Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Técnicas de Aprendizado de Máquina para Classificação de Células Mamárias Cancerígenas

Lucas R. Pereira[1]
IMECC/UNICAMP, Campinas, SP

Resumo.

Este artigo visa contribuir com o trabalho de classificação de células mamárias cancerígenas realizado pelos professores William Nick Street, William H. Wolberg e Olvi L. Mangasaria, utilizando técnicas de aprendizado de máquina modernas para classificação de câncer de mama benigno e maligno. O conjunto de dados contém medições nucleares das células em 569 pacientes, levando em consideração o formato, tamanho e textura, selecionando 10 características para cada núcleo com o valor médio, pior caso e erro padrão de cada medida. Para a classificação utilizamos métodos lineares como a regressão logística e máquina de suporte vetorial linear (SVM, do inglês *support vector machine*) e não lineares como o *gradient boosting* (GB), o qual se baseia em técnicas de *ensemble*. Aplicando os modelos lineares obtemos acurácia média igual a $96,57\%$ para ambos classificadores lineares e $98,27\%$ utilizando GB, evidenciando a alta eficiência de modelos modernos de aprendizado de máquina, o qual é discutido no decorrer do artigo. Esses avanços são promissores para o campo da classificação precoce de câncer de mama e têm o potencial de contribuir para o desenvolvimento de ferramentas mais precisas e confiáveis no diagnóstico dessa doença crítica.

Palavras-chave. Câncer de mama, SVM, Regressão Logística, Boosting, Aprendizado de máquina.

1 Introdução

No Brasil, excluindo os tumores de pele não melanoma, o câncer de mama é o mais incidente em mulheres de todas as regiões, com taxas mais altas nas regiões Sul e Sudeste. Para cada ano do triênio 2023-2025 foram estimados 73.610 casos novos, o que representa uma taxa ajustada de incidência de 41,89 casos por 100.000 mulheres [3].

Para auxiliar a classificação de tal enfermidade, utilizamos de técnicas de aprendizado de máquina como as SVM, regressão logística e GB. Técnicas que foram utilizadas para classificação binária nas categorias maligna e benigna. O conjunto de dados utilizado baseia-se no trabalho *Breast Cancer Wisconsin* [5] desenvolvido pelos professores William Nick Street, William H. Wolberg e Olvi L. Mangasarian, os quais disponibilizaram publicamente os dados que possibilitaram o treinamento dos modelos. No artigo é utilizado o método *Multisurface Method Tree* (MSM-T), alcançando uma acurácia notável de 97% ao empregar apenas três das trinta características disponíveis: textura média, pior área e pior maciez. Com o objetivo de superar esse desempenho, estamos explorando o uso de modelos mais robustos do que o utilizado.

2 Modelo

Para implementação da regressão logística, SVM e GB utilizamos da biblioteca *scikit-learn* integrada ao *Python*, para o treinamento e validação de todos os modelos, adotou-se uma divisão de

[1] l251341@dac.unicamp.br

70% para treinamento e 30% para validação em relação ao conjunto de dados original e implementação do método de validação cruzada, analisando a média e o desvio padrão dentre 10 subdivisões contendo a acurácia, precisão, revocação, *f1-score*, e a área sob a curva ROC. Para visualização do desempenho dos classificadores utilizamos o gráfico da precisão e revocação em função do limiar permitindo visualizar a troca inerente entre essas duas métricas, a medida que se ajusta o limiar, a precisão e a revocação geralmente se movem em direções opostas possibilitando a escolha do limiar para minimizar os falsos negativos, outra medida de desempenho é a curva ROC e a matriz de confusão, a área sob a curva ROC (AUC-ROC) é uma métrica comum para resumir o desempenho global do modelo. Quanto maior a AUC-ROC, melhor o modelo é em distinguir entre as classes. Um valor de AUC-ROC de 0,5 indica que o modelo não tem capacidade discriminativa, enquanto um valor de 1,0 representa um modelo perfeito, a matriz de confusão apresentar os casos que foram classificados corretamente e incorretamente, permitindo ter noção do quanto o classificador está acertando e errando. Por fim, utilizando características que não foram implementadas diretamente ao treinamento do modelo temos a previsão dos casos de câncer em relação ao raio e perímetro médio nuclear da célula. Para uma análise mais eficiente precisa do classificador, utilizamos o método de validação cruzada para a acurácia, precisão, revocação, *f1-score*, AUC-ROC. Para implementação da técnica utilizamos 10 divisões de validação cruzada ajudando a reduzir a variabilidade associada a uma única divisão aleatória dos dados em conjuntos de treinamento e validação. A validação cruzada ajuda a identificar se o modelo está com *overfitting*[2] em relação aos dados.

