Uso de Inteligência Artificial para a Identificação de Fatores Influenciadores do Câncer de Mama a partir de Dados Clínicos

Marina R. de Rezende[1], Tamiris S. Maimone[1], Mario O. de Menezes[1]

[1]Universidade Presbiteriana Mackenzie (UPM)
Rua da Consolação, 930 Consolação – 01.302-907 – São Paulo – SP – Brazil

31911838@mackenzista.com.br, 31901174@mackenzista.com.br,
mario.menezes@mackenzie.br

Abstract. *This paper seeks to understand how Artificial Intelligence, more specifically Machine Learning, is being used in the oncology area, mainly in the study of breast cancer. Using the Decision Tree algorithm, clinical variables were analyzed in order to identify their possible relationships and influences in relation to the number of deaths identified in a database. It was observed that the factors HR, ER, PgR and Her2 actually have an association when the tumor is present. In view of the results, it was noticed that more specific studies are needed, focused on the variables ER and PgR negative, HR and PgR positive, ER and Her2 negative.*

Keywords: *Artificial Intelligence, Machine Learning, Breast Cancer, Decision Tree.*

Resumo. *Este artigo busca entender como a Inteligência Artificial, mais especificamente o Machine Learning (Aprendizado de Máquina), está sendo utilizada na área oncológica, principalmente no estudo do câncer de mama. Através da utilização do algoritmo Árvore de Decisão, variáveis clínicas foram analisadas, a fim de identificar suas possíveis relações e influências em relação ao número de óbitos identificados em uma base de dados. Foi observado que os fatores HR, ER, PgR e Her2 de fato possuem uma associação quando há a presença do tumor. Diante dos resultados, percebeu-se que é necessário estudos mais específicos e focados nas variáveis ER e PgR negativos, HR e PgR positivos, ER e Her2 negativo.*

Palavras-chave: *Inteligência Artificial, Aprendizado de Máquina, Câncer de Mama, Árvore de Decisão.*

1. Introdução

A Inteligência Artificial (IA) na medicina tem sido muito requisitada, pois ela ajuda na detecção de doenças, prognósticos e tomadas de decisão clínica [Houssami et al. 2019]. Este projeto foca na aplicação da técnica de aprendizagem de máquina (um subconjunto da IA que permite que o computador aprenda através de um conjunto de dados que são treinados e testados [Ludermir 2021]) conhecida como Árvore de Decisão [Géron 2017] na área oncológica, a fim de identificar quais fatores influenciam em um maior número

de óbitos relacionados ao câncer de mama e quais as relações entre eles, permitindo, desta forma, que médicos possam utilizar esse conhecimento como suporte em diagnósticos.

O câncer de mama é uma doença causada pela multiplicação desordenada de células anormais da mama, que forma um tumor com potencial de invadir outros órgãos [INCA 2021]. Esta é uma doença que atinge mais as mulheres, respondendo por 22% de novos casos a cada ano [Lopes 2014] e que possui como fatores de risco a idade avançada, exposição prolongada aos hormônios, o excesso de peso e a história familiar ou de mutação genética [CRUK 2020]. Além disso, esta doença é uma das quatro principais causas de morte prematura (antes dos 70 anos de idade) na maioria dos países [INCA 2020]. Portanto, quanto mais precoce sua detecção, melhor. Dessa forma, os tratamentos são menos agressivos e os diagnósticos mais satisfatórios e precisos.

Mamas são glândulas que possuem como função principal a produção do leite, que se forma nos lóbulos e é encaminhado até os mamilos por pequenos canais chamados ductos. Quando as células da mama começam a se dividir de forma desordenada, um tumor maligno pode, normalmente, se instalar nos ductos, porém, em casos mais raros, pode acontecer de se alojar nos lóbulos [Souza 2013].

Em 2020, no Brasil, foram estimados 625.370 casos relacionados com algum tipo de câncer, incluindo homens e mulheres [INCA 2019]. Além disso, houve um aumento do índice de câncer, onde a estimativa para cada ano do triênio 2020-2022 aponta que ocorrerão 625 mil casos novos de câncer [INCA 2020]. A partir da necessidade de melhor detecção desses casos, é necessário também identificar o grau de extensão da doença o mais rápido possível, obtendo informações sobre o comportamento biológico do tumor, a previsão de seu provável desenvolvimento, a avaliação dos resultados do tratamento, entre outros [INCA 2011].

Nesta área, a IA auxilia no diagnóstico junto com o relatório do médico no encaminhamento de pacientes que apresentam risco de câncer, podendo ter seus dados utilizados em terapias personalizadas, através de informações genéticas do paciente [Dlamini et al. 2020]. Dessa forma, na área oncológica a Inteligência Artificial, através da análise de grandes volumes de dados e reconhecimento de padrões [Santos & Baeßler 2018], vem sendo utilizada para prevenção de doenças, diagnósticos precoces, tratamentos e suporte/monitoramento dos pacientes [Ávila-Tomás et al. 2020].

Por consequência, essa tecnologia também é utilizada em exames clínicos, sendo estes a base fundamental para que a mamografia, ultrassom ou raio-x (exames feitos para detectar câncer de mama) possam ser realizados [Bernardes et al. 2019] e, dessa forma, obter parâmetros que possam ser utilizados em relatórios e para auxiliar o médico na tomada de decisão.

Enquanto a ressonância magnética é utilizada em casos onde há implantes de silicone e em situações onde o câncer já está avançado [Houssami & Hayes 2009], a mamografia apresenta os nódulos (lesões redondas/ovais) e o ultrassom exibe se a lesão em questão é cística (que possui água) ou sólida (composta por um conjunto de células). A ultrassonografia oferece um grande potencial para ser uma alternativa viável para identificar câncer de mama precocemente em áreas com recursos mais limitados, pois

além de ser mais portátil, ela custa menos que a mamografia e é mais versátil em uma ampla quantidade de aplicações clínicas [Sood et al. 2019].

