Diagnóstico de Câncer de Mama com Aprendizado de Máquina: Estudo de Caso no UCI-Wisconsin

Messias Acacy[1], **Raldnei Miguel**[1], **Luciano de Souza Cabral** [1,2]

[1]Instituto Federal de Pernambuco (IFPE) - Jaboatão dos Guararapes - PE - Brasil

{mafl2,rms88}@discente.ifpe.edu.br, luciano.cabral@jaboatao.ifpe.edu.br

Abstract. *Machine learning has gained prominence in breast cancer detection, driven by the need for more accurate diagnoses. This study compares algorithms such as Logistic Regression (LR), Random Forest (RF), Latent Variable Modeling (LVM), and Artificial Neural Networks (ANN) using the UCI-Wisconsin (Diagnostic) dataset. The analysis employs metrics like accuracy, AUC, sensitivity, and specificity. Results show that LR balances accuracy and interpretability, making it viable for medical applications. The study explores how algorithmic optimization and feature selection impact prediction, contributing to medical informatics by providing a comparative framework. The findings help in selecting effective models, improving diagnostic accuracy, and fostering automated clinical support tools.*

Resumo. *O aprendizado de máquina tem se destacado na detecção do câncer de mama, impulsionado pela necessidade de diagnósticos mais precisos. Este estudo compara algoritmos como Regressão Logística (RL), Floresta Aleatória (RF), Modelagem de Variáveis Latentes (LVM) e Redes Neurais Artificiais (RNA) no conjunto de dados UCI-Wisconsin (Diagnostic). A análise utiliza métricas como acurácia, AUC, sensibilidade e especificidade. Os resultados mostram que a RL equilibra precisão e interpretabilidade, sendo viável para aplicações médicas. O estudo investiga como a otimização algorítmica e a seleção de características impactam a predição, contribuindo para a informática médica ao fornecer um framework comparativo. As descobertas auxiliam na escolha de modelos eficazes, melhorando a precisão diagnóstica e fomentando ferramentas automatizadas de apoio clínico.*

1. Introdução

O câncer de mama é o segundo tipo mais comum entre as mulheres no Brasil e no mundo, configurando-se como um grave problema de saúde pública [Pinto et al. 2018]. A detecção precoce e o tratamento em estágios iniciais são fundamentais, pois aumentam significativamente as chances de cura e reduzem os custos associados ao tratamento.

No entanto, a escassez de dados no Brasil dificulta a formulação de estratégias eficazes, levando a discussões empíricas promovidas por sociedades médicas, mídias, ONGs e universidades públicas. Essa lacuna de informações representa um desafio para legisladores e gestores de saúde na implementação de decisões custo-efetivas que visem à redução da mortalidade por câncer de mama [GEBRIM 2016].

Nesse contexto, a inteligência artificial surge como uma ferramenta promissora, com aplicações que vão desde cuidados de saúde até setores como transporte, finanças

e entretenimento. Embora ofereça benefícios significativos, sua implementação também levanta questões éticas e de segurança que precisam ser cuidadosamente abordadas.

O aprendizado de máquina, uma aplicação da inteligência artificial, permite que algoritmos identifiquem padrões em grandes conjuntos de dados, o que é útil na saúde. Ao melhorar a detecção de tumores, essa técnica pode aprimorar o diagnóstico e tratamento do câncer de mama, beneficiando profissionais e pacientes [Gupta et al. 2021].

Neste estudo, explora-se um conjunto de dados *Breast Cancer Wisconsin (Diagnostic)*[1] contido no repositório *UC Irvine (UCI) Machine Learning Repository*, selecionando as melhores variáveis e analisando a performance de diferentes técnicas de aprendizagem de máquina, com o intuito de mensurar qual seria o modelo de melhor desempenho para o problema em questão. Em adendo, efetuar comparativos com a literatura para averiguar se tal desempenho é competitivo. Com possibilidade de aplicação de técnicas de análise de dados e *tunning* de hiperparâmetros, para tornar os modelos ainda melhores.

