766 |
| Resistina | 3.21 | 6.882 | 10.828 | 14.726 | 17.755 | 82.1 | 12.391 | 0.841 | 0 | **0.002** |
| MCP.1 | 45.843 | 269.978 | 471.322 | 534.647 | 700.085 | 1698.44 | 345.913 | 0.647 | 0 | 0.504 |

Os resultados mostram que nenhum dos preditores apresenta uma distribuição normal dado que os p-valores são menores que 0.05 ($p < 0.05$). Em consequência deste resultado, o teste não paramétrico de Wilcoxon foi usado para identificar se os grupos tem diferenças estatisticamente significativas entre suas distribuições, com nível de significância de 5%, ou $p < 0.05$. Foram apontadas diferenças estatisticamente significativas para os dois grupos em quatro preditores: Glicose, Insulina, HOMA e Resistina, que estão com seus p-valores enfatizados em negrito. A Figura 1 apresenta as distribuições de densidade destes quatro preditores em relação aos diferentes grupos.

### 3.2 Modelos de Árvores de Decisão

As 116 amostras disponíveis foram divididas, de maneira estratificada de acordo com sua classe (ausente ou presente), na proporção de 80% dos dados para treino (com 42 amostras de ausente e 52 amostras de presente) e 20% dos dados para teste (com 10 amostras de ausente e 12 amostras de presente).

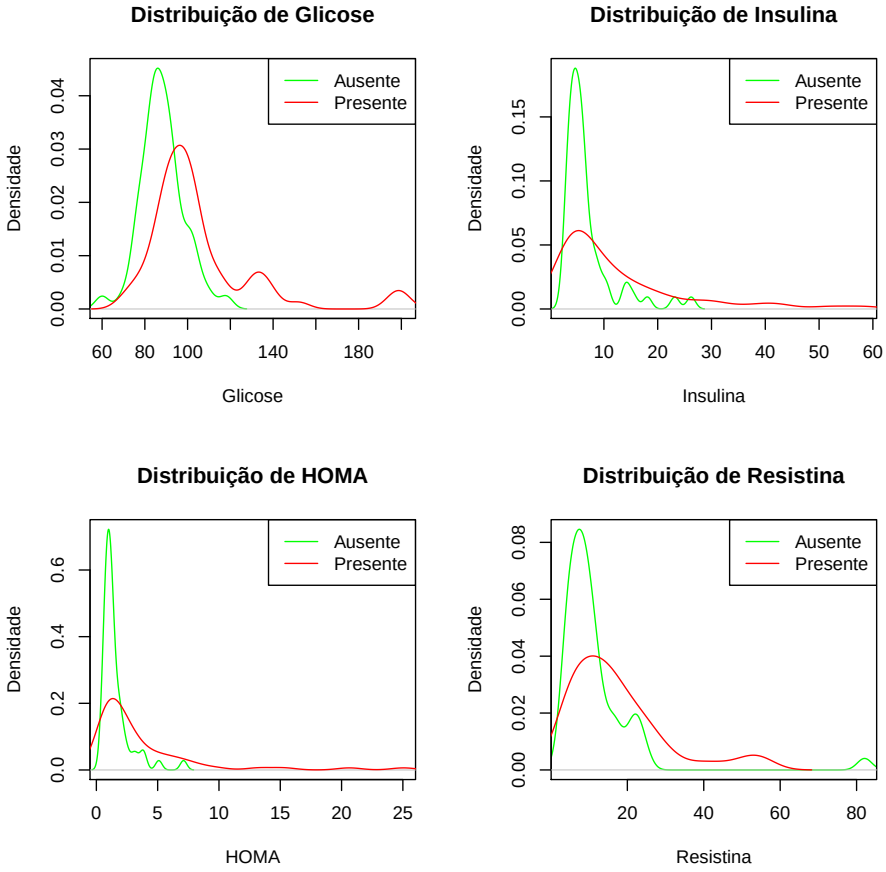


**Fig. 1.** Densidades por grupo dos quatro preditores que apresentaram diferenças estatisticamente significantes em suas distribuições.

Sobre os dados de treino, foi aplicado o algoritmo de classificação J48. O objetivo deste experimento foi criar modelos preditivos para o diagnóstico de câncer de mama. Foi encontrado um padrão de comportamento dos preditores que influenciam tal diagnóstico. O modelo foi aplicado aos dados de teste e foi avaliado a fim de inferir a sua capacidade preditiva. A Figura 2 apresenta o modelo de árvores de decisão criado pelo algoritmo J48. A árvore de decisão criada está baseada nos preditores Glicose, Resistina e Idade.

