Detecção de Câncer de Mama Baseada em Aprendizado de Máquina

Leandro Ivanildo da Silva[1]

[1]Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE - Campus jaboatão dos Guararapes)

Caixa Postal 54080-000 – Jaboatão dos Guararapes – PE – Brazil

lis4@discente.ifpe.edu.br

Abstract. *Early detection of breast cancer is a powerful tool to reduce its socioeconomic impact. Tumors can be classified as benign or malignant, representing critical conditions that affect global public health. The application of machine learning techniques, especially classification algorithms such as CatBoost, enables advances in the early and accurate detection of these tumors from clinical data. This study focuses on the use of CatBoost applied to the Breast Cancer Wisconsin dataset. Although artificial intelligence (AI) methods have shown remarkable results for this goal, their "black box" nature hinders their widespread adoption in clinical practice. To overcome this limitation, interpretability methods can be employed, allowing not only the evaluation of model performance but also the understanding of data features that influence the algorithm's decisions, making the results more transparent and reliable for clinical application.*

Keywords: *Machine Learning, CatBoost, Interpretability, Artificial Intelligence, breast cancer*

Resumo. *A detecção precoce do câncer de mama é uma ferramenta poderosa para reduzir seu impacto socioeconômico. Os tumores podem ser classificados como benignos ou malignos, representando condições críticas que afetam a saúde pública mundial. A aplicação de técnicas de aprendizado de máquina, especialmente algoritmos de classificação como o CatBoost, possibilita avanços na detecção precoce e precisa desses tumores a partir de dados clínicos. Este estudo foca no uso do CatBoost aplicado ao dataset Breast Cancer Wisconsin. Embora métodos de inteligência artificial (IA) tenham apresentado resultados notáveis para esse objetivo, sua natureza de "caixa preta" dificulta sua ampla adoção na prática clínica. Para superar essa limitação, métodos de interpretabilidade podem ser empregados, permitindo não apenas a avaliação do desempenho do modelo, mas também a compreensão das características dos dados que influenciam as decisões do algoritmo, tornando os resultados mais transparentes e confiáveis para aplicação clínica.*

Palavras-chave: *Aprendizado de máquina, CatBoost, Interpretabilidade, Inteligência Artificial, Câncer de mama*

1. Introdução

O câncer de mama tornou-se a neoplasia mais comum no mundo, ultrapassando o câncer de pulmão em número de diagnósticos, conforme dados recentes da Organização Mundial da Saúde (OMS) e da Agência Internacional de Pesquisa em Câncer (IARC). Essa doença representa um desafio significativo para a saúde pública global, especialmente devido ao seu impacto socioeconômico e à crescente incidência, que tende a aumentar nas próximas décadas [1][2].

Fatores como o aumento da expectativa de vida, obesidade e sedentarismo — intensificados pela pandemia de COVID-19 — são apontados como os principais contribuintes para esse cenário. O diagnóstico precoce é fundamental para melhorar o prognóstico e aumentar as chances de cura, que podem chegar a 90% quando a doença é detectada em estágios iniciais. Entretanto, a pandemia agravou o acesso a exames e tratamentos, resultando em atrasos que podem levar a quadros mais avançados e maiores índices de mortalidade [1].

Nesse contexto, a aplicação de técnicas de aprendizado de máquina para a detecção precoce do câncer de mama tem ganhado destaque, possibilitando a análise eficiente de dados clínicos para auxiliar no diagnóstico [2]. No entanto, a adoção clínica desses métodos é limitada pela dificuldade de interpretação dos modelos, geralmente considerados “caixas pretas”.

Portanto, este trabalho tem como objetivo aplicar o algoritmo CatBoost ao conjunto de dados Breast Cancer Wisconsin, avaliando seu desempenho na classificação de tumores e explorando métodos de interpretabilidade que tornem as decisões do modelo mais transparentes e confiáveis para uso clínico [3].