As próximas seções abordam Estado da Arte, Revisão da Literatura, com estudos anteriores sobre o desempenho dos modelos e o dataset utilizado. Em seguida, descreve-se a Metodologia, incluindo aquisição, pré-processamento, análise dos dados, modelagem e métricas de avaliação. Logo são apresentados os experimentos, resultados e discussões. Por fim, são apresentadas as contribuições, limitações, trabalhos futuros e as referências.

2. Estado da Arte

A previsão de câncer de mama tornou-se uma área de relevância nos últimos anos, especialmente com os avanços nas tecnologias de aprendizado de máquina e análise de dados.

Pesquisas mostram que métodos como Regressão Logística (LR), SVM, Redes Neurais (ANN) e Random Forest (RF) são eficazes na previsão do diagnóstico do câncer, oferecendo resultados confiáveis [Kourou et al. 2014].

Técnicas que treinam modelos com dados rotulados são comuns, e tanto a Regressão Logística quanto a SVM são reconhecidas por sua eficácia e confiabilidade. Além disso, a melhoria das previsões depende do tratamento adequado dos dados, já que a qualidade dos dados de entrada é crucial para o desempenho dos modelos [Chudik et al. 2023].

O aprendizado de máquina (ML) tem se consolidado como uma ferramenta importante na saúde, especialmente no diagnóstico do câncer de mama. O avanço tecnológico possibilitou o desenvolvimento de algoritmos para detecção precoce e precisa da doença, superando erros humanos e interpretações subjetivas [Esteva et al. 2017].

Pesquisas recentes têm explorado técnicas de ML, como SVM, Redes Neurais (ANN), *Random Forest* (RF) e Regressão Logística (LR). Esses algoritmos são eficazes no processamento de grandes volumes de dados médicos, como imagens e informações genéticas, auxiliando na tomada de decisões diagnósticas [Kourou et al. 2014].

2.1. Modelos de Aprendizagem de Máquina

As SVMs são amplamente utilizadas devido à sua capacidade de lidar com dados complexos e sua eficiência em tarefas de classificação. Elas são particularmente eficazes na

[1] https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

separação de diferentes categorias de dados, especialmente em conjuntos de dados pequenos ou equilibrados [Cortes and Vapnik 1995].

As ANNs buscam imitar o funcionamento do cérebro humano e são excelentes na detecção de padrões complexos. No entanto, elas requerem uma quantidade significativa de dados para operar de forma eficaz, além de ajustes nos parâmetros, como o número de camadas e neurônios [Lecun et al. 1998].

Baseado em árvores de decisão, o modelo Random Forest é simples de entender e eficaz na prevenção de erros causados pelo sobreajuste. Sua facilidade de aplicação e interpretação o torna bastante popular na área médica [Breiman 2001].

A Regressão Logística é um modelo mais simples, mas continua a ser uma opção válida, especialmente quando a relação entre os dados é direta e fácil de compreender [Hosmer et al. 2013].

3. Trabalhos Relacionados

Nesta seção, abordam-se os trabalhos que se concentraram na análise do desempenho de modelos de aprendizado de máquina no diagnóstico do câncer de mama, utilizando o conjunto de dados mencionado nas seções anteriores. Detalhes sobre esse conjunto de dados serão apresentados na seção de metodologia.

Nos últimos anos, diversos estudos têm-se concentrado na aplicação de técnicas de aprendizado de máquina, utilizando modelos populares como Regressão Logística (LR), Máquinas de Vetores de Suporte (SVM), Random Forest (RF) e Redes Neurais Artificiais (ANN). Para avaliar a eficácia desses modelos, foram empregadas métricas como Acurácia, Sensibilidade, Especificidade e AUC.

Nesse contexto, o estudo de [Sharma et al. 2018] comparou diferentes métodos de aprendizado de máquina, com foco em Máquinas de Vetores de Suporte (SVM) e Redes Neurais Artificiais (ANN) usando o mesmo dataset aqui trabalhado. O trabalho obteve uma acurácia de 96,66% utilizando o modelo SVM. A pesquisa avaliou o desempenho de vários algoritmos na detecção de câncer, selecionando esses métodos por sua capacidade de lidar com problemas complexos e fornecer resultados de alta precisão no diagnóstico. Para o modelo ANN, o artigo reportou uma acurácia de 93,06%.

