Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Predição do Câncer de Mama utilizando Rede Naive Bayes Gaussiana com Parâmetros *Fuzzy*

Luiz Henrique da Silva[1], Ronei Marcos de Moraes[2]
PPGMDS/UFPB, João Pessoa, PB

Resumo. A mamografia é o método mais eficaz para o rastreamento e identificação do câncer de mama precocemente. No entanto, sua interpretação errônea pode ocasionar procedimentos desnecessários. Algoritmos computacionais podem subsidiar a tomada de decisão médica e evitar tais procedimentos e gastos desnecessários ao sistema de saúde. Objetivou-se aplicar uma Rede Naive Bayes Gaussiana com parâmetros *Fuzzy* para predição da gravidade de lesões mamográficas. A proposta do presente estudo foi utilizar uma nova rede baseada em probabilidade *Fuzzy*, cujo treinamento e validação se deu por meio de conjunto de dados de domínio público. A avaliação da rede ocorreu segundo a análise competitiva com outras redes, com base na Acurácia e Coeficiente Kappa. Os resultados encontrados indicam que a Rede gerou resultados competitivos quando comparados com outros métodos presentes na literatura.

Palavras-chave. Redes Bayesianas, Probabilidade *Fuzzy*, Diagnóstico Precoce, Neoplasias Mamárias, Machine Learning.

1 Introdução

O câncer é uma grave questão de saúde pública em escala global e, particularmente, o Câncer de Mama (CM) é considerado o mais prevalente entre as mulheres de todas as faixas etárias no mundo [6]. No Brasil, a incidência na população feminina se mantém elevada, e exceto os tumores de pele não melanoma, o CM ocupa a primeira posição mais frequente em todas as Regiões do Brasil [13].

As estratégias para a detecção precoce do CM são subdivididas em basicamente duas formas: o diagnóstico precoce e o rastreamento, ambas objetivam identificar alterações sugestivas de câncer e colaborar com o tratamento. Nesse contexto, a mamografia é o método padrão ouro de rastreamento no Brasil, sendo indicada sua realização bienal em mulheres de 50 a 69 anos [1]. No entanto, na literatura, radiologistas e profissionais da saúde demonstram variação considerável na sua interpretação, o que significa, por vezes, procedimentos desnecessários que causam dor e desconforto ao paciente, além de gastos ao sistema de saúde [16].

Assim, identifica-se a necessidade do desenvolvimento de métodos que subsidiem no reconhecimento precoce do CM. Logo, o diagnóstico auxiliado por sistemas computacionais pode ajudar a reduzir o número de interpretações errôneas de mamografia e, portanto, reduzir consequências biopsicossociais ao paciente. Algoritmos de aprendizado de máquina têm se destacado como soluções potencialmente viáveis, principalmente por sua fácil aplicabilidade e poder de inferência em diversos cenários reais.

Estudos têm apresentado bons resultados na predição do CM baseados em algoritmos de aprendizado de máquina, como por exemplo [5], No entanto, há necessidade da comparação dos resultados

[1]luizenf2014.2@gmail.com
[2]ronei@de.ufpb.br

desses algoritmos com outros métodos afim de determinar qual melhor decisão a ser tomada. Além disso, esses algoritmos são capazes de extrair padrões e detectar tendências que são complexas para serem notadas por humanos ou outras técnicas convencionais [7].

Dessa forma, o presente estudo utilizou uma Rede Naive Bayes Gaussiana com Parâmetros Fuzzy (NBG-PF) e comparou seus resultados com outras redes presentes na literatura. Como aspecto inovador, destaca-se a ausência de estudos na literatura pesquisada sobre a aplicação da rede NBG-PF diante a problemática do CM no Brasil.