O ultrassom de mama vem sendo provado como uma ferramenta excepcionalmente eficaz para a detecção de anomalias na mama, sendo efetivo na detecção de cânceres pequenos e invasivos em tecidos mamários densos, sendo mais efetivo para mulheres abaixo de 35 anos, e conseguindo diferenciar massas benignas e malignas com uma alta acurácia. No geral, utilizar o ultrassom pode aumentar a detecção do câncer em até 17%. Sendo assim, o ultrassom pode ser considerado em alguns casos superior à mamografia, pois como não há radiação ele é mais conveniente e seguro, tanto para pacientes quanto para radiologistas, além de, devido a altas taxas de resultados falsos positivos na mamografia, há muitas biópsias que são realizadas sem necessidade [Cheng et al. 2009].

Entretanto, a IA pode apresentar dificuldades para gerar classificações precisas da doença ao ser aplicada nos exames mencionados anteriormente, fazendo com que o desenvolvimento de terapias específicas e ideais também seja afetado [Shimizu & Nakayama 2020].

Portanto, devido ao grande número de mortes - 17.825 em 2020 no Brasil [INCA 2021] - é necessário estudar campos da ciência que possam ser mais eficientes na detecção de casos de câncer de mama, como a Inteligência Artificial. Atualmente, as técnicas de IA utilizadas para análise e predição do câncer de mama são as Redes Neurais Artificiais, Máquinas de Vetores Suporte, o algoritmo *K Nearest Neighbors* e Árvores de Decisão.

1.1 Algoritmos

As Redes Neurais Artificiais (RNA) são uma técnica de IA inspiradas no processamento de informação do sistema nervoso central, recebendo informação, processando e devolvendo uma resposta através de sua arquitetura em camadas [Raschka 2015]. Sua estrutura e aplicação podem ser adaptadas para resolver diferentes problemas, como, por exemplo, análise de mamografias. Neste caso, normalmente é utilizado aprendizado supervisionado, sendo capaz de detectar a microcalcificação, ou seja, o acúmulo de partículas de cálcio que se convertem em sais e obtêm um formato sólido no tecido mamário [Nascimento et al. 2016]. Além disso, as RNAs podem ser utilizadas nas segmentações de imagens, classificando os dados em “microcalcificação” e “não-microcalcificação” e também no registro de imagens, onde as dimensões dos dados são utilizadas como entrada para treinamento do algoritmo [Mehdy et al. 2017].

As Máquinas de Vetores Suporte (SVM) têm como objetivo encontrar o melhor classificador de dados (a partir de um critério pré-definido) para determinados problemas, partindo de um conjunto de dados de treinamento em um hiperplano e dividindo-os de acordo com a maior margem apresentada (distância entre os hiperplanos e dados limites das classes) [Géron 2017]. O algoritmo separa os padrões pertencentes às classes do problema e faz a classificação binária de acordo com o classificador definido, sendo este escolhido através dos dados que estão próximos ao limite de cada classe, chamados de vetores de suporte. Na oncologia, ela é utilizada na etapa de classificação, após coletar os parâmetros relacionados ao tumor obtidos nos exames laboratoriais. Os dados são pré-

processados, podendo passar pelo processo de redução de dimensionalidade que extrai os atributos relevantes. A SVM utiliza os dados para classificar o tumor como maligno ou benigno.

O algoritmo *K Nearest Neighbors* (KNN) utiliza aprendizado supervisionado, ou seja, o modelo gerado por ele "aprende" a partir de dados de treinamento rotulados. Seu procedimento para classificação consiste em três partes: Comparar, Ordenar e Classificar. Dado um novo objeto, que não faz parte dos dados de treinamento, o KNN primeiro irá calcular a similaridade (através da distância euclidiana) desse objeto de classe desconhecida com os outros da base de treinamento. A similaridade obtida é ordenada de forma crescente, da mais similar para a menos similar. Após essa ordenação, é feita uma análise das classes dos objetos da base de treinamento, onde o rótulo em maior quantidade do(s) *k* vizinhos (objetos mais próximos) é usado para classificar o novo objeto. O objeto em questão recebe a classe predominante no(s) *k*(s) vizinhos mais próximo(s). Esse procedimento também é aplicado para classificar o câncer de mama, separando o conjunto de dados em duas classes: benigno e maligno. Então, é feito o cálculo da média e desvio padrão de cada dado e, em seguida, para cada classe do conjunto, de forma que se possa saber qual classe é mais similar à qual dado, dependendo da distância em que eles se encontram. A previsão final é dada através da classificação de cada dado, de acordo com o tipo de classe que os outros dados mais próximos a eles possuem [Grus 2019].

As Árvores de Decisão possuem aprendizado supervisionado e apresentam os resultados usando regras simples, do tipo Se/Então (If/Then). O objetivo é criar um modelo de predição do valor de uma variável através de regras de decisão (classificação) obtidas de acordo com as características dos dados. A partir de um conjunto de dados, é construída uma árvore que, ao fazer testes com relação às características dos dados de entrada, quando se chega na folha da árvore, consegue-se predizer o rótulo correspondente ao dado que está sendo analisado [Géron 2017]. Sua estrutura se baseia em uma árvore real, com *nós* (atributos descritivos/preditivos), *ramos* (valores) e *folhas* (atributos classificatórios). O *nó raiz* ou *nó pai* é o topo da árvore, seus ramos são as ligações entre *nós estruturais* ou *nós de decisão* e as folhas representam as classes, onde os dados são rotulados. Seguem uma estrutura hierárquica (em *camadas* ou *níveis*) e possuem um processo de decisão que consiste em: percorrer os registros da base de dados; a cada registro inicia-se uma análise a partir do nó raiz; segue-se com o registro pela árvore e caminha-se pelos ramos até um nó de decisão, chegando às folhas, sendo este o final do processo, ou seja, a classificação do registro de entrada [Vanderplas 2017; Grus, 2019].

