Modelos de Aprendizado de Máquina Aplicados à Detecção de Câncer de Mama

José de Deus e Silva Neto
PET-EE, FEEB
Universidade Federal do Pará (UFPA)
Belém, Brasil
jose.silva.neto@itec.ufpa.br

André Oliveira Carvalho da Silva
PET-EE, FEEB
Universidade Federal do Pará (UFPA)
Belém, Brasil
andre.carvalho.silva@itec.ufpa.br

Matheus Leão Campos
PET-EE, FEEB
Universidade Federal do Pará (UFPA)
Belém, Brasil
matheus.campos@itec.ufpa.br

Wendler Luis Nogueira Matos
PET-EE, FEEB
Universidade Federal do Pará (UFPA)
Belém, Brasil
wendler.matos@itec.ufpa.br

Orlando Fonseca Silva
PET-EE, FEEB
Universidade Federal do Pará (UFPA)
Belém, Brasil
orfosi@ufpa.br

***Resumo*—O objetivo deste trabalho é construir um algoritmo classificador binário de câncer de mama, baseado em dados de exame de sangue e antropométricos (Idade, Índice de Massa Corporal, Glicose, Insulina, Modelo de Avaliação da Homeostase, Leptina, Adiponectina, Resistina e Proteína Quimiotática de Monócitos-1) de 116 indivíduos. Para tal estudo, realizou-se comparativo de desempenho dos seguintes modelos de aprendizado de máquina: Árvore de Decisão, Floresta Aleatória, K-Vizinhos mais Próximos, Redes Neurais Artificiais, Máquinas de Vetores de Suporte e Regressão Logística. As metodologias utilizadas nos dados foram: validação cruzada por k-fold (k = 10); divisão dos dados em 80% treino e 20% teste. Para a primeira, avaliou-se média da acurácia e da sensibilidade. Na segunda, valores de acurácia, sensibilidade, especificidade e área sob a curva da característica de operação do receptor. Além disso, a partir da avaliação de normalidade pelo teste de Kolmogorov-Smirnov e do valor-p e coeficiente de correlação de Pearson ou Spearman, dependendo da variável de entrada, realizou-se o teste apenas com as variáveis com nível de significância de 5%, que foram: Glicose, Insulina, Resistina e Modelo de Avaliação da Homeostase. Como melhor classificador final, teve-se a Floresta Aleatória, no método treino/teste e com 9 variáveis, com 83,3% de acurácia, 100% de sensibilidade, 64% de especificidade e 0,881 de área sob a curva.**

***Palavras Chave*—Câncer de Mama, Classificação, Modelos de Aprendizado de Máquina.**

I. Introdução

O câncer de mama, comum entre as mulheres, atingiu cerca de 2,2 milhões de pessoas no ano de 2020, representando um total de 11,7% dos pacientes diagnosticados com câncer no mundo inteiro e sendo responsável por 6,9% das mortes ocasionadas por câncer nesse mesmo ano [1]. Desse modo, o diagnóstico precoce é essencial, já que a rapidez do mesmo é diretamente proporcional às oportunidades de cura do paciente [2]. Embora grande parte das mulheres conheçam a doença, é comum o sentimento de recusa em se realizar determinados exames precocemente (mamografia, ultrassonografia e autoexame), devido principalmente a: falta de recomendação média, ausência de sintomas visíveis e insegurança ou medo [3].

Estudos revelam que entre 10 e 30% das mulheres que possuem câncer de mama, foram diagnosticadas com tumores benignos, o que revela uma ineficácia ou má interpretação de alguns exames. Além disso, a maioria das mamografias são realizadas em tumores benignos. Com isso, percebe-se que esses exames podem valer-se de outras ferramentas, para auxiliarem na tomada de decisão, e o aprendizado de máquina pode oferecer grande utilidade e bom custo/benefício no processo de diagnóstico do câncer de mama [4].

Vários candidatos a biomarcadores de câncer de mama têm sido relatados ao longo dos anos [5]. O presente trabalho irá analisar as variáveis quantitativas em potencial: idade, Índice de Massa Corporal (IMC), Glicose, Insulina, Modelo de Avaliação da Homeostase (HOMA), Leptina, Adiponectina, Resistina e Proteína Quimiotática de Monócitos-1 (MCP-1).

Ante a tal problema de saúde pública, o trabalho analisa o desempenho de 6 algorítimos de Aprendizado de Máquina para classificação de dados, sendo eles: Árvore de Decisão, Floresta Aleatória, K - Vizinhos mais Próximos, Máquinas de Vetores de Suporte, Redes Neurais Artificiais e Regressão Logística [6]. O objetivo é discutir a precisão e eficiência em predizer a ocorrência do câncer de mama em indivíduos, a partir das variáveis de entrada [7], podendo servir como ferramenta auxiliadora para a comunidade médica, no diagnóstico precoce do câncer de mama, representando uma possível solução rápida e eficiente que organiza os pacientes de forma que isso permita a tomada de medidas mais direcionadas, facilitando o trabalho dos médicos envolvidos. Importante ainda ressaltar que os resultados obtidos por meio dos modelos utilizados não são o diagnóstico final dos indivíduos. Na seção II tem-se trabalhos relacionados. Na III a metodologia é detalhada, com a exposição da base de dados e dos modelos. Na seção IV, apresenta-se os resultados e as discussões. Por fim, na V conclui-se o trabalho.