2. Trabalhos Relacionados

Diversos estudos têm explorado o uso de técnicas de aprendizado de máquina para a detecção precoce do câncer de mama, demonstrando o potencial desses métodos em apoiar o diagnóstico clínico. O avanço das ferramentas de *machine learning* permitiu desenvolver modelos capazes de identificar padrões sutis nos dados médicos, auxiliando na distinção entre tumores benignos e malignos a partir de características extraídas de exames clínicos e laboratoriais.

Em [4], **SHARMA et al.** utilizaram o algoritmo Support Vector Machine (SVM) aplicado ao *Breast Cancer Wisconsin Dataset*, alcançando altas taxas de acurácia (96.66%) na classificação dos tumores. Os autores destacam que o desempenho do modelo depende fortemente da escolha dos parâmetros do kernel e do pré-processamento dos dados, o que pode limitar sua aplicação em contextos clínicos reais. No entanto, a interpretabilidade dos modelos ainda foi apontada como um desafio, uma vez que as decisões internas desses algoritmos são difíceis de compreender por profissionais da saúde [3].

Pesquisas mais recentes têm focado em algoritmos de boosting, como o XGBoost e o LightGBM, que demonstram resultados superiores em termos de precisão e tempo de treinamento [5]. Contudo, esses modelos exigem otimização manual e codificação de variáveis categóricas, o que pode tornar o processo mais complexo e suscetível a erros.

Nesse contexto, o CatBoost surge como uma alternativa promissora. Ele combina alto desempenho com tratamento nativo de variáveis categóricas, o que mitiga a necessidade de pré-processamento manual complexo. Além disso, o CatBoost incorpora mecanismos que reduzem o *overfitting* e simplificam o ajuste de hiperparâmetros [6].

Além disso, trabalhos como [3] enfatizam a importância da interpretabilidade dos modelos, especialmente em aplicações médicas. Métodos como SHAP permitem compreender a influência de cada característica clínica nas decisões do modelo, aumentando a transparência e confiabilidade das previsões. Essa abordagem torna o uso da inteligência artificial mais alinhado à prática médica, fornecendo suporte à decisão sem substituir o julgamento clínico.

Dessa forma, este trabalho propõe a aplicação do algoritmo CatBoost ao conjunto de dados *Breast Cancer Wisconsin*, com o objetivo de avaliar seu desempenho e explorar sua interpretabilidade, buscando contribuir para o desenvolvimento de sistemas de apoio ao diagnóstico mais precisos e compreensíveis.

3. Metodologia

Este estudo foi desenvolvido em quatro etapas principais: coleta de dados, pré-processamento, treinamento do modelo e interpretação dos resultados. O objetivo foi avaliar a capacidade do algoritmo CatBoost em classificar tumores de mama como benignos ou malignos, além de aplicar métodos de interpretabilidade para compreender as decisões do modelo.

3.1 Conjunto de Dados

Foi utilizado o Breast Cancer Wisconsin (Diagnostic) Dataset, amplamente empregado em pesquisas de aprendizado de máquina voltadas à detecção de câncer de mama. O conjunto contém 569 amostras e 30 atributos numéricos extraídos de imagens digitalizadas de aspirados por agulha fina (FNA) de massas mamárias, como raio, textura, suavidade, compacidade e simetria. Cada instância é rotulada como benigna (B) ou maligna (M).

3.2 Pré-processamento dos Dados

Os dados foram carregados e analisados em Python, utilizando as bibliotecas pandas, numpy e scikit-learn. As etapas de pré-processamento incluíram:

- Padronização dos atributos numéricos**:** Foi utilizado o **StandardScaler** do scikit-learn para padronizar resultando em média zero e desvio-padrão unitário. Esta etapa é essencial para o bom desempenho de algoritmos baseados em distância, como o Support Vector Machine (SVM), que foi incluído na análise de comparação.