Outro estudo [Hossin et al. 2023] utilizou o mesmo *dataset* selecionado neste estudo, teve como objetivo comparar oito algoritmos de Aprendizagem de Máquina, incluindo *Logistic Regression* (LR), *Random Forest* (RF), *K-Nearest Neighbors* (KNN), *Decision Tree* (DT), *Ada Boost* (AB), *Support Vector Machine* (SVM), *Gradient Boosting* (GB), e *Gaussian Naive Bayes* (GNB) para detecção do câncer de mama. Obtendo como melhor resultado a acurácia de 99,12% com o modelo de Regressão Logística (LR).

No estudo de [Yadav et al. 2019], foi aplicada a técnica de Máquinas de Vetores de Suporte (SVM) para prever a predisposição de uma pessoa ao câncer de mama. Os pesquisadores destacam que os custos associados a estudos desse tipo são elevados, justificando assim a escolha do conjunto de dados utilizado. Eles também avaliaram o desempenho do algoritmo, calculando a acurácia e aplicando o método de escala min-max para evitar problemas como outliers e overfitting. Além disso, utilizaram a Análise de Componentes Principais (PCA) para otimizar a seleção de parâmetros, com o objetivo de reduzir o número de variáveis e aumentar a precisão do modelo. Ao final, foi reportada

uma acurácia de 95,1% utilizando o modelo SVM com a seleção de características por meio de PCA.

Por fim, o estudo [Sidey-Gibbons and Sidey-Gibbons 2019] analisa a necessidade da capacidade de desenvolvimento da área de Inteligência Artificial aplicada à Saúde, provendo conceitos introdutórios de Aprendizagem de Máquina e um guia prático de utilizar e avaliar algoritmos utilizando ferramentas gratuitas e de domínio público. No final, o estudo reportou uma acurácia de 96% com uma sensibilidade de 97% utilizando o dataset *UCI-Winscosin Breast Cancer* (Diag).

Embora esses estudos forneçam *insights* valiosos sobre diversos aspectos do desempenho dos modelos aplicados ao problema do diagnóstico de câncer de mama usando o *dataset UCI-Winscosin*, viu-se uma oportunidade de atualizar o estudo no ano de 2024. Assim, o presente artigo visa preencher essa lacuna, concentrando-se nos fatores que podem influenciar a performance de modelos de Aprendizagem de Máquina sob o problema e *dataset* supracitados.

4. Metodologia

A metodologia adotada neste estudo consiste em três etapas principais: a seleção da base de dados da *UCI Machine Learning*, o pré-processamento dos dados e a replicação dos modelos de predição descritos na literatura, utilizando técnicas de normalização e otimização de hiperparâmetros.

4.1. Seleção da Base de Dados

Para a análise, foi selecionada a base de dados "*Breast Cancer Wisconsin*", que se concentra na avaliação das características de tumores mamários, classificando-os como malignos ou benignos. Este conjunto de dados é composto por 32 variáveis, onde cada instância é identificada por um número único (ID) e um diagnóstico (*Diagnosis*), que pode ser categorizado como M (Maligno) ou B (Benigno). As variáveis incluem características extraídas de núcleos celulares, medidas em três momentos distintos, resultando em um total de 32 atributos contínuos. Os dados foram coletados em três diferentes momentos, permitindo uma análise abrangente das características tumorais. A seguir, tem-se uma amostra e a descrição de alguns destes atributos.

Por fim, *diagnosis* teve seu valor alterado entre um (antes M) e zero (antes B) para realizar a previsão.