Os algoritmos mencionados anteriormente são objetos de estudo em diversos projetos na área. Portanto, a fim de obter maior aprofundamento sobre o tema proposto, a seguir são apresentados os trabalhos relacionados que utilizam técnicas de Aprendizado de Máquina para a análise de dados clínicos relacionados ao câncer de mama.

1.2 Trabalhos relacionados

O artigo *Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset* [Austria et al. 2019] faz uma comparação entre os métodos de Aprendizagem de Máquina, a fim de obter aquele que, dentre os estudados

(Floresta Aleatória, Árvore de Decisão, SVM, Regressão Logística, Naive Bayes, *Gradient Boosting* e KNN), é considerado o melhor. De acordo com os dados coletados (Glicose, Resistina, Idade e Índice de Massa Corporal), conclui-se que o *Gradient Boosting* obteve melhores resultados, com acurácia de 74,14%. Porém, o classificador com melhor tempo de treinamento foi o KNN (0.000598 segundos), enquanto para testes foi o SVM, com 0 segundos.

O artigo *Artificial intelligence (AI) and big data in cancer and precision oncology* [Dlamini et al. 2020] mostra que a Inteligência Artificial (IA) e o aprendizado de máquina influenciam significativamente a área da saúde, como no auxílio do diagnóstico precoce para intervenções, através do Sequenciamento de Nova Geração (NGS), que gera grandes conjuntos de dados que demandam recursos especializados em bioinformática para analisar os dados relevantes e clinicamente significativos. Com isso, mostra-se que por meio das aplicações de IA, diagnósticos de câncer e a previsão prognóstica são aprimorados com NGS e imagens médicas, além de também mostrar os desafios, focando no aumento da velocidade de dados de alto *throughput* e aumento da sensibilidade.

O artigo *Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice* [Houssami et al. 2019] foca em aprender técnicas aprimoradas para treinar e validar modelos de IA para o rastreamento de mamografia usando uma combinação de exames de rastreamento que representam cenários do mundo real (em termos de um espectro de resultados positivos e negativos achados em imagens e prevalência de câncer em populações). Mostra-se que devem ser validados usando grandes conjuntos de dados de triagem independentes de diversas populações, com a contribuição de especialistas em imagens e aqueles que trabalham no ambiente de triagem, para garantir a acurácia.

O artigo *Using Resistin, glucose, age and BMI to predict the presence of breast cancer* [Patrício et al. 2018] verifica, através de coletas e análises sanguíneas de rotina, como desenvolver e avaliar um modelo de predição potencialmente utilizado como biomarcador de câncer de mama, baseando-se em dados e parâmetros antropométricos. Todo este estudo sugere que Resistina e Glicose, em conjunto com a Idade e o IMC da pessoa, podem ser considerados um ótimo conjunto de candidatos para implementar em testes de triagem. Portanto, este procedimento visa aumentar a facilidade de diagnóstico do câncer de mama, podendo ter grande impacto na saúde de muitas mulheres.

O artigo *Artificial intelligence in oncology* [Shimizu & Nakayama 2020] descreve casos do aprendizado profundo, um subcampo de IA que é altamente flexível e suporta extração automática de recursos, sendo aplicado em várias áreas da pesquisa básica e clínica do câncer. Alguns desses casos mostram a resolução de problemas que eram visivelmente insolúveis e o artigo também aborda os obstáculos que devem ser superados antes de sua aplicação se tornar mais utilizada, como por exemplo, um conjunto de dados com anotação suficiente em bancos de dados de grande escala e também a tomada de decisão, já que por ter um grande número de parâmetros envolvidos, dificulta a compreensão dos detalhes de como o aprendizado profundo analisa dados e toma as decisões.

O artigo *Clinical and Imaging Surveillance Following Breast Cancer Diagnosis* [Flowers et al. 2012] apresenta os protocolos de imagem de instituições acadêmicas do

Texas, São Francisco, Los Angeles, Tampa e Boston, com foco na mamografia e ressonância magnética. Após tal estudo deduziu-se que, mesmo com as técnicas de imagens, o exame clínico continua sendo imprescindível para diagnósticos, tendo-as como seu complemento. Além disso, verificou-se que para detectar a recorrência de câncer nos estágios iniciais, os médicos devem estar familiarizados com o comportamento dos subtipos de câncer de mama e as diferentes abordagens para seu tratamento.

1.3 Problema de Pesquisa

Portanto, é necessário compreender quais desafios a Inteligência Artificial enfrenta atualmente no contexto da análise de variáveis que influenciam o câncer e como ultrapassá-los, de forma que novos conhecimentos sejam obtidos e, assim, permita que a medicina possa realizar novos estudos na área, através de outras combinações de biomarcadores, além dos mais utilizados na literatura. Neste contexto, a pergunta que será respondida nesta pesquisa é: *Como o Aprendizado de Máquina pode ser utilizado na identificação de possíveis relações entre variáveis clínicas e o câncer de mama?*

A motivação para realizar tal pesquisa surge devido ao alto índice de mortalidade relativo a este grupo de doenças, sendo necessário compreender como a tecnologia pode auxiliar em uma análise de fatores clínicos mais abrangente, identificando possíveis relações entre eles, de forma que possam ser mais estudados e, a partir disso, novas soluções identificadas.

Assim, este projeto está sendo realizado com o intuito de estudar e compreender como a Inteligência Artificial, através do Aprendizado de Máquina, pode contribuir na medicina através da aplicação do algoritmo Árvore de Decisão para entender quais as variáveis mais influenciam em mortes ocasionadas pelo câncer de mama e as relações entre elas, de forma que o exame de sangue possa ser utilizado como uma alternativa viável para um primeiro diagnóstico da doença.

Diante disto, o objetivo geral do trabalho é utilizar a técnica de Aprendizado de Máquina Árvore de Decisão para investigar a influência de variáveis disponíveis em uma determinada base de dados clínicos em mortes ocasionadas pelo câncer de mama e quais as relações entre elas.