II. Trabalhos Relacionados

O século XXI proporcionou vários avanços nas técnicas de análise e classificação de dados, e o aprendizado de máquina está sendo aplicado nas mais diversas áreas. Com isso, a detecção do câncer de mama pode ser feita de maneira mais precisa e menos onerosa [4]. Por conseguinte, os chamados biomarcadores são rotineiramente utilizados como atributos de modelos do aprendizado de máquina para a classificação do câncer de mama.

Em 2008, um algoritmo de regressão logísica utilizou-se de 2 entradas: antígeno específico 15-3 e proteína 3 de ligação ao fator de crescimento semelhante à insulina. A métrica Característica de Operação do Receptor [Receiver Operating Characteristics (ROC)] produziu uma Área Sob a Curva [Area Under the Curve (AUC)] de 0,86; além de uma sensibilidade de 85% e especificidade de 62%, ao predizer se os dados correspondiam a um paciente (tem a doença) ou controle (saudável) [8].

Dalamaga et al [9], em 2013, utilizaram a resistina sérica como biomarcador para o câncer de mama pós-menopausa, encontrando AUC de 0,71, com Intervalo de Confiança (IC) de 95%.

Os algoritmos: regressão logística, floresta aleatória e máquina de vetores de suporte, em 2018, analisaram os mesmos 9 atributos que serão utilizados nesse trabalho. Aplicaram a metodologia de validação por Monte Carlo. O modelo Regressão Logística alcançou 0,81 de aprendizado na curva ROC, 76% de sensibilidade e 86% de especificidade. Da mesma forma, para a Floresta Aleatória, esses valores foram de 0,83, 85% e 77%, respectivamente. Por fim, o modelo Máquinas de Vetor de Suporte, obteve os valores de 0,85, 81% e 84%. Porém, os melhores resultados foram obtidos utilizando apenas 4 variáveis, que tiveram melhor avaliação pelo Coeficiente de Gini: Resistina, Glicose, Idade e IMC. A sensibilidade variou entre 82 - 88%, enquanto especificidade 85 - 90% [10].

Em 2019, Pham e Pham [11] também aplicaram aprendizado de máquina para classificar o câncer, através dos atributos: glicose, IMC, resistina, idade, HOMA, leptina e adiponectina. Os modelos classificadores foram: floresta aleatória e regressão logística múltipla. O segundo apresentou os melhores resultados, com 75% de acurácia e 0,849 de aprendizado na curva ROC.

III. Metodologia

A. Dados

Foi utilizada a base de dados de Câncer de Mama de Coimbra, disponibilizada pelo repositório da UCI (Universidade da Califórnia Irvine), publicada em 06 de Março de 2018 [10], a qual possui dados de 116 indivíduos, obtidos por meio de exames de sangue e antropometria, sendo 64 destes diagnosticados com câncer (pelo exame de mamografia) e 52 saudáveis.

A base de dados possui 116 linhas (indivíduos) e 10 colunas, na qual a última coluna corresponde à classe de saída (1 = saudável e 2 = paciente), e as restantes são variáveis quantitativas: idade, IMC, Glicose, Insulina, HOMA, Leptina, Adiponectina, Resistina e MCP-1. Em uma tentativa de reduzir a quantidade de variáveis de entrada, o que poderia implicar na necessidade de se realizar menos exames nas pessoas, utilizou-se o Software BioEstat 5.0 e precisou-se inicialmente fazer uma avaliação de normalidade dos atributos/variáveis. Para tal, aplicou-se o teste de Kolmogorov - Smirnov, com nível de significância de 5%. As variáveis normais, cujo valor-p foi maior que esse nível e que permitiram o uso da correlação de Pearson, foram: Idade, IMC e MCP-1; enquanto que as anormais, com valor-p menor que o nível e que permitiram a correlação de Spearman: Glicose, Insulina, HOMA, Leptina, Adiponectina e Resistina. A Tabela I apresenta a média (desvio padrão) das variáveis normais, assim como a mediana (amplitude interquartil) para as anormais; também mostra o valor-p (adotou-se nível de significância de 5%) e o coeficiente de correlação (Pearson ou Spearman, dependendo do atributo).

Ao se avaliar os resultados, além do teste com todas as 9 variáveis, fez-se também com 4: Glicose, Insulina, HOMA e Resistina, pois foram as que apresentaram valor-p, na correlação, menor que 5%; com coeficientes de correlação acima de 0,2. O atributo mais próximo foi o IMC, com 0,1326 de correlação, porém com valor-p excendendo o permitido em 10,59%. Um outro ponto importante é que a análise de todos os algorítimos fora realizada com auxílio da plataforma Google Collaboratory, a partir da linguagem Python [12].