O conjunto de dados mamográficos investigados neste estudo foi coletado no Instituto de Radiologia do Universidade Erlangen-Nuremberg entre 2003 e 2006, e disponibilizado no *UCI Machine Learning Repository* [8]. Diversos métodos já foram empregados neste conjunto de dados e, portanto, alguns deles foram utilizados para comparação com os resultados do presente trabalho. Uma contribuição essencial deste estudo é prover uma nova metodologia que trate a incerteza do conhecimento presente na medição das variáveis relacionadas ao CM, além da validação da rede NBG-PF para estudos futuros.

2 Fundamentação Teórica

2.1 Diagnóstico precoce e Rastreamento do câncer de mama

As principais ferramentas para diagnóstico da doença são a mamografia e o exame clínico, além de outras como ultrassonografia, ressonância, exames de sangue, raio-X, biópsia, exames citopatológico e histopatológico, no entanto, apesar de todos essas metodologias de diagnóstico, o principal desafio se configura em ter um diagnóstico precoce da doença [12].

Ainda mais, independente da faixa etária da mulher, a educação em saúde se traduz como uma importante ferramenta para potencialização o reconhecimento dos sinais e sintomas suspeitos de CM, além do esclarecimento acerca do fluxo nos serviços de saúde. Os principais sinais e sintomas estão relacionados com alterações na pele da mama, nódulos anormais, lesões sem motivo prévio e aumento excessivo associado a presença de edema [9].

Associado ao diagnóstico precoce, o rastreamento traz diversos aspectos positivos, por exemplo, o melhor prognóstico da doença e diminuição da morbimortalidade. É sabido que regiões onde há presença de rastreamento eficiente a mortalidade por CM tem diminuído [15]. No entanto, uma má interpretação de exames, como a mamografia, pode ocasionar desde episódios de ansiedade e pânico até procedimentos desnecessários no paciente.

Dessarte, considera-se como faixa etária de alto risco mulheres entre 50 e 69 anos de idade. Logo, estudos comprovam que o fator genético também possui forte relação com o surgimento do CM, devendo-se a avaliação ser realizada com cautela, considerando as características individuais de cada mulher [15]. A continuidade do cuidado inclui o retorno da mulher ao serviço de saúde com o laudo da mamografia e, se há indícios para suspeita da patologia, realiza-se o encaminhamento para investigação diagnóstica [1].

2.2 Rede Gaussiana Naive Bayes com Parâmetros *Fuzzy*

α-corte Esta sessão descreve a Rede NBG-PF, utilizada com o objetivo de classificar pacientes em dois grupos: benigno ou maligno, a mesma foi constituída baseada na teoria proposta por [2]. A rede é fundamentada na distribuição gaussiana, e seus parâmetros são descritos por números *Fuzzy*. Formalmente, sabendo que o evento A pertence a uma variável aleatória $X \in R^n$, com função de distribuição dada por $f_x(x, \Theta)$, onde Θ representa um vetor com parâmetros *Fuzzy*, e sabendo que as variáveis aleatórias modeladas estatisticamente pela rede se aproximam da distribuição gaussiana

DOI: 10.5540/03.2025.011.01.0461

Fuzzy, dada por $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, cujos parâmetros são números *Fuzzy*, para um α-corte $\in [0, 1]$, $\mu \in \tilde{\mu}[\alpha]$ e $\sigma^2 \in \tilde{\sigma}^2[\alpha]$.

Segundo [11], é necessário utilizar a distribuição gaussiana com média zero e variância um para calcular a probabilidade da variável X assumir valores entre $[c, d]$

$$\mathcal{P}(c \leq X \leq d)[\alpha] = \left\{ \int_{z_1}^{z_2} f_x(x;\, 0,\, 1)\, dx \,|\, \mu \in \tilde{\mu}[\alpha],\, \sigma^2 \in \tilde{\sigma}^2[\alpha] \right\} \tag{1}$$

onde: $z_1 = \frac{c-\mu}{\sigma}$; $z_2 = \frac{d-\mu}{\sigma}$ e $f(x;\, 0, 1)$ representam a densidade gaussiana com média zero e variância um.