Como objetivos específicos, tem-se a aplicação da árvore de decisão à base com o objetivo de identificar quais variáveis possuem maior influência em registros de óbitos, como esses fatores se comportam juntos e qual a situação atual de estudos sobre essas variáveis na área.

Além da Introdução, este artigo possui as seguintes seções: Metodologia, Resultados e Conclusão, que serão apresentados a seguir.

2. Metodologia

O presente projeto propõe a utilização do algoritmo Árvores de Decisão para investigar as influências das variáveis disponíveis em uma base de dados clínicos de câncer de mama e como podemos relacioná-las. Para este tipo de dado as árvores de decisão são adequadas, com a vantagem de que permitem a interpretação das regras de decisão geradas [Austria et al. 2019]. Além disso, é uma técnica de fácil implementação,

compreensão e de rápido processamento, podendo ser uma alternativa viável para ser utilizada em diagnósticos de câncer de mama [Bifet et al. 2017].

Para isso, utilizou-se a base de dados clínicos do Programa de Pesquisa de Imagem de Mama da Universidade da Califórnia em São Francisco. Esta base possui 76.2GB no total, apresentando também imagens, e está disponível para download no site *The Cancer Imaging Archive* (TCIA), do *National Cancer Institute* (NIH) (https://wiki.cancerimagingarchive.net/display/Public/ISPY1). Diante de todas as bases pesquisadas para utilização, esta foi a escolhida por ser uma base longitudinal, ou seja, possui a variável ID e sua sequência de medidas (com a progressão da doença em cada paciente) e que apresenta a variável alvo (status). Porém, não foi utilizada toda a base de dados em questão, e sim uma base construída de acordo com um dos arquivos (extensão csv) disponibilizados.

As características clínicas dessa base foram observadas/medidas para 221 pacientes com câncer de mama, variando entre os status "Vivo" e "Morto". Os atributos utilizados estão mencionados na tabela a seguir.

TABELA 1. Informações sobre a base de dados clínicos da Universidade da Califórnia em São Francisco

Atributo	Valores
ID	1001-1239
Idade	26-68
Receptor de Estrogênio Positivo (ER)	0, 1
Receptor de Progesterona Positivo (PgR)	0, 1
Receptor Hormonal Positivo (HR)	0, 1
Receptor tipo 2 do fator de crescimento epidérmico humano (Her2)	0, 1
Lateralidade	1 (Mama esquerda), 2 (Mama direita)
Status	7 (Vivo), 8 (Morto)

Para poder utilizar a Árvore de Decisão como o método de pesquisa (por ser conhecido e de fácil implementação e compreensão), foi necessário unir colunas de 2 tabelas diferentes, através do método *merge*( ), disponível na biblioteca *pandas*. Primeiramente, as tabelas estavam separadas contendo *age, ERpos, PgRpos, HR Pos, Her2MostPos, Laterality* em uma tabela, e o status em outra. Depois, para evitar possíveis erros, houve a limpeza da base, retirando dados nulos e status com resultados iguais a "*lost*", ou seja, "perdido".

Após preparar a base de dados, o algoritmo da árvore de decisão começou a ser programado. Para isso, houve a separação dos dados em dois grupos, havendo dados para

a coordenada x (*age, ERpos, PgRpos, HR Pos, Her2MostPos, Laterality*) e para a coordenada y (*sstat* - status) do gráfico.

Em seguida, o balanceamento de dados foi feito, devido ao fato de que a base utilizada apresentava uma quantidade desbalanceada na variável alvo, com mais casos apresentando "vivo" como status, de forma que, se não fosse tratado, o algoritmo poderia perder sua capacidade de predição, pois seria aplicado a um conjunto de dados cuja distribuição de instâncias entre classes é desbalanceada [Beckmann 2010]. A estratégia utilizada para fazer tal balanceamento foi a *Oversampling,* que aumenta a amostra da categoria minoritária utilizando técnicas estatísticas para replicação das amostras minoritárias, duplicando-as ou gerando novas amostras a partir das amostras reais [Binuesa 2020]. Para isso, utilizou-se o algoritmo de pré-processamento SMOTE (*Synthetic Minority Oversampling Technique*), que aumenta o número de instâncias de classes minoritárias através da introdução de novos exemplos dessas classes na base de dados, auxiliando o classificador a melhorar sua capacidade de generalização [Fernández et al. 2018].

A partir disso, os dados para treinamento e teste foram separados, seguindo a proporção de 70% e 30%, respectivamente. Com isso, uma matriz de confusão foi utilizada para comparar a quantidade de acertos e erros do algoritmo [Vanderplas 2017], chegando ao resultado de 71 acertos e 18 erros. A partir deste resultado, as técnicas utilizadas para medir a performance do modelo foram: *erro*, *acurácia*, *precisão* e *recall*.

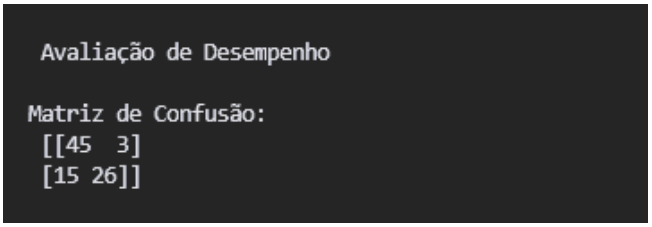

Figura 1. Matriz de Confusão gerada, onde a diagonal principal representa os acertos do algoritmo.

A taxa de erro é utilizada para avaliar os erros cometidos pelo modelo de acordo com as classificações erradas observadas na matriz de confusão. Para calcular a taxa deve-se somar o total de erros ocorridos e dividi-los pelo número total de experimentos realizados [De Castro & Ferrari 2016]:

$$ERR = \frac{quantidade\ \ de\ erros\ ocorridos}{total\ de\ experimentos\ realizados}$$

A taxa de erro do algoritmo foi de 22.47%.