2.1 Regressão Logística e Máquina de suporte vetorial

Implementando a regressão logística, realizamos o treinamento com o auxílio do parâmetro de regularização *Ridge*, a regularização *Ridge* contribui para reduzir o *overfitting*, que é o fenômeno em que um modelo se ajusta demais aos dados de treinamento e, consequentemente, não generalizando bem para novos dados, ainda, assim, utilizamos o *solver*[3] *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS) pois permiter suporte a regularização *Ridge*, que é robusto em relação a *outliers*[4] e não assume uma distribuição específica dos dados, este *solver* tem alto desempenho e convergência rápida em muitos casos e oferece uma alta precisão na estimativa dos coeficientes do modelo, isso é particularmente útil quando a precisão dos coeficientes é crítica, como em aplicações médicas. Por fim a técnica de regressão logística foi treinada variando o parâmetro parâmetro de regularização C da através da ferramenta *GridSearch*[5], obtendo o valor $C = 26.5$ como melhor parâmetro de regularização o qual desempenhou um ótimo resultado no conjunto de validação.

Para o classificador da regressão logística, dado o conjunto de treinamento $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{398}, y_{398})\} \subset \mathbb{R}^3 \times \{0, 1\}$, o vetor de parâmetro $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3) \in \mathbb{R}^{3+1}$ da regressão logística pode ser determinado minimizando a entropia binária cruzada dada por

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{3}\sum_{i=1}^{3}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \tag{1}$$

em que $\hat{p}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ representa a probabilidade estimada pelo modelo de $\boldsymbol{x}_i$ pertencer a classe 1, onde $f_{\boldsymbol{\theta}}(x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3)$ em que a função logística $\sigma : \mathbb{R} \to [0, 1]$ é dada por

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2}$$

[2] *Overfitting* é um fenômeno em modelagem estatística e aprendizado de máquina no qual um modelo se ajusta excessivamente aos dados de treinamento.

[3] *Solver* é um método computacional projetado para resolver um problema matemático ou otimização.

[4] *Outliers* são pontos de dados que se afastam significativamente do padrão geral de um conjunto de dados.

[5] *GridSearch* é uma técnica usada em aprendizado de máquina para encontrar os melhores hiperparâmetros para um modelo, que são configurações ajustáveis que não são aprendidas pelo modelo durante o treinamento

Note que a probabilidade de $\boldsymbol{x}$ pertencer a outra classe é $1 - f_\theta(\boldsymbol{x})$. Logo podemos definir um classificador binário $\phi : \mathbb{R}^n \to \{0, 1\}$ como segue:

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \geq 0.5 \\ 0 & \text{caso contrário} \end{cases} \tag{3}$$

A SVM linear define um classificador cuja função de decisão é caracterizada pelo hiperplano que maximiza a margem de separação entre as classes. Utilizando o conjunto $\mathcal{T} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, 398\} \subset \mathbb{R}^3 \times \{-1, 1\}$, vamos assumir que os pares $(\boldsymbol{x}_i, y_i)$ são linearmente separáveis. A SVM linear é obtida maximizando a margem de separação entre as classes, o vetor de pesos $\boldsymbol{w} \in \mathbb{R}^3$ e o viés $b \in \mathbb{R}$ são determinados resolvendo o problema de otimização:

$$\begin{cases} \text{maximize}_{\boldsymbol{w},b}\ 2r \\ \quad \text{sujeito à} \qquad \boldsymbol{w}^T\boldsymbol{x}_i + b \geq r, \forall \boldsymbol{x}_i \in \mathcal{C}^+ \\ \qquad\qquad\quad\ \ \boldsymbol{w}^T\boldsymbol{x}_i + b \leq -r, \forall \boldsymbol{x}_i \in \mathcal{C}^- \end{cases} \tag{4}$$

onde $\mathcal{C}^+ = \{\boldsymbol{x}_i : y_i = +1\}$, $\mathcal{C}^- = \{\boldsymbol{x}_i : y_i = -1\}$ e $r > 0$ define a margem de separação que pode ser escrita em termos de $\boldsymbol{w}$ e b:

$$\begin{cases} \text{minimize}_{\boldsymbol{w},b}\ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ \quad \text{sujeito à} \qquad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad \forall\ i = 1, \ldots, m \end{cases} \tag{5}$$

As amostras $\boldsymbol{x}_i$ tais que $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$ são chamado **vetores de suporte**, pois determinam o hiperplano ótimo.