Tabela I: Parâmetros estatísticos, para 64 pacientes e 52 de controle, além do teste de correlação e consequente valor-P

Variáveis de entrada	Pacientes	Controle	Valor-p da correlação	Coeficiente de correlação
Distribuição Normal	**Média (Desvio Padrão)**	**Média (Desvio Padrão)**		
Idade (anos)	56,67 (13,49)	58,07 (18,95)	0,6425	-0,0436
IMC (Kg/m²)	26,98 (4,62)	28,31 (5,42)	0,1559	-0,1326
MCP-1 (pg/dL)	563,01 (384,00)	499,73 (292,24)	0,3292	0,0914
Distribuição Anormal	**Mediana (Amplitude Interquartil)**	**Mediana (Amplitude Interquartil)**		
Glicose (mg/dL)	98,5 (17)	87 (10,5)	< 0,0001	0,4561
Insulina (µU/mL)	7,58 (11,65)	5,48 (2,69)	0,0257	0,2071
HOMA	2,05 (3,42)	1,13 (0,89)	0,0025	0,278
Resistina (ng/mL)	14,37 (14,85)	8,92 (6,21)	0,0015	0,2904
Adiponectina (µg/mL)	8,44 (6,77)	8,12 (5,36)	0,7658	0,028
Leptina (ng/mL)	18,87 (24,97)	21,49 (24,87)	0,9472	0,0062

B. Validação cruzada e divisão treinamento/teste

Foram aplicadas 2 metodologias para futura avaliação de desempenho dos algoritmos. A primeira consistiu na validação cruzada pelo técnica k - fold , que utilizou toda a base de dados e consistiu em realizar 'k' partições (1 para teste e k-1 para treino), alternando os dados de treino e teste por 'k' vezes. Fez-se uso do k = 10 em todos os 6 modelos de classificação [13]. Já na segunda configuração, a base de dados foi dividida entre treino e teste, na seguinte proporção: 20% para teste e 80% para o treinamento, pelo comando 'split', com uso do parâmetro 'stratify' e 'random state', sendo este último igual a 300 (escolha aleatória), para garantir que, cada vez que o código fosse inicializado, a divisão fosse a mesma.

C. Métricas de desempenho

A matriz de confusão, Figura 1, para a problemática abordada, consiste em uma matriz 2X2, atribuída à classificação binária. As linhas representam as saídas reais, presentes na

base de dados; já as colunas as saídas preditas pelo algoritmo. Na diagonal principal constam os dados classificados de maneira correta pelo algoritmo (Verdadeiro Positivo = pessoa doente classificada como doente e Verdadeiro Negativo = pessoa saudável classificada como saudável) e na diagonal secundária estão os dados classificados de maneira errônea (Falso Positivo = pessoa saudável classificada como doente e Falso Negativo = pessoa doente classificada como saudável). A partir dessa matriz, pode-se calcular diversas métricas, que são formas de analisar o desempenho dos modelos.

Na validação cruzada, as avaliações de desempenho dos modelos foram obtidas pela média da acurácia e da sensibilidade nas 10 iterações. Já na divisão treino/teste, utilizou-se a acurácia, sensibilidade, especificidade e a Área sob a Curva (AUC), sendo a Curva Característica de Operação do Receptor (ROC).

A acurácia, representada pela Equação 1, indica uma performance geral do modelo, avaliando, dentre todas as classificações, quantas o modelo classificou corretamente (entre doentes e saudáveis), enquanto que a sensibilidade, Equação 2, avalia a quantidade de acertos das pessoas que tem câncer, em relação a todas as que tem câncer. Já a especificidade, Equação 3, se refere aos acertos dos indivíduos saudáveis, em relação a todos os saudáveis. É importante ressaltar que, para aplicações de detecção de doenças, o algoritmo precisa ter uma boa eficiência na sensibilidade, pois esta medida representa as pessoas que realmente têm a doença. Portanto, o erro deve ser mínimo para essa métrica.

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN} \tag{1}$$

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{3}$$

	Valor predito	
Valor Real	Sim	Não
Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 1: Matriz de Confusão

Outra métrica inserida no algoritmo foi a Curva AUC-ROC, sendo ROC um gráfico da sensibilidade (taxa dos verdadeiros positivos) em função da taxa dos falsos positivos (1 - especificidade). Já a AUC é a área sob a curva ROC, possuindo uma variação de 0 a 1, na qual quanto mais próximo de 1 a estiver, melhor é o modelo em predizer se o paciente realmente tem a doença, ou seja, o modelo está separando corretamente as classes.

D. Modelos de Aprendizado de Máquina

Para definir os parâmetros mais adequados de cada modelo de aprendizado de máquina, aplicou-se o ”Grid Search”, que permuta entre vários valores escolhidos previamente (sob pesquisa bibliográfica e/ou escolha pessoal) e retorna os que resultaram em melhor desempenho. Para a situação atual de detectar doença, a métrica base foi a sensibilidade.

1) K Vizinhos Mais Próximos: O algorítimo K - Vizinhos Mais próximos é um dos modelos mais simples de aprendizado supervisionado [14]. Seu método de classificação não necessita de tempo de treinamento e baseia-se em distâncias calculadas entre dois pontos, avaliadas pelo modelo assumindo que os pontos de mesma classe estariam localizados próximos uns dos outros. Dessa forma, os vizinhos mais próximos ditarão a presença de câncer de mama nas amostras em análise.