A *acurácia* define, dentre todas as classificações realizadas, quantas o modelo classificou corretamente, indicando uma performance geral do modelo. É uma medida muito utilizada para avaliar o desempenho de classificadores. Trata-se do total de acertos obtidos dividido pelo total de experimentos, podendo ser calculada também como o inverso do erro [Mariano et al. 2020]:

$$ACC = \frac{quantidade\ \ de\ acertos\ obtidos}{total\ de\ experimentos\ realizados} = 1 - ERR$$

O algoritmo apresentou uma acurácia de 77.53%.

A *precisão* significa que as informações que o modelo classificou estão muito próximas do valor real, sendo toleráveis as possíveis diferenças. É utilizada quando os resultados *falsos positivos* são considerados mais prejudiciais que os *falsos negativos*, pois se o modelo é considerado eficaz, ele deve ser preciso em suas classificações; se não for, principalmente na área de estudo da medicina, os resultados podem prejudicar vidas. Seu cálculo se dá através da proporção de acertos dentre aqueles classificados como positivos [Mariano et al. 2020], como apresenta a equação a seguir.

$$PRE = \frac{quantidade\ de\ \ verdadeiros\ positivos}{verdadeiros\ positivos\ +\ falsos\ positivos}$$

A precisão do algoritmo para a classificação de "vivos" e "mortos" foi igual a 72.58% e 88.89%, respectivamente.

A *revocação* (também conhecida por *recall*) apresenta a relação entre os experimentos classificados como determinada classe e aqueles que realmente pertencem à classe em questão. É utilizada em situações nas quais os falsos negativos são considerados mais prejudiciais que os falsos positivos, sendo muito importante no presente trabalho, pois é necessário que o modelo construído encontre a maioria dos pacientes que apresentam câncer de mama, devido ao fato de que uma classificação errada, onde o paciente que possui o tumor maligno é classificado como benigno, pode resultar problemas. Portanto, trata-se da proporção de verdadeiros positivos dentre o total de acertos [De Castro & Ferrari 2016]. Sua equação é representada a seguir.

$$REC = \frac{quantidade\ de\ verdadeiros\ positivos}{verdadeiros\ negativos\ +\ verdadeiros\ positivos} = \frac{quantidade\ de\ \ verdadeiros\ positivos}{total\ de\ acertos}$$

Em relação à revocação, o algoritmo apresentou para a classe "vivos" um resultado igual a 93.75%, porém, para a classe "mortes", o valor foi de 58.54%. Este último resultado pode ter sido influenciado pelo fato de que a base de dados foi balanceada, havendo o aumento da amostra da categoria relacionada.

Após analisar o desempenho, aplicou-se o algoritmo de seleção de atributos univariado *Select K-Best*, disponível na biblioteca *sklearn*, que seleciona os atributos com as maiores pontuações baseados no cálculo estatístico qui-quadrado, realizando combinações lineares de uma matriz de amostras e atributos com o intuito de encontrar quais que mais impactam a classificação, baseando-se em sua variância [Villar 2021]. O algoritmo demonstrou que os fatores que mais influenciaram na classificação foram o ER (Receptor de Estrogênio) e o HR (Receptor Hormonal).

Com isso, o gráfico do algoritmo foi feito, tendo a árvore com uma altura igual a 4 e utilizando o índice Gini para medir a impureza dos *nós*. Este índice mede a heterogeneidade dos dados, tendo como *nós* puros aqueles com resultados mais próximos a 0 [Barbosa et al. 2012]. Quanto mais próximo a 1, mais o *nó* é impuro, ou seja, há diferenças significativas, desigualdades nos dados representados no *nó* em questão. Para visualizar o gráfico (que foi exportado em um arquivo com extensão .dot), utilizou-se o seguinte link: https://tinyurl.com/bdf4mn2f.

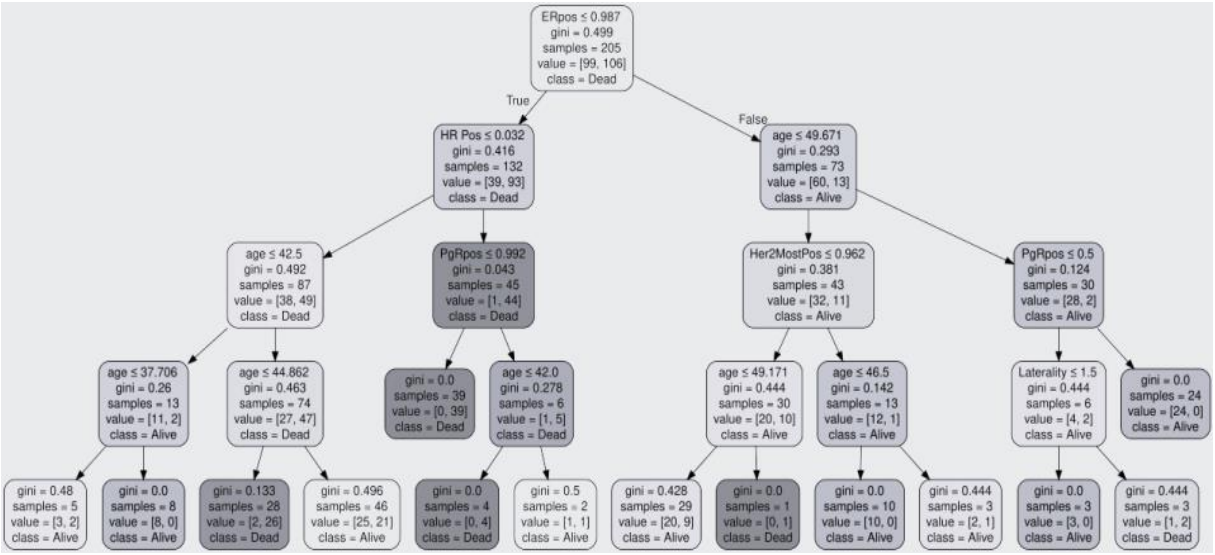

Figura 2. Gráfico da Árvore de Decisão gerado.