Implementando ambos classificadores, obtemos resultados iguais que estão amostrados nas Tabelas 1 à 4 e na Figura 1, em que (C.V.) e (C.T.) representam o conjunto de validação e treinamento respectivamente.

Tabela 1: Desempenho médio da regressão logística utilizando a validação cruzada.

Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$96,57\%$	$97,17\%$	$98,18\%$	$0,9657$	$0,9859$	Validação
$91,97\%$	$91,81\%$	$95,58\%$	$0,9197$	$0,9775$	Treinamento

Tabela 2: Desvio padrão da regressão logística utilizando a validação cruzada.

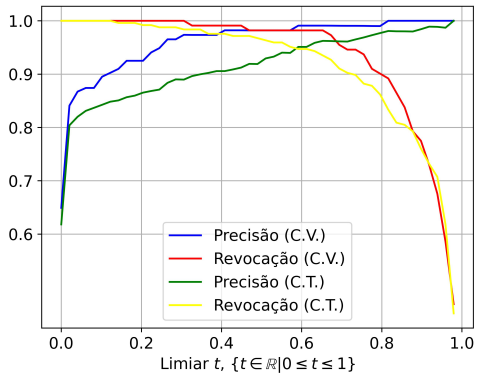
Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$5,77\%$	$6,24\%$	$5,45\%$	$0,0577$	$0,0336$	Validação
$4,43\%$	$2,69\%$	$6,31\%$	$0,0443$	$0,0220$	Treinamento

Tabela 3: Desempenho médio da SVM utilizando a validação cruzada.

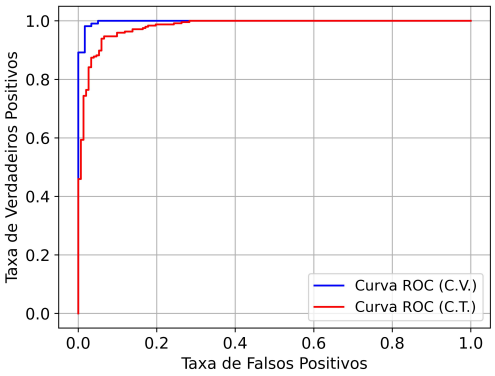
Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$96,57\%$	$97,17\%$	$98,18\%$	$0,9657$	$0,9859$	Validação
$91,22\%$	$90,40\%$	$95,98\%$	$0,9122$	$0,9704$	Treinamento

Tabela 4: Desvio padrão da regressão logística utilizando a validação cruzada.

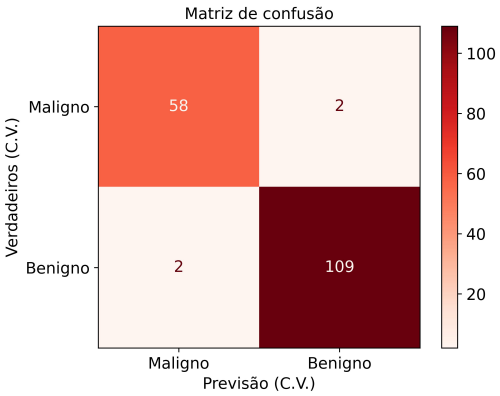
Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$5,77\%$	$6,24\%$	$5,45\%$	$0,0577$	$0,0336$	Validação
$4,63\%$	$2,66\%$	$6,20\%$	$0,0463$	$0,0239$	Treinamento



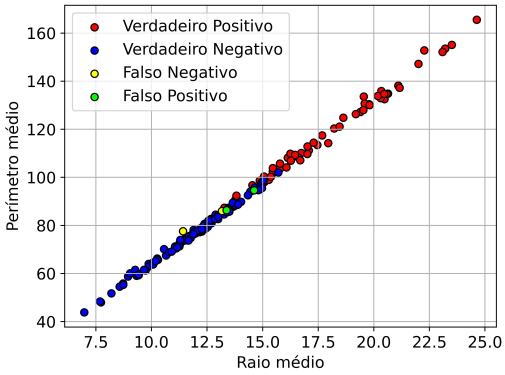
(a) Desempenho classificadores, evidenciando as métricas de precisão e revocação em relação a diferentes valores do limiar de decisão t.