- Divisão dos dados em 80% para treino e 20% para teste, utilizando o método train_test_split com semente aleatória fixa (random_state=42 e stratify=y) para garantir a reprodutibilidade.

3.3 Treinamento e Seleção do Modelo

Neste estudo, foram treinados e comparados dois modelos: o CatBoostClassifier e o Support Vector Machine (SVM).

O CatBoostClassifier foi construído com a seguinte estratégia de otimização: um alto número de iterações (iterations=3000) combinado com o uso de *early stopping* (early_stopping_rounds=500). Esta técnica interrompe o treinamento quando a métrica de avaliação não melhora no conjunto de teste, controlando o *overfitting* de forma eficiente [7]. A configuração final de hiperparâmetros utilizada foi:

Hiperparâmetro	**Valor**	**Descrição**
iterations	3000	Número máximo de árvores (estimadores) a serem construídas. Um valor alto é usado com o *early stopping*.
learning_rate	0.03	Controla o tamanho do passo para a otimização. Valores menores geralmente levam a um modelo mais robusto.
depth	8	Profundidade máxima de cada árvore de decisão. Controla a complexidade do modelo.

l2_leaf_reg	5	Coeficiente de regularização L2 aplicado às folhas. Ajuda a evitar o *overfitting*.
eval_metric	'F1'	Métrica usada para avaliação de *early stopping*. A F1-score é relevante em problemas de classificação com classes desbalanceadas.
early_stopping_rounds	500	Número de iterações sem melhoria na eval_metric no conjunto de teste antes de interromper o treinamento.

A avaliação de ambos os modelos foi feita por meio de métricas críticas para o contexto clínico: **Acurácia** (proporção de classificações corretas), **Sensitivity (Revocação)** (taxa de verdadeiros positivos, crucial para identificar corretamente casos malignos) e **Especificidade (Specificity)** (taxa de verdadeiros negativos), além da **Área sob a Curva ROC (AUC)**, obtidas no conjunto de teste. O uso dessas métricas visa uma avaliação abrangente, uma vez que, em aplicações médicas, a alta Sensibilidade é vital para minimizar falsos negativos [4].

3.4 Interpretabilidade e Análise de Resultados

Para interpretar as decisões do modelo CatBoost, utilizou-se a biblioteca SHAP. A análise de importância das variáveis e a contribuição de cada atributo para a previsão final foram realizadas por meio das seguintes visualizações:

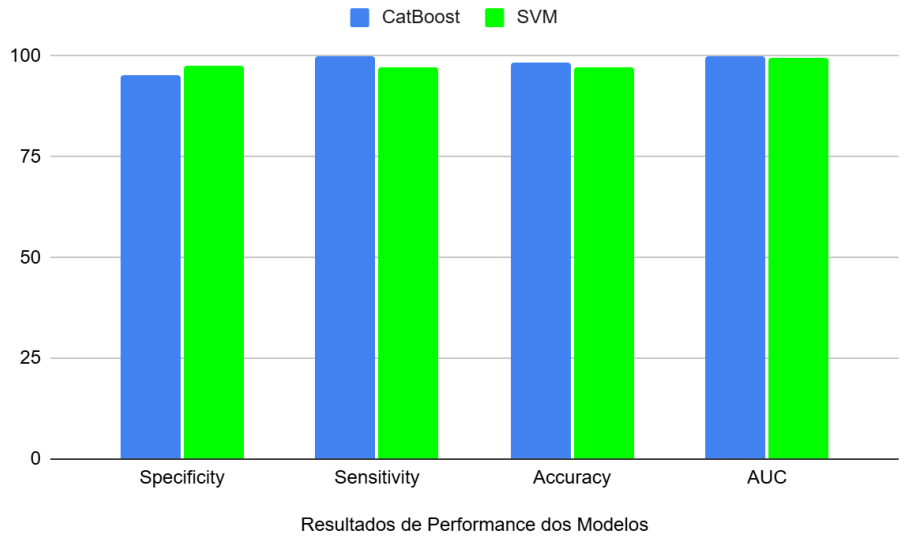
- **Summary Plot** (para importância global das características);
- **Decision Plot** (para análise de predições individuais);
- **Waterfall Plot** (para decomposição do impacto das variáveis em casos específicos).