4.2. Pré-processamento dos Dados

O pré-processamento dos dados incluiu a normalização utilizando o método *StandScaler*, que visa padronizar as variáveis para que tenham média zero e desvio padrão igual a um. Além disso, foi aplicada a técnica de *GridSearch* para otimização dos hiperparâmetros dos modelos de predição, com o objetivo de aprimorar as métricas de desempenho. Essa abordagem sistemática garante que os modelos sejam ajustados de forma a maximizar a acurácia e a robustez das previsões.

4.3. Ambiente Exploratório

O ambiente para realização da análise exploratória e experimentos, com foco na utilização dos modelos para este estudo, inclui o uso do *Random Forest* (RF), Regressão Logística

Tabela 1. Descrição das Variáveis do Conjunto de Dados

Variável	Descrição
Raio (*radius*)	Média das distâncias do centro aos pontos no perímetro.
Textura (texture)	Desvio padrão dos valores de escala de cinza.
Perímetro (*perimeter*)	Comprimento do perímetro do núcleo.
Área (*area*)	Área total do núcleo celular.
Suavidade (*smoothness*)	Variação local nos comprimentos do raio.
Compactação (*compactness*)	Razão entre o quadrado do perímetro e a área menos 1.0.
Concavidade (*concavity*)	Severidade das partes côncavas no contorno.
Pontos côncavos (*concave points*)	Número de porções côncavas no contorno.
Simetria (*symmetry*)	Grau de simetria do núcleo celular.
Dimensão fractal (*fractal dimension*)	Aproximação baseada na "linha costeira" do contorno menos 1.

(LR), Máquinas de Vetores de Suporte (SVM) e Redes Neurais Artificiais (ANN), conforme a literatura.

A implementação desses modelos foi realizada utilizando a biblioteca Scikit-learn, em um ambiente de desenvolvimento Google Colab. Essa escolha de ferramentas permite uma execução eficiente e escalável dos algoritmos, facilitando a comparação dos resultados obtidos com os descritos na literatura.

5. Experimentos, Resultados e Discussão

Para a execução dos experimentos, foi empregada a técnica de validação cruzada (*cross_validation*), na qual o conjunto de dados foi dividido em 5 partes (*cv=5*). Em cada iteração, 80% dos dados foram utilizados para treinamento, enquanto os 20% restantes foram reservados para teste e validação.

O conjunto de dados contém 30 características (features), além de um identificador único (ID) e um alvo (Target), que classifica as lesões como Malignas ou Benignas, indicando o risco associado à análise dos atributos. No total, o conjunto de dados possui 569 instâncias, resultando em aproximadamente 114 instâncias alocadas para testes e validação. Abaixo tem-se a visualização dos resultados das acurácias dos modelos.

Os resultados dos testes indicam que o modelo de Regressão Logística (LR) apresentou desempenho superior em comparação aos demais modelos levantados pelos trabalhos relacionados, alcançando uma acurácia 0,87% maior do que a do segundo modelo mais eficaz, a Máquina de Vetores de Suporte (SVM) (Ver Fig. 1). Essa evidência sugere que a utilização de Regressão Logística, combinada com técnicas de otimização como *GridSearch* e normalização com *StandScaler*, pode resultar em desempenhos satisfatórios. Abaixo têm-se visualizações das sensibilidades e especificidades dos modelos.

Por outro lado, as Redes Neurais Artificiais (ANN) apresentaram uma acurácia 0,88% inferior à da SVM, enquanto o Random Forest (RF) também teve uma performance 0,88% abaixo da ANN. É importante ressaltar que os resultados obtidos são influencia-