(b) Curva ROC utilizada para medir o desempenho geral do classificador, essas curvas apresentam o valor da área igual a 0,974 e 0,919 para o C.V. e C.T. respectivamente.



(c) Matriz de confusão utilizada na análise de classificações do C.V., proporcionando uma visão detalhada da precisão das classificações, incluindo tanto as instâncias corretamente classificadas quanto as incorretamente classificadas.



(d) O gráfico apresenta o perímetro médio em relação ao raio médio de células. As células malignas corretamente classificadas estão destacadas em vermelho, as células benignas corretamente classificadas em azul, as células incorretamente classificadas como benignas em amarelo, e as células incorretamente classificadas como malignas em cor lima.

Figura 1: Implementação da regressão logística para classificação binária envolvendo células cancerígenas.

A Figura 1a apresenta o desempenho do classificador, a alta precisão evita a classificação incorreta de exemplos malignos como benignos e a alta revocação é importante pois indica que o modelo está minimizando a quantidade de casos de câncer de mama maligno que estão passando despercebidos (falsos negativos). No contexto de um classificador de câncer de mama, ter uma alta taxa de revocação é geralmente desejável, uma vez que é mais crítico perder casos de câncer maligno, mesmo que isso signifique que algumas das classificações sejam falsos positivos (ou seja, classificando algumas células benignas como malignas), em um sistema de detecção de doenças, é

essencial que o modelo identifique corretamente o máximo de casos positivos possível, mesmo que isso signifique aceitar alguns falsos positivos, indicando que o limiar eficiente em nossa classificação entre 0,4 e 0,6. Na Figura 1b temos a curva ROC, em nossos experimentos, obtemos o valor da área igual a 0,974 no C.V. o que evidência o alto desempenho do classificador. Tanto na Figura 1c quando na Figure 1d temos casos positivos para câncer maligno sendo classificados como negativos, partindo para a necessidade de um modelo ainda mais robusto e eficiente.

Utilizando a validação cruzada, temos as Tabelas 1 à 4, vemos que todos os parâmetros se mantiveram iguais tanto para a regressão logística quando para SVM se mantendo alto mas ainda insuficiente não conseguindo superar o valor bem estabelecido de 97%, para tal temos o próximo modelo.

2.2 Gradient Boosting

Partindo para um modelo mais robusto para classificação de características lineares e não lineares temos o *Gradient Boosting*, que consiste na construção de um *ensemble*[6] de forma sequencial na tentativa de melhorar o desempenho do modelo antecessor, no GB os preditores são ajustados usando o resíduo (diferença entre a previsão e o valor correto) do anterior, uma vez que todos classificadores foram treinados, suas predições são então combinadas através do voto da maioria [1], segue a demonstração do modelo apresentada em [2].

Dado o C.T. $\mathcal{T} = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, 398\} \subset \mathbb{R}^3 \times \{-1, 1\}$, utilizamos 100 árvores de decisão. De forma similar a regressão logística, a previsão do ensemble da árvore de decisão é realizada utilizando a função *sigmoid*.

$$P(y = 1|\mathbf{x}, f) \stackrel{\text{def}}{=} \frac{1}{1 + e^{-f(\mathbf{x})}} \tag{6}$$

onde $f(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{m=1}^{100} f_m(\mathbf{x})$ e f_m é a árvore de regressão.

O algoritmo começa com o modelo inicial constante $f = f_0 = \frac{p}{1-p}$, onde $p = \frac{1}{100}\sum_{i=1}^{100} y_i$. Então a cada iteração, m, uma nova árvore f_m é adicionada ao modelo. Para encontrar o melhor f_m, a derivada parcial g_i do modelo atual é calculada para cada $i = 1, \ldots, 100$, $g_i = \frac{dL_f}{df}$, onde f é o modelo de classificador *ensemble* construído com a iteração anterior $m - 1$. Calculando a derivada, obtemos $f = \frac{1}{e^{f(\mathbf{x_i})}+1}$, então substituímos y_i com a correspondente derivada parcial g_i, construímos uma nova árvore f_m. Então encontramos o passo ótimo ρ_m como:

$$\rho_m \leftarrow \arg\max_{\rho} L_{f+\rho f_m} \tag{7}$$

No fim da iteração m, adicionamos ao modelo f adicionando a nova árvore f_m. Iteramos até $m = 100$, então paramos e retornamos o modelo de *ensemble* f.

$$f \leftarrow f + \alpha\rho_m f_m \tag{8}$$

Computando o resíduo, podemos encontra quão bem (ou mal) a característica de cada amostra está desempenhando pelo modelo atual f. Os três principais hiperparâmetros no GB são, o número de árvores (m), taxa de aprendizado (α) e quão profundo são as árvores. Essas três características afetam diretamente a acurácia do modelo.