O código do algoritmo, o arquivo com extensão .dot gerado e a imagem acima estão armazenados em um repositório no *GitHub*. podendo ser acessados através do link: https://github.com/TamiMaimone/TCC_BreastCancer.

3. Resultados

A partir do gráfico gerado, os seguintes resultados foram observados: Um alto número de mortes ocorreu quando os fatores ER e PgR (Receptor de Progesterona) deram negativos (e com a presença da variável HR), causando cerca de 39 mortes e sem a ocorrência de sobreviventes; enquanto a combinação das variáveis HR e PgR resultou na morte de pacientes com menos de 42 anos, os fatores ER e Her2 (Receptor tipo 2 do fator de crescimento Epidérmico Humano), juntos, sucederam na morte de pessoas acima de 47 anos; ao relacionar o fator ER com o Her2, observou-se que o tumor causou maior influência em óbitos quando alocado na mama direita; assim como demonstrado pelo algoritmo *Select K-Best,* os fatores ER e HR apresentam alta influência na construção do gráfico, devido ao alto número de mortes quando ambos estão presentes (independente se as variáveis são positivas ou negativas).

Estudos apontam que há relação entre as variáveis ER negativo e PgR negativo com o câncer inflamatório, com taxas que demonstram que a baixa presença de ambos os fatores pode gerar maior influência em diagnósticos deste carcinoma do que outros fatores, como Her2 positivo, por exemplo [Castro et al. 2013]. Ao analisar o alto número de mortes através do algoritmo, percebeu-se que é necessário haver a realização de maiores estudos para o câncer de mama inflamatório, tumor este que é uma forma rara e

agressiva de câncer de mama invasivo [Sánchez 2010] e que apresenta prognóstico desfavorável, com fatores de risco ainda pouco compreendidos [Faldoni 2018]. Além disso, analisar também a relação da quantidade de receptores hormonais com a presença deste câncer, se sua presença em maior quantidade pode influenciar em possíveis tratamentos.

Em relação aos fatores HR e PgR, ambos com resultados positivos, foi analisado que há ocorrências destes fatores no câncer de mama Luminal B, sendo caracterizado também por baixa/moderada expressão de genes expressos pelas células epiteliais luminais [Oliveira et al. 2017]. Este resultado corrobora com estudos na literatura que identificam que mulheres mais jovens podem apresentar maior incidência de tumores luminais (tanto A quanto B) em relação às mulheres mais velhas. Porém, em 2012 o Dr. Stephen Johnston, em sua palestra no maior evento de oncologia do mundo, o congresso anual da *American Society of Clinical Oncology* (ASCO), relatou que ainda não se sabe quais são as pacientes que realmente se beneficiam do tratamento. Ou seja, os fatores de predição que temos atualmente são insuficientes [Tiezzi 2014]. Portanto, além da necessidade de haver maiores investimentos em pesquisas e tratamentos para o câncer Luminal B, principalmente em mulheres mais jovens, através dos resultados obtidos pode-se inferir que também é necessário realizar um estudo mais amplo com pacientes abaixo de 42 anos no Brasil, de forma que a terapia antiestrogênica - tratamento que possui bons resultados para este tipo de câncer de mama [Penatti 2019] - seja mais conhecida popularmente pela sociedade, para que todos tenham maior conhecimento sobre este tipo de tratamento.

Ao analisar os resultados da combinação dos fatores ER e Her2 negativo, observou-se que a maioria sobreviveu, porém houve a classificação notável de mortes, sendo necessário o estudo sobre esta relação. Conclui-se que esta combinação também resulta no subtipo Luminal B, apresentado anteriormente, e que resultou no falecimento de pacientes mais jovens, com menos de 49 anos. Já para os resultados acima desta idade, estudos apontam que há relação com a pós-menopausa, com maior ocorrência em pacientes com síndrome metabólica [Motoki 2020] - que representa a associação de diversos fatores de risco cardiovascular, como hipertensão arterial e distúrbios do metabolismo glicídico e lipídico [Meirelles 2013] -, além de se associar com o tumor do subtipo Luminal B Her2 negativo [Motoki 2020]. Com isso, pode-se concluir que é preciso realizar mais estudos sobre a relação dos fatores que afetam a síndrome metabólica com o biomarcador Her2, e se de fato um fator influencia o outro, para que, em caso afirmativo, novos tratamentos possam ser aplicados e diagnósticos mais precisos possam ser efetuados, principalmente para mulheres na pós-menopausa.

4. Conclusão

Este projeto propôs a realização de um estudo sobre o uso da técnica de Inteligência Artificial aplicada à análise de câncer de mama. Mais especificamente, foi proposto o uso do algoritmo árvore de decisão para a análise de dados clínicos associados ao câncer de mama.

A partir disso, chegou-se à seguinte conclusão: Os fatores Receptor de Progesterona e, principalmente, os Receptores Hormonais e de Estrogênio, também coletados a partir de exames de sangue, demonstraram grande influência na detecção do

câncer de mama, além de um alto número de mortes quando o ER e PgR, ambos negativos, estiveram presentes.

Com a utilização da árvore de decisão, encontramos que a relação entre as variáveis ER negativo e PgR negativo pode gerar influência nos diagnósticos de câncer inflamatório. Porém, este é um tumor ainda pouco compreendido. Por isso, através de estudos sobre a quantidade de receptores hormonais relacionados à presença deste câncer, pode-se encontrar novos resultados que gerem prognósticos mais favoráveis e novas opções de tratamento. Além disso, é necessário analisar o biomarcador Her2, principalmente em mulheres que estão na pós-menopausa, e qual sua relação com a síndrome metabólica. Já para mulheres mais jovens, estudar o possível desenvolvimento de novos exames para que o tipo de câncer Luminal B seja identificado precocemente.