Para realizar tal classificação, é feita a escolha do parâmetro K [15], que simboliza o número de vizinhos mais próximos (não confundir com o ”k - fold” da validação cruzada). Nesse sentido, o número (inteiro) escolhido precisa satisfazer certas condições como:

- Pequeno o suficiente para evitar que a amostra seja erroneamente classificada.
- Grande o suficiente para garantir que o modelo não sofra ”overfitting” (perda de generalização).

O modelo K-NN foi testado, com o parâmetro ‘k’ na faixa de 1 a 10, dentre esses, o valor 7 obteve melhores resultados. Outros parâmetros de destaque são a métrica e o peso. Métrica é a forma como as distâncias são calculadas entre os vizinhos e a amostra, a mais comum sendo a Euclidiana, que generaliza a menor distância entre dois pontos em ‘n’ dimensões. No entanto, a métrica utilizada foi a distância de Manhattan [16], na qual seu cálculo afirma que a soma das projeções ortogonais dos segmentos de reta que une dois pontos é constante.

Por sua vez, o peso determina se vizinhos mais próximos terão mais relevância no momento de determinar a classe da amostra, ou se todos terão relevância uniforme. Ambos foram avaliados neste modelo, e o peso por maior relevância teve maior eficiência.

2) Regressão Logística: A regressão logística tem como objetivo gerar um modelo que, por meio de observações de variáveis independentes (atributos de entrada), é capaz de prever a probabilidade de um evento ocorrer, que, normalmente, é representado por uma variável binária [17].

Para a criação desse modelo, foram utilizados dois parâmetros: solucionador e número máximo de iterações. O solucionador é o algoritmo que será utilizado pelo modelo e o segundo parâmetro define a quantidade de vezes na qual o solucionador será executado. Variações destes parâmetros, escolhidos de maneira arbitrária, foram testadas, são elas, para o solucionador, ‘lbfgs’, ‘newton-cg’, ‘liblinear’, ‘sag’ e ‘saga’. Para o número máximo de iterações, os valores utilizados foram 500, 1000, 1500 e 2000. Após o uso da ferramenta ‘GridSearch’, os valores utilizados para os parâmetros citados foram: Solucionador = ’lbfgs’ e número máximo de iterações = 500.

3) Máquina de Vetor de Suporte: Para que se entenda o funcionamento da Máquina de Vetor de Suporte, é necessário o conhecimento de 4 conceitos: O hiperplano de separação, o hiperplano de margem máxima, a margem suave e a função Kernel. O hiperplano de separação é o equivalente a uma linha de separação, no entanto, em uma dimensão maior que 2. Seu papel é definir a fronteira entre as amostras, de modo que semelhantes estejam juntos e ao serem inseridos novos dados, estes sejam classificados corretamente [18].

O hiperplano de margem máxima é o que diferencia a Máquina de Vetor de Suporte dos demais classificadores com base em hiperplano. Existem diversas maneiras de separar dois grupos, no entanto, o hiperplano de margem máxima é considerado o melhor. Para que isso aconteça, o classificador calcula a menor distância entre duas amostras de diferentes grupos e encontra o valor médio, em seguida traça o hiperplano [18].

A margem suave ocorre quando o classificador utiliza uma margem que aceita classificações erradas para os dados de treino para que, durante os testes, as classificações tenham um baixo percentual de erro. Tal situação ocorre em bases de dados em que ao menos uma amostra de um grupo X, se encontra próxima das amostras de um grupo Y [18].

A função Kernel soluciona os problemas em que a criação de um hiperplano ou linha seja impossível sem que haja erros de classificação futuramente, por meio do aumento da dimensão dos dados. Por exemplo, para um conjunto de dados unidimensional em que não é possível traçar o hiperplano, a função Kernel aumenta para duas dimensões, sendo a nova dimensão, o quadrado dos valores iniciais. Dessa forma é possível traçar um hiperplano correspondente para aquele conjunto de dados, que anteriormente, era impossível [18].

Para este modelo, foram utilizados três parâmetros: Kernel, Gamma e C. O kernel é o algoritmo utilizado pelo modelo, Gamma é coeficiente para kernel “rbf”, ”poly”e “sigmoid”e o C é o parâmetro de regularização. Variações destes parâmetros, escolhidos de maneira arbitrária, foram utilizadas, são elas, para o Kernel, “poly”, “linear”e “rbf”, para o Gamma, valores de 1, 10, 100, 150, 200 e “scale”. Para o C, valores de 1, 10, 100 e 1000. Após o uso da ferramenta ‘GridSearch’, os valores escolhidos para os parâmetros utilizados foram: Kernel = ‘poly’, C = 10 e gamma = 1.

4) Árvore de Decisão: O algoritmo da árvore de decisão consiste em classificar dados através da análise de seus atributos, com a finalidade de representar de forma eficiente o conhecimento obtido através do conjunto de entrada. No modelo em questão, os nós representam os testes feitos sobre os valores dos atributos, já os arcos indicam a possível saída para um determinado teste e, por fim, as folhas mostram a classificação final da árvore sobre o conjunto de dados [19].