Implementando o classificador, temos:

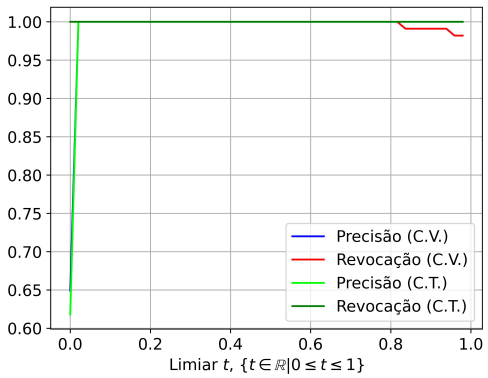
Tabela 5: Desempenho médio do GB utilizando a validação cruzada.

Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$98,27\%$	$97,56\%$	100%	$0,9827$	$0,9985$	Validação
$94,48\%$	$94,48\%$	$96,78\%$	$0,9448$	$0,9757$	Treinamento

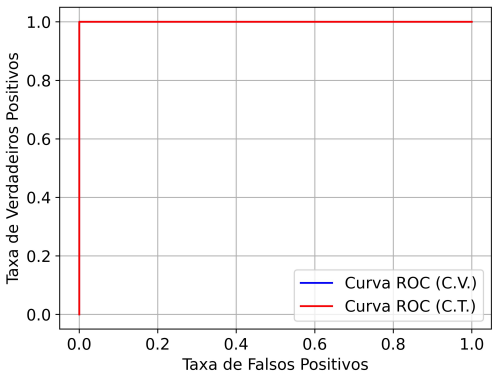
[6] *Ensemble* é um comitê de modelos de aprendizado de máquina que apresenta uma predição de forma coletiva.

Tabela 6: Desvio padrão do GB utilizando a validação cruzada.

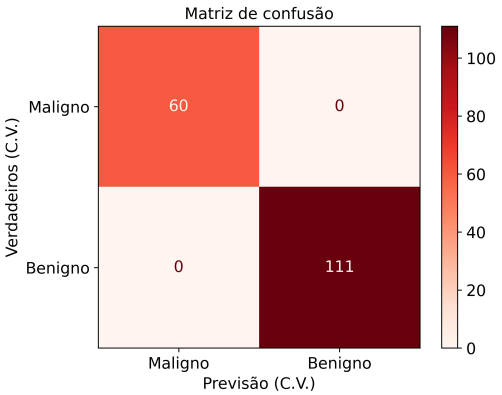
Acurácia	Precisão	Revocação	F1-Score	ROC-Score	Conjunto
$2,65\%$	$3,72\%$	0%	$0,0265$	$0,0045$	Validação
$3,83\%$	$2,48\%$	$5,00\%$	$0,0383$	$0,0283$	Treinamento



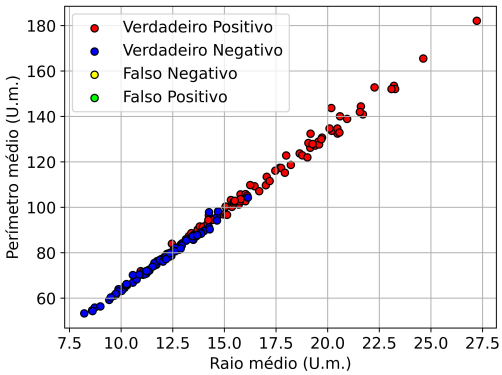
(a) Desempenho classificadores, evidenciando as métricas de precisão e revocação em relação a diferentes valores do limiar de decisão t.



(b) Curva ROC utilizada para medir o desempenho geral do classificador, está curva apresenta uma área igual a 1.0 para o C.V. e C.T..



(c) Matriz de confusão utilizada na análise de classificações do C.V., proporcionando uma visão detalhada da precisão das classificações, incluindo tanto as instâncias corretamente classificadas quanto as incorretamente classificadas.