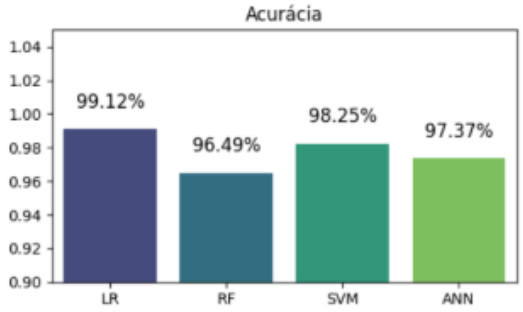

Figura 1. Comparação da acurácia dos modelos

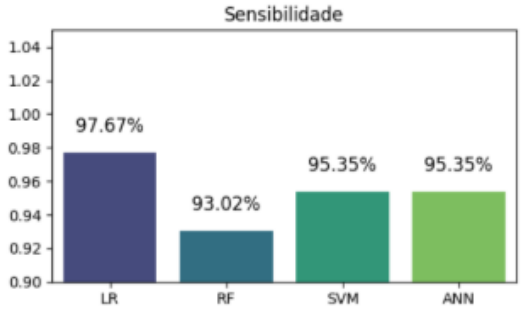

Figura 2. Comparação da sensibilidade dos modelos

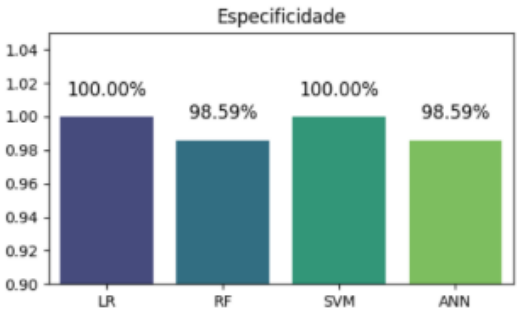

Figura 3. Comparação da especificidade dos modelos

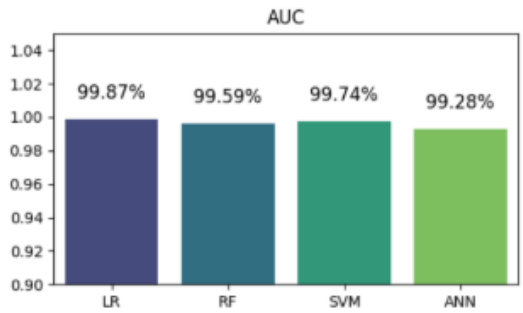

Figura 4. Comparação da AUC dos modelos

dos pelas características dos dados utilizados, podendo variar conforme a qualidade, a diversidade e a natureza da base de dados, bem como os parâmetros escolhidos para cada modelo. Essa variabilidade destaca a importância de uma análise cuidadosa dos dados e da seleção adequada dos modelos para a obtenção de resultados robustos e confiáveis.

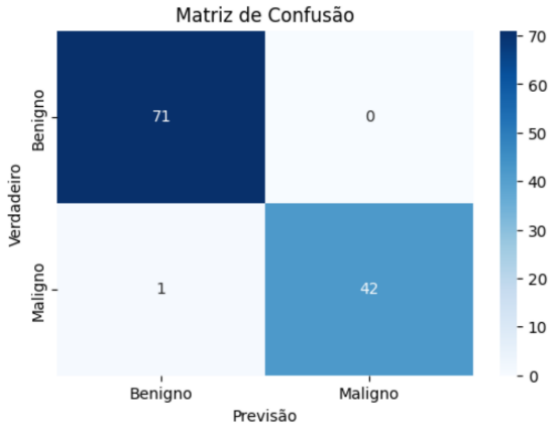

Figura 5. Matriz de confusão do melhor modelo analisado.

Esses resultados de acurácia fornecem uma visão geral do desempenho relativo de cada algoritmo em relação à tarefa de previsão de presença de tumores mamários malignos ou benignos. No entanto, é importante lembrar que a acurácia não é a única métrica relevante. Outros aspectos, como sensibilidade, especificidade, AUC e a matriz de confusão, também são essenciais para uma avaliação mais completa do desempenho dos modelos.

Para tanto, apresentaram-se também tais visualizações para corroborar com a análise e discussão efetuada. No caso da Matriz de Confusão, observa-se uma forte acurácia, com apenas 1 instância errada sugerida pelo melhor modelo.

6. *Minimum Viable Product (MVP)*

Nossa solução consiste em um algoritmo de Aprendizagem de Máquina que utiliza as técnicas de *GridSearch* e *StandardScaler* para otimizar os resultados, demonstrando uma melhoria significativa em várias métricas em comparação com os resultados apresentados nos artigos analisados. Este avanço é especialmente relevante no contexto da análise clínica, onde a precisão das previsões pode impactar diretamente o diagnóstico e o tratamento de pacientes.