A Equação (1) fornece os α-cortes necessários para $\mathcal{P}(c \leq X \leq d)[\alpha]$, gerando como resultado um número *Fuzzy* de formato triangular. Dessa maneira, do ponto de vista *Fuzzy*, a saída é a fuzzyficação do valor clássico, denodado por $\tilde{M}$ e definido por:

$$\tilde{M}[\alpha] = \left\{ \int_{-\infty}^{+\infty} x f_x(x;\, \mu,\, \sigma^2)\, dx \,|\, \mu \in \tilde{\mu}[\alpha],\, \sigma^2 \in \tilde{\sigma}^2[\alpha] \right\}. \tag{2}$$

Similarmente, a variância denotada por $\tilde{V}$ é definida como:

$$\tilde{V}[\alpha] = \left\{ \int_{-\infty}^{+\infty} (x-\mu)^2 f_x(x;\, \mu,\, \sigma^2)\, dx \,|\, \mu \in \tilde{\mu}[\alpha],\, \sigma^2 \in \tilde{\sigma}^2[\alpha] \right\}. \tag{3}$$

A partir da hipótese Naive Bayes, assumindo forte independência para a variável aleatória X, a função discriminante g que define a Rede GNB-PF é dada por:

$$\begin{aligned} g(w_i, X_1, X_2, \ldots, X_n) = log[\mathcal{P}(w_i) \,|\, X_1, X_2, \ldots, X_n)] = \\ log\left[\frac{1}{S} \mathcal{P}(w_i) \prod_{k=1}^{n} \mathcal{P}(X_k|w_i) \right] = \\ log\left(\frac{1}{S}\right) + log[\mathcal{P}(w_i)] + \sum_{k=1}^{n} log[\mathcal{P}(X_k|w_i)] \end{aligned} \tag{4}$$

onde: $\frac{1}{S}$ é constante e não relevante no processo de decisão. O segundo fator é a probabilidade da classe w_i, que pode ser facilmente estimada pela razão entre o número de elementos nos dados da amostra que pertencem a essa classe e o total de elementos na amostra, o que equivale à estimativa de probabilidade para um evento clássico.

Assim, se as probabilidades *a priori* $\mathcal{P}(w_i)$ forem as mesmas para todas as possíveis decisões, esse fator também não é relevante. Para o terceiro fator dessa equação, é utilizado o valor da densidade gaussiana, cujos α-cortes são dados por:

$$\mathcal{P}(X_k \,|\, w_i)[\alpha] = \left\{ f_X, w_i(x; 0, 1) \,|\, \mu \in \tilde{\mu}[\alpha],\, \sigma^2 \in \tilde{\sigma}^2[\alpha] \right\} \tag{5}$$

onde: $x = \frac{X_k}{\mu}$, $f(x; 0, 1)$ significa a densidade gaussiana padrão com média zero e variância um e $0 \leq \alpha \leq 1$. A utilização do logaritmo diminui o custo computacional e simplifica as equações anteriores, obtendo-se:

$$
\begin{aligned}
log\mathcal{P}(X_k \mid w_i)[\alpha] = log\big\{ f_X, w_i(x;0,1) \mid \mu \in \tilde{\mu}[\alpha],\, \sigma^2 \in \tilde{\sigma}^2[\alpha] \big\} \\
log\left\{ \frac{1}{\sqrt[2]{2\pi\sigma_j}}\, exp\left[-\frac{(X_j-\mu_j)^2}{2\sigma_j} \right] \right\}[\alpha] = \\
log(1) - log[\sqrt[2]{2\pi\sigma_j}] - log[\sqrt[2]{\sigma_j}][\alpha] + \\
\left\{ -\frac{X_j-\mu_j)^2}{2\sigma_j} \right\}[\alpha].
\end{aligned}
\tag{6}
$$