Foram utilizados 3 parâmetros na implementação do algoritmo da árvore de decisão: critério, estado aleatório e profundidade máxima. O critério é a função que mede a qualidade de uma divisão, são suportados dois parâmetros: gini (impureza de gini) e entropia (ganho de informação). Além disso, a árvore possui a variante da profundidade máxima da árvore, que vai definir até onde a árvore será ramificada e os parâmetros variam de 0 até infinito. Por fim, a árvore de decisão possui também o parâmetro estado aleatório, onde o principal objetivo é controlar a aleatoriedade dos dados, ou seja, se for atribuído o número 80 para o estado aleatório, a saída dos dados será sempre a mesma [7]. No trabalho, foram utilizados os seguintes valores no modelo: profundidade máxima = 2,critério = ”entropia”e estado aleatório = 100. Vale ressaltar que tais parâmetros foram obtidos através da ferramenta ”GridSearch”.

5) Floresta Aleatória: O modelo Floresta Aleatória caracteriza-se por realizar classificação ou regressão baseado no modelo da Árvore de Decisão [20]. Porém, algumas diferenças surgem ao analisar os critérios para ramificação dos nós.

Em sua aplicação, o algoritmo seleciona aleatoriamente as características que irão compor as raízes das árvores, constituindo, assim, diferentes modelos. Em seguida, as ramificações são realizadas a partir dos mesmos cálculos de impureza presentes no modelo da Árvore de Decisão. Ao final desse processo, os dados de teste serão classificados sob os critérios de “n” árvores (avaliadas de 1 a 300) e, por análise de moda estatística, será inferida a classe da amostra. Neste trabalho 19 árvores realizaram melhor a classificação.

Um parâmetro importante a destacar a respeito desse modelo é o ato de criar um “bootstrap”, que significa a geração de um sub-conjunto de dados [21]. Nela, o algoritmo seleciona aleatoriamente amostras dos dados de treino, possivelmente até repetidas, e os aplica na concepção das árvores. Esse parâmetro tem a função de reduzir a ocorrência de “overfitting” e melhorar a estabilidade do algorítimo. Outro parâmetro utilizado foi o de máxima profundidade das árvores, responsável por ditar quantas subdivisões cada nó fará, ou seja, o número máximo de sub-classificações feitas antes de classificar finalmente a amostra. Foi avaliado na faixa de 1 a 10, com o número 7 obtendo maior performace.

Por fim, com o fito de evitar a aleatoriedade de resultados, a declaração do modelo ainda conta com um parâmetro chamado “random-state”, ou “estado-aleatório”, cuja função é padronizar as seleções de entradas de treino. Nesse sentido, avaliando o intervalo de 1 a 500, o valor 50 foi escolhido.

6) Redes Neurais Artificiais: As Redes Neurais Artificiais (RNA’s) são constituídas por unidades simples (neurônios) e se baseiam em funções matemáticas não-lineares, de modo a obter uma organização e generalização dos dados. De maneira similar ao sistema nervoso biológico, os neurônios são organizados por uma ou mais camadas, se interligando por inúmeras conexões (sinapses). Na Rede Neural Artificial, as sinapses representam o peso sináptico, responsável pela ponderação dos dados de entrada em cada neurônio [22].

O processo de aprendizagem de uma rede neural ocorre através de inúmeras iterações e correções sucessivas dos pesos sinápticos. Tal correção só é possível após a rede fornecer uma saída e realizar a comparação com a saída real, o que representa a função erro. Em seguida, a rede irá propagar os

dados de volta para a entrada e fazer a correção dos pesos aplicados, etapa denominada de “backpropagation” [23].

Para a implementação do algoritmo, 6 parâmetros foram utilizados: solucionador, tamanho de camadas ocultas, taxa de aprendizagem inicial, função de ativação, número máximo de iterações e estado aleatório. A melhor combinação foi: solucionador = “adam”, uma camada oculta, com 3 neurônios, taxa de aprendizagem inicial = 0.1, ativação = “logística”, número máximo de iterações = 200 e estado aleatório = 100 [7].

IV. Resultados e Discussões

Esta seção será dividida em duas. Na primeira, serão apresentados os resultados da aplicação da validação cruzada k-fold; e na segunda da divisão treinamento/teste. Ressalta-se que foram realizados duas considerações: utilizando todos os 9 atributos de entrada; apenas 4 variáveis, estatisticamente significativas com relação a saída binária classificatória. A ideia é verificar se essa redução de dimensionalidade vai melhorar/piorar ou manter os mesmos valores de métricas de desempenho que utilizando todas as 9.

A. Validação Cruzada

Os dados mostrados na Figura 2 e Figura 3 representam a média das métricas acurácia e sensibilidade, respectivamente, das 10 iterações, já que utilizou-se a validação cruzada (para avaliar a capacidade de generalização de cada um dos seis modelos), pelo método k - fold, com um valor de ‘k’ igual a 10, um valor que experimentalmente admite poucos erros de predição (“bias” e variância).