O produto mínimo viável (MVP) está disponível no link [2]. Em consonância com o Manifesto para a Ciência Aberta da UNESCO [UNESCO 2022], o código utilizado nos experimentos e na implementação do MVP está acessível a todos por meio do seguinte link [3]. Este repositório contém o código-fonte do MVP e dos experimentos realizados, promovendo a transparência e a colaboração na pesquisa.

Com a disponibilização de um ambiente de experimentação 100% aberto, que inclui o conjunto de dados, os experimentos realizados e o código do MVP, busca-se incentivar investigações adicionais, críticas construtivas e melhorias por parte da comunidade científica. Essa abordagem não apenas fortalece a pesquisa colaborativa, mas também contribui para o avanço do conhecimento na área de IA aplicada à saúde.

7. Considerações Finais

Conclui-se que é viável aprimorar as métricas de previsão apresentadas nos artigos analisados, evidenciando que a qualidade e a quantidade dos dados utilizados têm um impacto significativo nos resultados obtidos. A análise realizada demonstra que a escolha de uma base de dados com características mais robustas pode levar a melhorias nas métricas de desempenho dos modelos.

Além disso, a exploração de diferentes abordagens e implementações, como a análise da distribuição de frequência dos dados, representa uma oportunidade promissora para investigar variações nas métricas de previsão. Essa abordagem pode revelar *insights* valiosos sobre as singularidades dos dados e suas influências nos resultados.

Entretanto, é importante reconhecer algumas limitações deste estudo. A dependência de um único conjunto de dados pode restringir a generalização dos resultados. Assim, recomenda-se a realização de experimentos com múltiplas bases de dados para validar as conclusões aqui apresentadas.

Sugere-se, para trabalhos futuros, investigar técnicas avançadas de pré-processamento, modelagem e algoritmos de Aprendizagem de Máquina mais complexos. A integração de métodos de ensemble e redes neurais profundas pode melhorar a acurácia das previsões, avançando o conhecimento e aprimorando as práticas de diagnóstico clínico.

[2] https://wdbc-ads.streamlit.app/

[3] https://colab.research.google.com/drive/1hrFVa0NwL04aJr21iTQFKrEY0Eqk389n?usp=sharing

Referências

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Chudik, A., Pesaran, M. H., and Sharifvaghefi, M. (2023). Variable selection in high dimensional linear regressions with parameter instability. Technical report, CESifo Working Paper, No. 10223, Center for Economic Studies and ifo Institute (CESifo).

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118.

GEBRIM, L. H. (2016). A detecção precoce do câncer de mama no brasil. *Cadernos de Saúde Pública*, 32.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R., and Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, 25.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression.* John Wiley & Sons, Inc., 3rd edition.

Hossin, M., Shamrat, F., Bhuiyan, M. R., Hira, R., Khan, T., and Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the wisconsin dataset. *Bulletin of Electrical Engineering and Informatics*, 12:2446–2456.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8 – 17.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324.

Pinto, K. R. T. d. F., Mattias, S. R., Lima, N. d. M., Santos, I. D. d. L., Bernardy, C. C. F., and Sodré, T. M. (2018). Câncer de mama: sentimentos e percepções das mulheres diante do diagnóstico / breast cancer: feelings and perceptions of women before the diagnosis. *Revista de Pesquisa Cuidado é Fundamental Online*, 10(2):385–390.

Sharma, A., Kulshrestha, S., and Daniel, S. (2018). Machine learning approaches for cancer detection. *International Journal of Engineering and Manufacturing*, 8:45–55.

Sidey-Gibbons, J. A. M. and Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19.

UNESCO (2022). Recomendação da unesco sobre ciência aberta. Technical report, UNESCO Office Brasilia.

Yadav, A., Jamir, I., Jain, R., and Sohani, M. (2019). Breast cancer prediction using svm with pca feature selection method. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pages 969–978.