Em relação às medidas de desempenho, conclui-se que a Árvore de Decisão é um algoritmo válido para classificar dados clínicos, permitindo identificar quais variáveis (fatores) exercem maior influência na identificação do câncer de mama e as estimativas de sobrevivência e óbito que estão relacionadas a cada uma delas. Demonstrou-se também que a ideia do algoritmo pode ser aplicada em outras bases de dados, de modo que novos resultados sejam obtidos e diferentes classificações sejam feitas, auxiliando em novos estudos na área, de forma que pesquisadores tenham maior foco nas pesquisas e, assim, encontrem soluções mais rapidamente.

Ainda há desafios nessa área da tecnologia, como classificações incorretas e algoritmos com dados heterogêneos, levando a resultados errôneos (algo prejudicial, principalmente na área da saúde). Porém, ao utilizar a IA em resultados de exame de sangue, já podemos obter resultados que auxiliem os médicos. Sua identificação pode não ser tão precisa no primeiro momento, mas esta técnica fará com que o profissional perceba algum fator que ele não tenha visto sozinho, de forma que os diagnósticos se tornem mais precisos.

Portanto, as variáveis clínicas HR, ER, PgR e Her2 apresentam importante influência em diagnósticos de câncer de mama e devem ser coletadas em exames, de modo que se possa identificar tumores precocemente. Além disso, a realização de mais estudos nesta área também é essencial para encontrar novas relações entre estes fatores e, assim, gerar maiores opções de tratamento, para que, ao longo dos anos, o número de casos e mortes relacionado a esta doença diminua cada vez mais.

Referências

Austria, Y., Lalata, J., Goh, M. e Goh, J. (2019) “Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset”, https://www.researchgate.net/publication/337193772_Comparison_of_Machine_Learning_Algorithms_in_Breast_Cancer_Prediction_Using_the_Coimbra_Dataset, Maio.

Ávila-Tomás, J., Mayer-Pujadas, M. e Quesada-Varela, V. (2020) “La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas”, https://www.sciencedirect.com/science/article/pii/S0212656720301463, Maio.

Barbosa, J., Carneiro, T., Tavares, A. (2012) “Métodos de Classificação por Árvores de Decisão Disciplina de Projeto e Análise de Algoritmos”,

http://www.decom.ufop.br/menotti/paa111/files/PCC104-111-ars-11.1-JulianaMoreiraBarbosa.pdf, Outubro.

Beckmann, M. (2010) "Algoritmos Genéticos como Estratégia de Pré-Processamento em Conjuntos de Dados Desbalanceados", http://objdig.ufrj.br/60/teses/coppe_m/MarceloBeckmann.pdf, Outubro.

Bernardes, N., Fonseca de Sá, A., Facioli, L., Ferreira, M., Rigolim de Sá, O. e Costa, R. (2019) "Câncer de Mama X Diagnóstico", https://idonline.emnuvens.com.br/id/article/view/1636, Maio.

Bifet, A., Zhang, J., Fan, W., Zhang, C., Qian, J., Holmes, G. e Pfahringer, B. (2017) "Extremely Fast Decision Tree Mining for Evolving Data Streams", https://dl.acm.org/doi/pdf/10.1145/3097983.3098139, Maio.

Binuesa, F. (2020) "Previsão de Mortalidade em Cirurgia Cardíaca Congênita utilizando Aprendizagem de Máquina", https://repositorio.fei.edu.br/bitstream/FEI/3126/1/fulltext.pdf, Outubro.

Cancer Research UK, "Risk factors". https://www.cancerresearchuk.org/about-cancer/breast-cancer/risks-causes/risk-factors, Maio.

Castro, C., Bolaños, G., Montero, M., Mora, G. (2013) "Cáncer de Mama Inflamatorio: Un Reto Diagnóstico y Terapéutico", https://www.scielo.sa.cr/pdf/mlcr/v30n1/art10v30n1.pdf, Outubro.

Cheng, H.D., Shan, J., Ju, W., Guo, Y. e Zhang, L. (2009) "Automated breast cancer detection and classification using ultrasound images: A survey", https://www.sciencedirect.com/science/article/abs/pii/S0031320309002027, Março.

De Castro, L. e Ferrari, D. (2016) "Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações", São Paulo, Editora Saraiva, https://app.minhabiblioteca.com.br/#/books/978-85-472-0100-5/, Maio.

Dlamini, Z., Francies, F., Hull, R. e Marima, R. (2020) "Artificial intelligence (AI) and big data in cancer and precision oncology", https://www.sciencedirect.com/science/article/pii/S200103702030372X, Maio.

Faldoni, F. (2018) "Identificação de Marcadores Moleculares no Carcinoma Inflamatório de Mama", https://accamargo.phlnet.com.br/Doutorado/2018/FLCFaldoni/FLCFaldoni.pdf, Outubro.

Fernández, A., García, S., Herrera, F., Chawla, N. (2018) "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary", https://www.jair.org/index.php/jair/article/view/11192/26406, Outubro.

Flowers, C., Mooney, B. e Drukteinis, J. (2012) “Clinical and Imaging Surveillance Following Breast Cancer Diagnosis”, https://ascopubs.org/doi/pdf/10.14694/EdBook_AM.2012.32.220, Maio.

Géron, A. (2017), Hands-on Machine Learning with Scikit-Learn and Tensor Flow, Sebastopol, O'Reilly.

Grus, J. (2021) “Data Science do Zero”, Rio de Janeiro, Editora Alta Books, https://app.minhabiblioteca.com.br/#/books/9788550816463/, Maio.

Houssami, N. e Hayes, D. (2009) “Review of Preoperative Magnetic Resonance Imaging (MRI) in Breast Cancer”, https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.20028, Maio.
Houssami, N., Kirkpatrick-Jones, G., Noguchi, N. e Lee, C. (2019) “Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI’s potential in breast screening practice”, https://www.tandfonline.com/doi/pdf/10.1080/17434440.2019.1610387?needAccess=true, Maio.