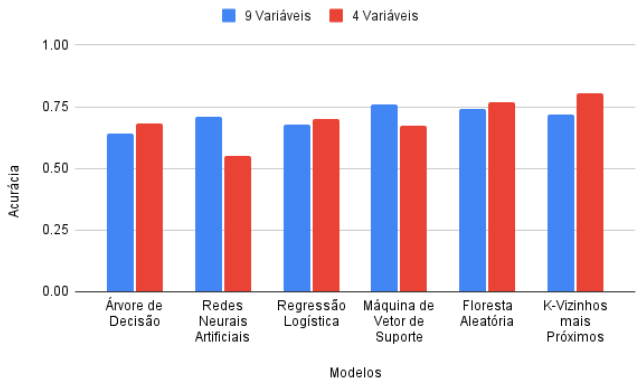

Figura 2: Acurácias da validação cruzada dos modelos

De acordo com a Figura 2, o modelo de Máquina de Vetor de Suporte obteve a maior acurácia, para 9 variáveis, na validação cruzada (76%). Para 4 variáveis, o modelo com maior acurácia foi o K - Vizinhos mais Próximos (80,45%). Percebe-se, então, que mesmo com menos atributos quantitativos de aprendizado (que reduz o processamento do computador), foi possível obter uma melhor acurácia.

Em termos de sensibilidade, Figura 3, o modelo de Regressão Logística alcançou o melhor resultado para 9 variáveis, em relação aos demais modelos, atingindo 86.4%. Para 4 variáveis, o modelo Redes Neurais Artificiais apresentou maior sensibilidade, 100%. Novamente obteve-se melhor desempenho para o teste de redução de dimensionalidade, representando um aumento de 13,6%.

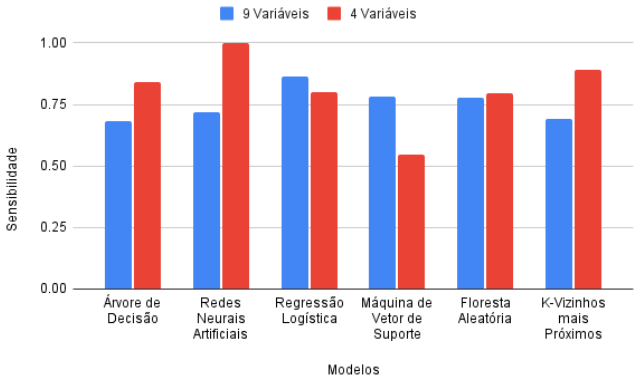

Figura 3: Sensibilidades da validação cruzada dos modelos

B. Treinamento/Testes

1) Métricas pela matriz de confusão: A Tabela II, representa os valores de desempenho para 9 e 4 variáveis. Ressalta-se que, para a presente aplicação, a métrica sensibilidade tem uma importância consideravalmente alta, pois é necessário reduzir ao máximo a possibilidade de errar a classificação de indivíduos com câncer de mama. Por consequência, pode-se permitir valores relativamente mais baixos de especificidade, pois se uma pessoa saudável for classificada como doente, o máximo que irá acontecer é ser pedido um exame a mais, haja vista que é necessário recordar que a saída dos modelos propostos não representam a verdade absoluta, e sim são ferramentas de auxílio. Por outro lado, se alguém doente for definido como saudável, já seria um erro grave. Em termos de acurácia (total de acertos de doentes e saudáveis) e considerando todos os atributos, o modelo de Redes Neurais Artificiais obteve o melhor desempenho (83,33%). Além disso, o algoritmo de Redes Neurais Artificiais alcançou 100% de especificidade (acerto dos saudáveis). Em relação à sensibilidade (acerto dos doentes), os modelos: Floresta Aleatória e Árvore de Decisão, atigiram o melhor resultado possível (100%), no entanto, a Floresta Aleatória atingiu uma acurácia maior (83,3%) e por este motivo foi escolhido como o melhor modelo para 9 variáveis.

Em relação ao desempenho dos modelos com a utilização das 4 variáveis, o modelo Máquina de Vetor de Suporte obteve o melhor desempenho em relação à acurácia (79%). Dentre os seis, três modelos atingiram 100% de sensibilidade, são eles: Regressão Logística, Árvore de Decisão e Redes Neurais Artificiais. Por último, no quesito especificidade, Máquina de Vetor de Suporte alcançou 91%, a maior entre os modelos.

Novamente, valendo-se da sensibilidade e do balanço por acurácia, entende-se que a Árvore de Decisão conseguiu maior eficácia na tarefa classificatória proposta. É notável, também,

verificar que a redução do número de variáveis utilizadas não resultou em melhora geral tão significativa, analisando todos os modelos.

Tabela II: Resultados dos testes

Modelos	Acurácia	Sensibilidade	Especificidade
	9 | 4	9 | 4	9 | 4
Regressão Logística	62,5% | 71%	92% | 100%	27% | 36%
Máquina de Vetor de Suporte	83% | 79%	77% | 69%	91% | 91%
K-Vizinhos mais Próximos	79,16% | 67%	77% | 77%	82% | 55%
Floresta Aleatória	83,3% | 75%	100% | 69%	64% | 82%
Árvore de Decisão	75% | 75%	100% | 100%	45% | 45%
Redes Neurais Artificiais	83,33% | 54%	69% | 100%	100% | 0%

2) Curva AUC-ROC: A curva ROC é a representação gráfica da relação entre a taxa de verdadeiro positivo e a taxa de falso positivo. A AUC varia de 0 a 1, sendo que, quanto mais próximo de 1 estiver (distante, para a parte superior, da predição aleatória AUROC = 0,5, mais generalista no aprendizado é o modelo. As Figuras 4 e 5 mostram que os modelos que alcançaram os maiores valores de AUC, para 9 e 4 variáveis, respectivamente, foram: Floresta Aleatória (AUROC = 0,881) e Regressão Logística/Máquina de Vetor de Suporte (AUROC = 0,860). Por outro lado, o modelo da Árvore de Decisão obteve o pior resultado (para 9 variáveis), com AUROC = 0,825 e, para 4 variáveis, o modelo com menor valor foi a Rede Neural Artificial, com AUROC = 0,119, que representa um valor muito pequeno de aprendizado.