Na Figura 3, é apresentada a superfície de decisão do ramo esquerdo da árvore, que considera Glicose e Resistina, e como as amostras de teste deveriam ter sido classificadas, de acordo com o sombreamento das áreas, sendo verde para ausente e vermelho para presente. Conforme o modelo da Figura 2, quando o índice de Glicose da paciente é menor que 92 e o nível de Resistina é menor

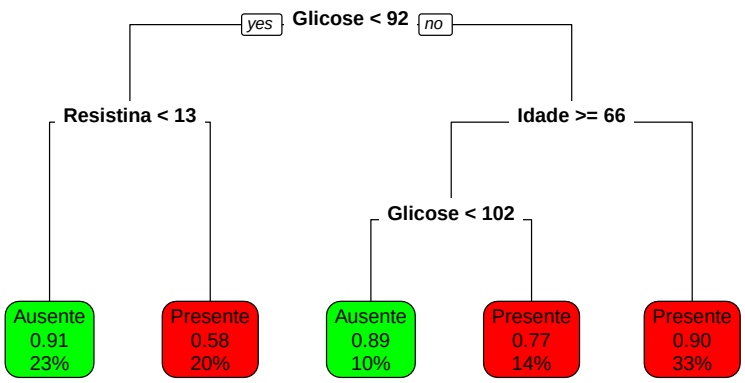


**Fig. 2.** Árvore de Decisão.

que 13, a decisão a ser tomada é que o paciente não teria indícios de Câncer de Mama. Esta decisão tem taxa de acerto de 0.91, ou 91%, e 23% dos dados de treino atendem os valores limite para se enquadrarem nesta regra. A regra foi representada na área sombreada em verde na Figura, e mostra que houve 1 erro em 8 amostras de teste. Na outra regra deste ramo, quando a Glicose é menor que 92 e a Resistina não é menor que 13, a decisão a ser tomada é que o paciente teria indícios de Câncer de Mama. Esta decisão tem taxa de acerto de 0.58, ou 58%, e 20% dos dados de treino atendem os valores limite para se enquadrarem nesta regra. A regra foi representada na área sombreada em vermelho na Figura e mostra que houve 1 erro em 2 amostras de teste. É importante salientar que esta é a regra com menor taxa de acerto (0.58) dentre todas as regras do modelo.

Complementarmente, na Figura 4, são apresentadas a superfície de decisão do ramo direito da árvore, que considera Glicose e Idade, e como as amostras de teste deveriam ter sido classificadas. Conforme o modelo da Figura 2, quando o índice de Glicose da paciente não é menor que 92 e a Idade não é maior que 66, a decisão a ser tomada é que o paciente teria Câncer de Mama. Esta decisão tem taxa de acerto de 0.90, ou 90%, e 33% dos dados de treino atendem os valores limite para se enquadrarem nesta regra. A regra foi representada na área sombreada em verde na Figura, e mostra que houve 1 erro em 8 amostras de teste. Por outro lado, quando a Glicose não é menor que 92 e a Idade é maior que 66, novamente a Glicose deve ser analisada e, caso ela não seja menor que 102, a decisão a ser tomada é que o paciente haveria indícios de Câncer de Mama. Esta decisão tem taxa de acerto de 0.77, ou 77%, e 14% dos dados de treino atendem os valores limite para se enquadrarem nesta regra. Finalmente, quando a Glicose está entre 92 e 102, e a Idade é maior que 66, o modelo indica que o diagnóstico seria negativo, sendo que esta regra teria taxa de 0.89 de acerto, definida com base na análise de 10% dos dados de treino. Nos dados de teste, há 1 erro para 2 amostras.