Como já citado anteriormente, as constantes presentes na equação (6) não são relevantes, então pode-se simplificar a função discriminante g:

$$
g(w_i, X_1, X_2, \ldots, X_n)[\alpha] = log[\mathcal{P}(w_i)][\alpha] + \left[-\frac{(X_j-\mu_j)^2}{2\sigma_j} \right][\alpha] - log[\sqrt[2]{\sigma_j}][\alpha], \tag{7}
$$

admitindo-se que as probabilidades *a priori* $\mathcal{P}(w_i)$ possam ser diferentes para todas as possíveis decisões, o que é o caso mais geral. Portanto, a regra de decisão para a rede GNB-PF pode ser dada por:

$$
X \in w_i \;\; SE\, g(w_i, X_1, X_2, \ldots, X_n)[\alpha] > g(w_j, X_1, X_2, \ldots, X_n) \tag{8}
$$

para todo $i \neq j$, onde i e j são subconjuntos de Ω e $g(.)[\alpha]$ são computados pela Equação (7)

3 Materiais e Métodos

O conjunto de dados, *Mammographic Mass*, consiste em 961 instâncias com 516 classificados como benignos e 445 como malignos, com a avaliação do BI-RADS, a idade do paciente e três atributos do BI-RADS, juntamente com a classe, ou seja, massas do tipo benignas ou malignas identificadas em mamografias digitais coletadas no Instituto de Radiologia da Universidade Erlangen-Nuremberg [8].

Utilizou-se a Rede NBG-PF descrita na sessão (2) por meio do pacote *FuzzyClass* [3], presente no *Software* R e disponível em `<https://cran.r-project.org/web/packages/FuzzyClass/index.html>`. Comparou-se os resultados encontrados com outras redes presentes na literatura, a saber: Rede Naive Bayes (NB), Rede Fuzzy Gaussiana Naive Bayes (FGNB), Rede Fuzzy Exponencial Naive Bayes (FENB) e Rede Neural Multilayer Perceptron (RNMP).

O desempenho de classificação da GNB-PF foi avaliado por meio da Acurácia e Coeficiente Kappa. A acurácia é um bom preditor para o grau de correção no treinamento do modelo e, de modo geral, como o modelo pode se comportar. Pode ser definida como a medida da correta predição em relação aos erros. Formalmente:

$$
Acurácia = \frac{VP + VN}{(VP + FP + VN + FN)} \tag{9}
$$

onde: VP = Verdadeiro Positivo; VN = Verdadeiro Negativo; FP = Falso Positivo; FN = Falso Negativo [14].

O coeficiente de Kappa indica a concordância das interpretações pela matriz de confusão, sua escala é definida de acordo com os intervalos: 0,01 a 0,20, levemente concordante; 0,21 a 0,40 razoavelmente concordantes; 0,41 a 0,60 moderadamente concordantes; 0,61 a 0,80 substancialmente concordantes e 0,81 a 0,99 quase perfeitamente concordantes [4].

O coeficiente de Kappa é dado da seguinte forma:

$$Kappa = \frac{P(O) - P(E)}{1 - P(E)} \tag{10}$$

onde: $P(O)$ = proporção observada de concordâncias (soma das respostas concordantes dividida pelo total); $P(E)$ = proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total) [4].

4 Resultados e Discussões

Aplicou-se a Rede NBG-PF ao conjunto de dados *Mammographic Mass*, onde dividiu-se o banco em 70% para treinamento (672 casos) e 30% para teste (289 casos), com dois desfechos rotulados como 0 = Benigno e 1 = Maligno, e seus resultados foram comparados com outras redes presentes na literatura.

Os resultados apresentados na Tabela 1 demonstram que a Rede NBG-PF apresentou resultados competitivos, mas a RNMP possuiu maiores acurácia e Coeficiente Kappa.

Tabela 1: Performance das redes analisadas.