Instituto Nacional de Câncer (INCA) (2011), ABC do câncer : Abordagens Básicas para o Controle do Câncer, Flama, Rio de Janeiro.

Instituto Nacional de Câncer (INCA), “Câncer de mama”, https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama, Abril.

Instituto Nacional de Câncer (INCA), “Estatísticas de câncer 2020”, https://www.inca.gov.br/numeros-de-cancer, Abril.

Instituto Nacional de Câncer (INCA), “Estimativa 2020”, https://www.inca.gov.br/estimativa/introducao, Março.

Instituto Nacional de Câncer (INCA) (2019), Estimativa 2020: Incidência de Câncer no Brasil, Fox Print, Rio de Janeiro.

Lopes, L. (2014) “Saúde da Mulher: Prevenção e Cuidados do Câncer de Mama”, https://repositorio.ufsc.br/bitstream/handle/123456789/172888/Liana%20Mayra%20da%20Silva%20e%20Souza%20SMNL%20-%20TCC.pdf?sequence=1&isAllowed=y, Maio.

Ludermir, T. (2021) “Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências”, https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd/?format=pdf&lang=pt, Novembro.

Mariano, D., Marques, L., Silva, M., et al. (2021), Data Mining, Grupo A, Porto Alegre.

Mehdy, M., Ng, P., Shair, E., Saleh, N., Gomes, C. (2017) “Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer”, https://downloads.hindawi.com/journals/cmmm/2017/2610628.pdf, Maio.

Meirelles, R. (2013) “Menopausa e síndrome metabólica”, https://www.scielo.br/j/abem/a/sPJDYwf8T5DLgFWSFwwcgLk/?format=pdf&lang=pt, Outubro.

Motoki, A. (2020) “Associação entre a Síndrome Metabólica e o Perfil Imuno-histoquímico do Câncer de Mama em Mulheres na Pós-menopausa”, https://repositorio.unesp.br/bitstream/handle/11449/192926/motoki_ah_me_bot.pdf?sequence=5&isAllowed=y, Outubro.

Nascimento, C., Silva, S., Silva, T., Pereira, W., Costa, M. e Filho, C. (2016) “Breast tumor classification in ultrasound images using support vector machines and neural networks”, https://www.scielo.br/j/reng/a/3qzxTpPnVPHLVqSXFcyR4JN/?format=pdf&lang=en, Maio.

Oliveira, C., Dias, C., Fecury, A., Dendasck, C. (2017) “Atualização Sobre os Principais Aspectos Relacionados ao Câncer de Mama”, https://www.researchgate.net/profile/Carla-Dendasck/publication/333940586_Atualizacao_Sobre_os_Principais_Aspectos_Relacionados_ao_Cancer_de_Mama/links/5ea97de0a6fdcc705097da08/Atualizacao-Sobre-os-Principais-Aspectos-Relacionados-ao-Cancer-de-Mama.pdf, Outubro.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. e Caramelo, F. (2018) “Breast Cancer Coimbra Data Set”, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra, Maio.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. e Caramelo, F. (2018) “Using Resistin, glucose, age and BMI to predict the presence of breast câncer”, https://bmccancer.biomedcentral.com/track/pdf/10.1186/s12885-017-3877-1.pdf, Maio.

Penatti, V. (2019) “Imunoterapia no Câncer de Mama. Revisão de Literatura”, http://www.pensaracademico.facig.edu.br/index.php/repositoriotcc/article/view/1848/1460, Outubro.

Raschka, S. (2015), Python Machine Learning, Packt Publishing, Birmingham, UK.

Rodrigues, D., Rocha, M., Trevisan, D., Prata, D. e Silva, M. (2019) “Proposta de Método para Redução do Conjunto de Regras de Associação Resultantes do Algoritmo Apriori”**,** http://www.ojs.unirg.edu.br/index.php/1/article/view/2788/1541, Maio.

Sánchez, F. (2011) “Análise da expressão do HER-2 no câncer de mama e sua correlação com outros fatores prognósticos”, https://repositorio.unesp.br/bitstream/handle/11449/99208/buitragosanchez_f_me_botfm.pdf?sequence=1&isAllowed=y, Outubro.

Santos, D. e Baeßler, B. (2018) “Big data, artificial intelligence, and structured reporting”, https://link.springer.com/content/pdf/10.1186/s41747-018-0071-4.pdf, Maio.

Shimizu, H. e Nakayama, K. (2020) “Artificial intelligence in oncology”, https://onlinelibrary.wiley.com/doi/epdf/10.1111/cas.14377, Maio.

Sood, R., Rositch, A., Shakoor, D., Ambinder, E., Pool, K., Pollack, E., Mollura, D., Mullen, L. e Harvey, S. (2019) “Ultrasound for Breast Cancer Detection Globally: A Systematic Review and Meta-Analysis”, https://pubmed.ncbi.nlm.nih.gov/31454282, Março.

Souza, A. (2013) “Análise Quantitativa dos Parâmetros Físicos e Radiométricos de Procedimentos Radioterápicos em Tumores de Mama”, https://repositorio.unesp.br/bitstream/handle/11449/121393/000807077.pdf?sequence=1&isAllowed=y, Maio.

Tiezzi, D. (2014) “A busca pela cura do câncer de mama: deveríamos começar tudo de novo?”, https://www.scielo.br/j/rbgo/a/RsT8GczLryVG6BJm7VzzChG/?format=pdf&lang=pt, Outubro.

Vanderplas, J. (2017), Python Data Science Handbook. Sebastopol: O'Reilly.

Villar, G. (2021) “Redução da Dimensionalidade em Dados da Saúde por meio de combinação de algoritmos de Seleção de Atributos”, https://repositorio.unesp.br/bitstream/handle/11449/216220/villar_gho_tcc_sjrp.pdf?sequence=4&isAllowed=y, Outubro.

Yuan, X. (2017) “An improved Apriori algorithm for mining association rules”, https://aip.scitation.org/doi/pdf/10.1063/1.4977361, Maio.