Como a ideia era comparar o desempenho entre 9 e 4 atributos de entrada, os parâmetros encontrados pela ferramenta "Grid Search", na situação inicial (com todas as variáveis), foram aplicados na redução para 4. Com isso, valores mais extremos, como o da RNA, tornam-se possíveis.

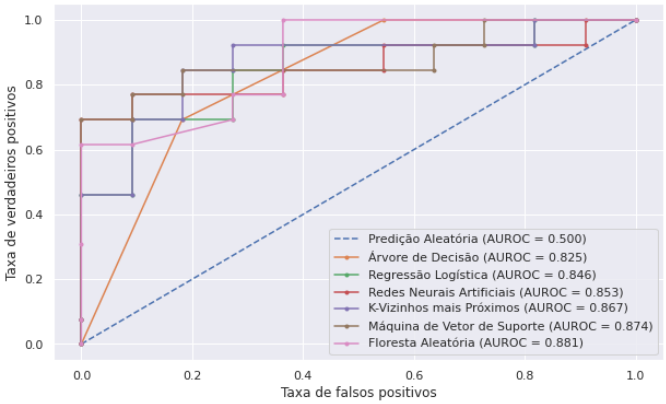

Figura 4: Curva ROC e valor AUC, para 9 variáveis

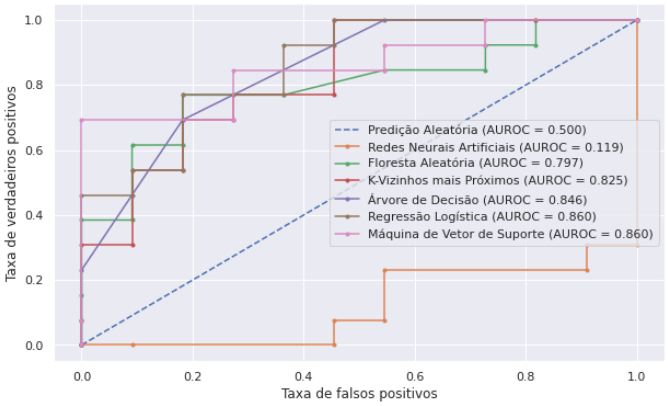

Figura 5: Curva ROC e valor AUC, para 4 variáveis

Dessa forma, em comparação com o artigo [10], no qual o seu melhor modelo (Floresta aleatória) obteve uma sensibilidade e especificidade de 85% e 77%, respectivamente, além de AUROC = 0,85, o modelo por Redes Neurais Artificiais presente no trabalho, por exemplo, apresentou 100% de sensibilidade para 4 variáveis na validação cruzada. Além disso, na divisão treino/teste para 9 atributos, 2 modelos alcançaram 100% de sensibilidade, sendo esses os modelos da Floresta Aleatória e Árvore de Decisão, enquanto que na ultilização de 4 atributos, 3 modelos também obtiveram o máximo desempenho: Regressão Logística, Árvore de Decisão e Redes Neurais Artificiais. Ademais, a especificidade superior a 77% foi alcançada pelo Modelo Máquina de Vetor de Suporte para as divisões de 9 e 4 atributos. Por fim, vale ressaltar que o modelo da Regessão Logística para a AUROC alcançou um resultado de 0,86, superando o modelo do artigo citado.

A AUROC = 0,881, obtida pela Floresta Aleatória (com 9 entradas), superou os valores de 0,86; 0,71 e 0,849; encontrados, respectivamente, por [8], [9] e [11].

A Tabela III expõe os melhores modelos, distinguindo entre os dois métodos avaliados no trabalho (validação cruzada com k = 10 e 80% treino - 20% teste), e a quantidade de atributos de entrada (9 - todos possíveis e 4 - após a análise dos valores-p e correlação). Para validação cruzada, a Rede Neural Artificial, com 4, foi melhor por apresentar 100% de sensibilidade e a Regressão Logística, com 9, por atingir 86,4%. Por treino/teste, com 4, a Regressão Logística obteve 100% de sensibilidade e AUROC 0,860 e, com 9, a Floresta Aleatória alcançou 100% de sensibilidade e AUROC 0,881. Entre os dois métodos de avaliação, a escolha final de melhor topologia é a Floresta Aleatória, haja vista que a Rede Neural Artificial, por validação e 4 atributos, teve pouco mais de 50% de acurácia, e a Floresta Aleatória teve 83,3%.