Redes	Acurácia (%)	Kappa (%)
NBG-PF	84,08	68,02
NB	77,57	54,76
FGNB	80,28	60,46
FENB	50,17	3,0
RNMP	85,42	70,8

Ao comparar os resultados, foi possível perceber que a rede NBG-PF obteve melhores resultados do que as redes NB, FGNB e FENB. Porém possuiu resultado ligeiramente inferior quando comparado com a RNMP.

A Rede NBG-PF avaliou corretamente o maior quantitativo de acertos da decisão "Benigno" em relação as outras redes. No entanto, seu desempenho foi inferior a duas redes em relação a decisão "Maligno". Dessa forma, é plausível afirmar que a rede obteve resultados competitivos em comparação com as demais analisadas. Ressalta-se que há possibilidade da rede NBG-PF obter melhores resultados em outros cenários como, por exemplo, uma análise por meio de um conjunto de dados constituído por variáveis com distribuição que se aproxime da normal.

Os autores em [11] argumentam ainda que características intrínsecas ao método podem influenciar no resultado final, o que significa dizer que questões como o formato do número *Fuzzy*, dado o número *Fuzzy* possui o conhecimento a respeito da informação tratada, utilizado pela rede e/ou a dependência em relação aos métodos de ordenação podem afetar o resultado.

Os resultados da presente pesquisa foram competitivos quando comparados com estudos encontrados na literatura como, por exemplo, em estudo [10], que utilizou uma Árvore de Decisão, Rede Neural Artificial e Máquinas de Vetores de Suporte para classificação de mamografias com o intuito de identificar a decisão "Benigno" ou "Maligno". Os resultados em termos de acurácia foram 78,12%, 80,56% e 81,25% respectivamente. A acurácia da Rede NBG-PF, utilizada na presente pesquisa, foi de 84,08%. Além disso, o treinamento da Rede NBG-PF se dá em apenas um passo, enquanto que nos outros métodos se dá de forma interativa.

O presente estudo, possuiu como aspecto inovador a utilização de uma Rede NBG-PF direcionada à detecção do CM, que por sua vez demonstrou resultados competitivos com outras redes encontradas na literatura pesquisada. Logo, a rede possui o potencial para subsidiar a tomada de

decisão médica com o intuito de reduzir os gastos em saúde e número de procedimentos desnecessários.

5 Considerações Finais

No presente artigo, identificou-se dois desfechos possíveis, "Benigno" e "Maligno", a partir de massas mamográficas, com base na aplicação de uma Rede NBG-PF, utilizando o *software* R e o pacote *FuzzyClass*. Observou-se que a Rede NBG-PF gerou resultados satisfatórios com uma acurácia de 84,08%, e Coeficiente Kappa de 68,02%.

Quando comparado com outras redes, seu desempenho foi considerado competitivo, tendo como vantagem o maior acerto na decisão "Benigno" e seu treinamento em apenas um passo. Como proposta futura, vale destacar a aplicação da rede com a seleção de atributos do conjunto de dados, buscando uma melhor descrição do desfecho.

Como limitação, ressalta-se que neste artigo o foco principal foi apenas em identificar o desfecho em benigno ou maligno, sem considerar variáveis indiretamente relacionadas e fatores inerentes à coleta dos dados.

Agradecimentos

Agradecemos à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo apoio financeiro fornecido para esta pesquisa.

Referências

[1] Brasil, Ministério da Saúde e Instituto Nacional de Câncer José Alencar Gomes da Silva. **Diretrizes para a detecção precoce do câncer de mama no Brasil**. 2015.

[2] J. J. Buckley. **Fuzzy Probability and Statistics**. Vol. 196. Springer-Verlag, 2006, pp. 1–270. ISBN: 3-540-30841-5. DOI: 10.1007/3-540-33190-5.

[3] J. A. Ferreira e R. M. Moraes. "FuzzyClass: A family of Fuzzy and Non-Fuzzy probabilistic-based classifiers". Em: **Journal of Open Source Software** 8 (88 ago. de 2023), p. 5613. ISSN: 2475-9066. DOI: 10.21105/joss.05613.