3.5 Ferramentas e Ambiente de Execução

O ambiente de desenvolvimento utilizado foi o Visual Studio Code, com execução local em Python 3.10. As principais bibliotecas empregadas foram: catboost, shap, pandas, scikit-learn, numpy e matplotlib. Para fins de apresentação interativa, foi utilizado o *framework* Streamlit.

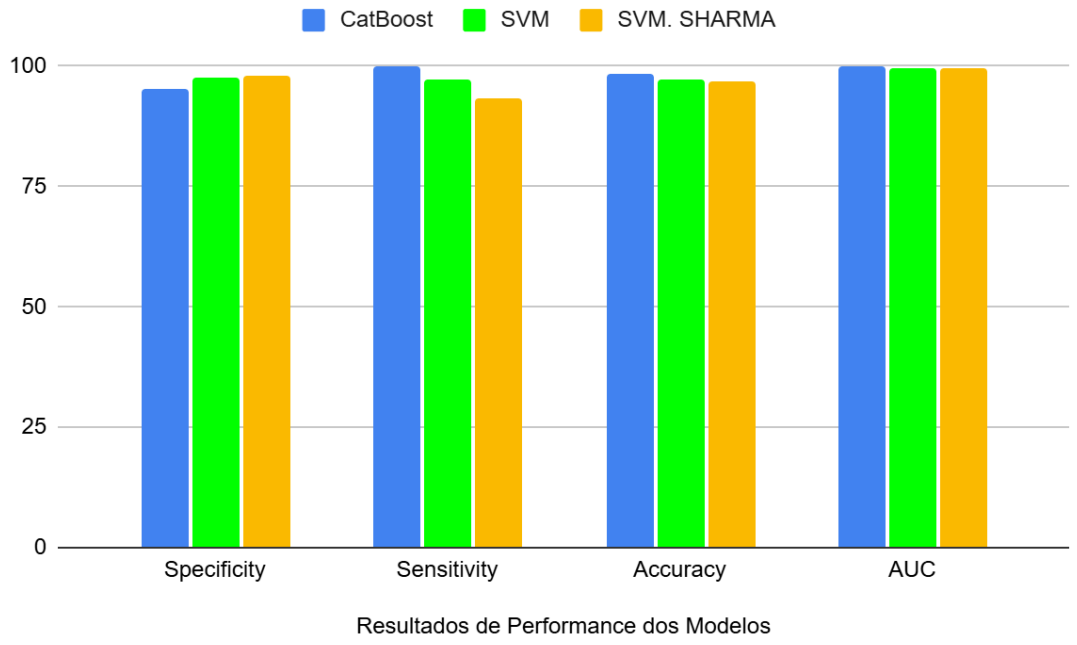
4. Resultados e Discussão

Após o treinamento e comparação dos modelos, o desempenho foi avaliado no conjunto de testes (20% dos dados). Os resultados de *performance* obtidos estão resumidos na Tabela 1:



Os resultados mostram que o CatBoost atingiu a maior Acurácia (98,25%) e o maior valor de AUC (0.9983), confirmando a superioridade dos algoritmos de *boosting* em relação ao SVM para este conjunto de dados. Em particular, a Sensitivity (Revocação) de 100 indica que o modelo CatBoost conseguiu identificar corretamente todos os casos malignos (verdadeiros positivos) no conjunto de teste, um resultado clinicamente crítico.