A matriz de confusão apresentada na Tabela 2 sumariza as classificações realizadas nos dados de teste. Pode ser evidenciado que foram realizadas, nos

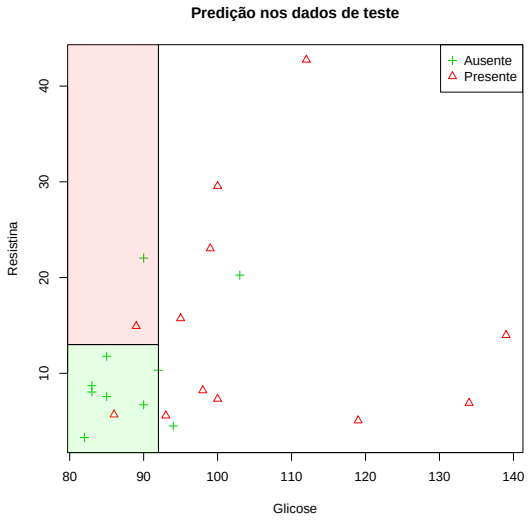


**Fig. 3.** Superfície de Decisão do ramo esquerdo da árvore para pacientes com glicose menor que 92.

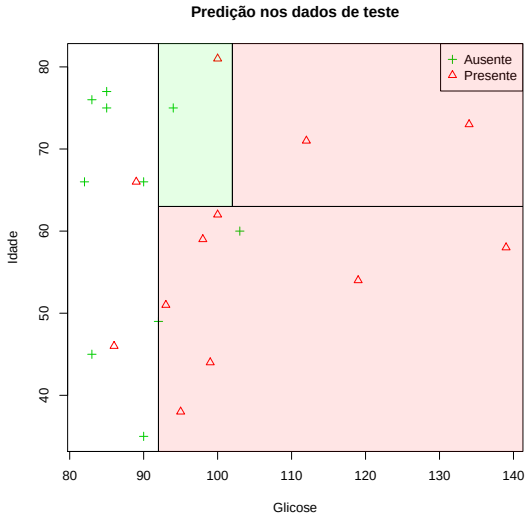


**Fig. 4.** Superfície de Decisão do ramo direito da árvore para pacientes com glicose maior ou igual a 92.

dados de teste, 17 classificações corretas, sendo 7 para a classe ausente e 10 para

a classe presente. Além disso, foram feitas 5 predições incorretas, sendo 3 para a classe Presente e 2 para a classe Ausente.

**Table 2.** Matriz de Confusão

| | Referência | |
|---|---|---|
| Predição | Ausente | Presente |
| Ausente | 7 | 2 |
| Presente | 3 | 10 |

**Table 3.** Métricas do Modelo

| Métrica | Valor |
|---|---|
| Acurácia | 0.7727 |
| 95% CI | (0.5463, 0.9218) |
| Taxa sem Informação | 0.5455 |
| P-Valor [Acc > NIR] | 0.02455 |
| Kappa | 0.5378 |
| P-Valor Teste Mcnemar's | 1.00000 |
| Sensibilidade | 0.8333 |
| Especificidade | 0.7000 |
| Valor Preditivo Positivo | 0.7692 |
| Valor Preditivo Negativo | 0.7778 |
| Prevalência | 0.5455 |
| Taxa de Detecção | 0.4545 |
| Detecção de Prevalência | 0.5909 |
| Acurácia Balanceada | 0.7667 |
| Classe 'Positiva' | Presente |

A Tabela 3 apresenta as métricas de avaliação do modelo sobre os dados de teste. Os resultados mostram que o modelo atingiu uma acurária de 77.27%, calculado a partir de dezessete classificações corretas em 22 amostras de teste. Com intervalo de confiança de 95%, estima-se que o modelo pode atingir acurária de 54,63% a 92,18%. O aumento da quantidade de amostras de teste poderia contribuir para uma redução da amplitude desta acurácia [10]. A taxa sem informação de 54.55% indica a proporção de amostras da maior classe, o que seria o baseline caso o modelo tentasse classificar todas as amostras como a classe mais numerosa. O p-valor deste modelo é inferior a 0.05, indicando que os resultados obtidos com as 22 amostras de teste também devem se repetir em populações maiores em mais de 95% das vezes. O coeficiente de Kappa é de 0.5378, indicando um nível de concordância moderado das decisões do modelo em relação ao diagnóstico real. A sensibilidade indica que 83,33% dos pacientes positivos foram identificados pelo modelo, calculado a partir de dez classificações corretas em doze amostras de teste positivas. A especificidade indica que 70% dos

pacientes negativos foram identificados pelo modelo, calculado a partir sete classificações corretas em dez amostras de teste negativas. O valor preditivo positivo de 76,92% indica a probabilidade de que amostras preditas como positivas pelo modelo realmente tenham a doença, e foi calculada a partir de dez predições corretas em treze predições positivas. O valor preditivo negativo de 77,78% indica sete predições corretas em nove amostras apontadas como negativas. A taxa de detecção de 45,45% indica quantas amostras positivas foram identificadas pelo modelo, sendo calculada razão entre as dez amostras positivas detectadas e as 22 amostras da base de teste. A detecção de prevalência de 59,09% é calculada como a taxa entre as treze amostras preditas como positivas e as 22 amostras da base de testes. A acurácia balanceada de 76,67%, usada em dados com quantidades de amostras diferentes em cada classe, é dada pela média aritmética da Sensibilidade e da Especificidade.