[4] J. L. Fleiss e J. Cohen. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability". Em: **Educational and Psychological Measurement** 33 (3 out. de 1973), pp. 613–619. ISSN: 0013-1644. DOI: 10.1177/001316447303300309.

[5] P. Gupta e S. Garg. "Breast Cancer Prediction using varying Parameters of Machine Learning Models". Em: **Procedia Computer Science** 171 (2020), pp. 593–601. ISSN: 18770509. DOI: 10.1016/j.procs.2020.04.064.

[6] F. Hauth, C. De-Colle, N. Weidner, V. Heinrich, D. Zips e C. Gani. "Quality of life and fatigue before and after radiotherapy in breast cancer patients". Em: **Strahlentherapie und Onkologie** 197 (4 2021), pp. 281–287. ISSN: 1439099X.

[7] M. I. Jordan e T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". Em: **Science** 349.6245 (2015), pp. 255–260.

[8] E. Matthias. **Mammographic Mass**. UCI Machine Learning Repository. DOI: 10.24432/C53K6Z. 2007.

[9] A. Migowski, G. A Silva, M. K. Dias, M. D. P. E. Diz, D. R. Sant'Ana e P. Nadanovsky. "Diretrizes para detecção precoce do câncer de mama no Brasil. II - Novas recomendações nacionais, principais evidências e controvérsias". Em: **Cadernos de Saúde Pública** 34 (6 jun. de 2018). ISSN: 1678-4464. DOI: 10.1590/0102-311x00074817.

[10] S. A. Mokhtar e A. M. Elsayad. **Predicting the Severity of Breast Masses with Data Mining Methods**. 2013. arXiv: 1305.7057 [cs.LG].

[11] R. M. Moraes, J. A. Ferreira e L. S. Machado. "A New Bayesian Network Based on Gaussian Naive Bayes with Fuzzy Parameters for Training Assessment in Virtual Simulators". Em: **International Journal of Fuzzy Systems** 23 (3 abr. de 2021), pp. 849–861. ISSN: 1562-2479. DOI: 10.1007/s40815-020-00936-4.

[12] A. C. V. Ramos, L. S. Alves, T. Z. Berra, M. P. Popolin, M. A. M. Arcoverde, L. T. Campoy, J. F. Martoreli, L. V. Lapão, P. F. Palha e R. A. Arcêncio. "Estratégia Saúde da Família, saúde suplementar e desigualdade no acesso à mamografia no Brasil". Em: **Revista Panamericana de Salud Pública** 42 (2018). ISSN: 1020-4989. DOI: 10.26633/RPSP.2018.166.

[13] M. O. Santos, F. C. S. Lima, L. F. L. Martins, J. F. P. Oliveira, L. M. Almeida e M. C. Cancela. "Estimativa de Incidência de Câncer no Brasil, 2023-2025". Em: **Revista Brasileira de Cancerologia** 69 (1 fev. de 2023). DOI: 10.32635/2176-9745.rbc.2023v69n1.3700.

[14] J. A. Swets. "Measuring the Accuracy of Diagnostic Systems". Em: **Science** 240 (4857 jun. de 1988), pp. 1285–1293. ISSN: 0036-8075. DOI: 10.1126/science.3287615.

[15] J. G. Tomazelli e G. A. Silva. "Rastreamento do câncer de mama no Brasil: uma avaliação da oferta e utilização da rede assistencial do Sistema Único de Saúde no período 2010-2012*". Em: **Epidemiologia e Serviços de Saúde** 26 (4 nov. de 2017), pp. 713–724. ISSN: 1679-4974. DOI: 10.5123/S1679-49742017000400004.

[16] E. E. Usang, M. Alakhras e P. Brennan. "Errors in mammography cannot be solved through technology alone". Em: **Asian Pacific journal of cancer prevention: APJCP** 19.2 (2018), p. 291.