(d) O gráfico apresenta o perímetro médio em relação ao raio médio de células. As células malignas corretamente classificadas estão destacadas em vermelho, as células benignas corretamente classificadas em azul e não apresenta nenhuma classificação incorreta.

Figura 2: Implementação do GB para classificação binária envolvendo células cancerígenas.

Assim como nos modelos anterior, a Figura 2a apresenta o desempenho do classificador, apresentando a precisão e revocação quase que constante para todos os pontos do limiar t, uma precisão e revocação de 1 para todos os limiares é incomum e geralmente não reflete a realidade tem cons-

ciência de tal fenômeno, utilizamos a validação cruzada para verificar a vericidade do resultado obtido, vemos pela Tabela 5 que a precisão foi um pouco abaixo de 100% porém se manteve muito alta, com precisão igual a $97,56\%$ e revocação igual a 100% tais resultados demonstram a confiabilidade do classificador. Na Figura 2b temos a curva ROC, em nossos experimentos, obtemos o valor da área igual a 1.0 no C.V. e aplicando a validação cruzada obtemos a curva ROC igual a 0.9985 com desvio padrão de $0,0045$ reafirmando o alto desempenho do classificador. Tanto na Figura 2c quando na Figure 2d não temos casos positivos para células cancerígenas maligna sendo classificados como negativas nem casos de células cancerígenas benignas sendo classificadas como positivas, obtendo a partir da validação cruzada a acurácia igual a $98,27\%$ com desvio padrão igual a $2,65\%$ conseguindo superar mesmo que de forma discreta o modelo bem estabelecido dos professores.

3 Conclusão

Neste estudo, investigamos a aplicação de diferentes algoritmos de aprendizado de máquina — regressão logística, máquina de suporte vetorial linear e *gradient boosting* — na tarefa de classificação de câncer de mama. Todos os classificadores apresentaram desempenho consistente, destacando-se o *gradient boosting*, que superou o trabalho de referência *Nuclear features extraction for breast tumor diagnosis* [4], o qual reportou acurácia de 97%. Em nossa abordagem, o modelo obteve acurácia de $98,27\%$ e pontuação ROC de $0,9985$, evidenciando sua robustez e precisão.

Esses resultados indicam que os modelos propostos são capazes de realizar classificações confiáveis quanto à gravidade do câncer, distinguindo de forma eficaz entre tumores benignos e malignos a partir de um método não invasivo. A superioridade do *gradient boosting* pode ser explicada por sua capacidade de lidar com conjuntos de dados complexos e não lineares, adaptando-se de maneira dinâmica aos padrões subjacentes. Por outro lado, embora mais simples, a regressão logística e a máquina de suporte vetorial linear também alcançaram desempenho expressivo, com acurácia de $96,57\%$, sugerindo a existência de certo grau de linearidade no conjunto de dados analisado.

Assim, este estudo não apenas demonstra a eficiência dos modelos desenvolvidos, mas também reforça a relevância contínua do aprendizado de máquina no aprimoramento de métodos de diagnóstico médico não invasivos, oferecendo soluções de alta eficiência, confiabilidade, segurança e rapidez. Tais avanços contribuem diretamente para a detecção precoce do câncer de mama, potencialmente beneficiando um maior número de pacientes e impactando positivamente a qualidade de vida da população.

Referências

[1] Christopher M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 657–660. ISBN: 0387310738.

[2] A. Burkov. **The Hundred-Page Machine Learning Book**. Andriy Burkov, 2019, pp. 82–87. ISBN: 9781999579517. URL: `https://books.google.com.br/books?id=0jbxwQEACAAJ`.

[3] INCA. **Dados e Números sobre Câncer de Mama - Relatório Anual 2023**. Online. Acessado em 24/10/2023, `https://www.inca.gov.br/publicacoes/relatorios/dados-e-numeros-sobre-cancer-de-mama-relatorio-anual-2023`.

[4] Olvi L. Mangasarian, W. Nick Street e William H. Wolberg. "Breast Cancer Diagnosis and Prognosis via Linear Programming". Em: **Operations Research** 43.4 (1995), pp. 570–577. ISSN: 0030364X, 15265463. URL: `http://www.jstor.org/stable/171686` (acesso em 24/10/2023).

[5] Mangasarian Olvi Wolberg William e Street Nick. **Breast Cancer Wisconsin (Diagnostic)**. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B. 1995.