## 4 Conclusões

Este trabalho utilizou uma base de dados pública de exames de sangue de rotina, coletados pelo Departamento de Ginecologia do Centro Hospitalar e Universitário de Coimbra (CHUC) entre 2009 e 2013 [5]. Usando estes dados, a partir de análises estatísticas, foi identificado que concentrações de Glicose, Insulina, HOMA e Resistina diferem significativamente entre pacientes com e sem câncer de mama, corroborando com os resultados encontrados em um trabalho anterior. Além disso, o modelo de classificação criado neste trabalho mostrou que Glicose, Resistina e Idade também podem ser usados como fatores de triagem para identificação de potenciais pacientes a desenvolverem câncer de mama. Os resultados mostraram que o modelo atingiu uma acurácia de 77,27%, com uma sensibilidade de 83,33% em dados de teste. Tais resultados mostram o potencial do uso do modelo aqui proposto para triagem de pacientes com câncer de mama, a partir da dados de exame de sangue de rotina.

Como limitações deste trabalho, pode ser citado que todas as amostras utilizados são provenientes de um mesmo local, com pacientes inseridos em um mesmo contexto ambiental. A coleta de dados similares em locais e contextos diferentes poderia contribuir para aumentar a variabilidade e viabilizar a criação de modelos mais robustos e com maior capacidade preditiva. Outra limitação importante é o baixo número de amostras utilizadas para criação dos modelos. A utilização de bases de dados com mais amostras poderia aumentar a robustez do trabalho por diminuir o Intervalo de Confiança dos resultados, que foi estimado entre 54% e 92%, com as 116 amostras disponíveis. Da mesma forma, uma maior quantidade de amostras teria impacto no aumento do coeficiente de Kappa, que foi de 0.5378.

Dentre os trabalhos futuros, está planejada a utilização de modelos não lineares de redes neurais artificiais para investigação de sua capacidade preditiva visando a otimização não só da acurácia, mas principalmente da sensibilidade.

## References


1. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Coordenação de Prevenção e Vigilância Estimativa 2020 : Incidência de Câncer no Brasil. Rio de Janeiro. http://www2.inca.gov.br. Acessado em 16 Mai 2021.
2. National Institute of Health. National Cancer Institute. www.cancer.gov/types/breast. Acessado em 15 Mai 2021.
3. Engelbrecht, A. P. : Computational Intelligence: An introduction. 2. ed. John Wiley & Sons, Chichester (2007)
4. Camargo, S. S. ; Azambuja, R. C. C. ; Feijó, J. O. ; Corrêa, M. N. ; Schneider, A. ; Cardoso, F. F. : Modelagem Computacional de Indicadores Metabólicos para Estudo de Eficiência Reprodutiva em Vacas de Corte. In: Anais Eletrônicos do X Congresso Brasileiro de Agroinformática, pp. 857–866. Sociedade Brasileira de Agroinformática, Ponta Grossa-PR, (2015)
5. Patrício, M., Pereira, J., Crisóstomo, J. et al. : Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer **18**(29) (2018) https://doi.org/10.1186/s12885-017-3877-1
6. Breast Cancer Coimbra Data Set, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra. Acessado em 13 Mai 2021
7. Konar, A. : Artificial Intelligence and Soft Computing: behavioral and cognitive modeling of the human brain. CRC Press, Boca Raton (2000)
8. Kecman, V. : Learning and Soft Computing: support vector machines, neural networks, and fuzzy logic models. MIT Press, Cambridge (2001)
9. Quinlan, J. R. : C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, New York (1993)
10. Panagiotakos, D. B. "Value of p-value in biomedical research." The open cardiovascular medicine journal vol. 2 (2008): 97-9. doi:10.2174/1874192400802010097