Tabela III: Melhores modelos de aprendizado

Métodos de avaliação			
Validação Cruzada, k=10		80% Treino - 20% teste	
4	9	4	9
Redes Neurais Artificiais	Regressão Logística	Regressão Logística	Floresta Aleatória

V. Conclusões

O trabalho se propôs em testar diferentes modelos de aprendizado de máquina, na tarefa de detectar a presença de câncer de mama, em 116 indivíduos, a partir de dados de exame de sangue e antropométricos. Foram aplicadas duas técnicas: validação cruzada, com k = 10, e divisão dos dados em 80% treino e 20% teste. Por teste de Kolmogorov-Smirnov e consequente correlação (Pearson ou Spearman), considerando nível de significância de 5%, também realizou-se testes com a redução, de 9 atributos, para 4, sendo estes: Glicose, Insulina, HOMA e Resistina. A programação foi realizada na plataforma Google Collab. As métricas de avaliação de desempenho foram acurácia e sensibilidade para a validação cruzada, e acurácia, sensibilidade, especificidade e AUROC para treino/teste. Os parâmetros dos modelos de aprendizado foram encontrados por pesquisa bibliográfica prévia e escolha pessoal, além de utilizar o comando "GridSearch", da Linguagem Python.

Os resultados evidenciaram, de modo geral, uma aproximação dos resultados para 9 e 4 atributos, não representando grande melhoria em se economizar alguns exames. Como resultado final, o modelo de aprendizagem de máquina Floresta Aleatória, com 19 árvores, 7 subdivisões e estado aleatório 50, obteve melhor avaliação geral dentre todos, obtendo 100% de sensibilidade, 83,3% de acurácia, 64% de especificidade e 0,881 de AUROC.

Ressalta-se, então, o êxito nos objetivos do presente trabalho, e a possibilidade de se realizar os exames nas clínicas de saúde no Brasil, enviar automaticamente os valores numéricos para tabelas no Software Excel, e desenvolver uma rotina em Python que já utilize esses dados na entrada do algoritmo de Floresta Aleatória, para que, em tempo real e com boa velocidade, a resposta do programa possa auxiliar na tomada de decisão por parte do profissional da área da saúde.

Agradecimentos

Ao Programa de Educação Tutorial - Engenharia Elétrica (PET-EE), da Universidade Federal do Pará (UFPA), pelo apoio e financiamento.

Referências

[1] Globocan, (2020, Dez), "Breast cancer data in 2020." https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf.

[2] F. Nascimento, M. Pitta, and M. Rêgo, "Análise dos principais métodos de diagnóstico de câncer de mama como propulsores no processo inovativo," *Arquivos de Medicina*, vol. 29, pp. 153–159, 2015.

[3] G. Santos and R. Chubaci, "O conhecimento sobre o câncer de mama e a mamografia das mulheres idosas frequentadoras de centros de convivência em são paulo (sp, brasil).," *Ciência Saúde Coletiva*, vol. 16, pp. 2533–2540, 2011.

[4] M. e. L. F. SILVA, R. e LEAL, "Predição do câncer de mama com aplicações de modelos de inteligência computacional.," *Tendências em Matemática Aplicada e Computacional.*, vol. 20, no. 2, 2019.

[5] K. D. Cole, H. J. He, and L. Wang, "Breast cancer biomarker measurements and standards," *Proteomics Clin Appl*, 2013.

[6] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310–1315, Ieee, 2016.

[7] F. Pedregosa and et al. (2011), "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] H. L. Hwa and et al, "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models," *J Eval Clin Pract*, 2008.

[9] M. Dalamaga and et al, "Serum resistin: a biomarker of breast cancer in postmenopausal women? association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters," *ClinBiochem*, 2013.

[10] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, no. 1, pp. 1–8, 2018.

[11] H. Pham and D. H. Pham, "A novel generalized logistic dependent model to predict the presence of breast cancer based on biomarkers," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 1, p. e5467, 2020.

[12] E. Bisong, *Building machine learning and deep learning models on Google cloud platform*. Springer, 2019.

[13] F. e. A. M. Holsbach, N. e andogliatto, "Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis," *Ciência Saúde Coletiva*, vol. 19, no. 4, pp. 1295–1304, 2014.

[14] E. Z. G. Max, "Seleção de instâncias baseado em aprendizado de métricas para k vizinhos mais próximos," 2016.

[15] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[16] P. Mulak and N. Talhar, "Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset," *International Journal of Science and Research*, vol. 4, no. 7, pp. 2101–2104, 2015.

[17] L. A. Gonzalez, "Regressão Logística e suas Aplicações," Universidade Federal do Maranhão, 2018.

[18] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[19] J. Bonini, "Aplicação de algoritmos de Árvore de decisão sobre uma base de dados de câncer de mama," *Revista ComInG - Communications and Innovations Gazette*, vol. 1, no. 1, pp. 57–67, 2016.

[20] R. Forests and L. Breiman, "Statistics department university of california berkeley," 1999.

[21] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational statistics & data analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[22] M. e. L. F. SILVA, R. LEAL, "Predição do câncer de mama com aplicação do modelo de inteligência computacional," *Tendências em Matemática Aplicada e Computacional*, vol. 20, pp. 229–240, 2021.

[23] C. de, A. de, R. Poppi, and C. Mello, "Redes neurais e suas aplicações em calibração multivariada," *Química Nova*, vol. 24, 12 2001.