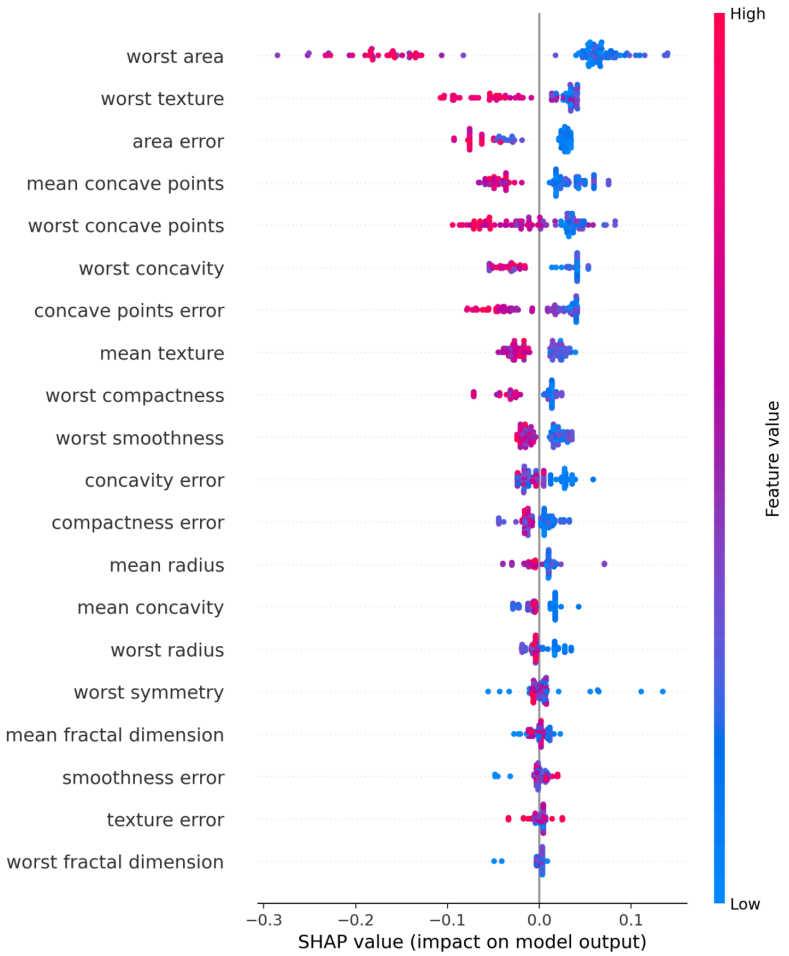
Quando comparado ao trabalho de SHARMA et al. [4], que obteve uma acurácia de **96,66%** com o algoritmo SVM no mesmo *dataset*, este estudo superou tal resultado tanto com o modelo SVM (96,49%, resultado similar) quanto, de forma notável, com o CatBoost (98,25%). A superioridade do CatBoost demonstra o avanço em relação a métodos de *Machine Learning* mais tradicionais, solidificando a premissa de que algoritmos de *ensemble* podem extrair mais conhecimento de conjuntos de dados clínicos.



Análise de Importância das Variáveis

Para interpretar as decisões do modelo, foi utilizada a biblioteca SHAP, permitindo identificar quais atributos tiveram maior influência nas previsões. As principais características relevantes, conforme o *Summary Plot* do SHAP, foram:

- **Worst area** (Pior área);
- **Worst texture** (Pior textura);
- **Area error** (Erro padrão da área);
- **Mean concave points** (Média de pontos côncavos);
- **Worst concave points** (Pior pontos côncavos).



A análise do gráfico revela que:

- **Pior área (Worst area):** É a característica mais importante. Valores altos (vermelho) tendem a mover a previsão para o lado positivo do eixo SHAP, indicando uma maior probabilidade de classificação como maligno.
- **Pior textura (Worst texture):** Valores altos (vermelho) também aumentam a chance de ser maligno, enquanto valores baixos (azul) diminuem.
- **Erro de área (Area error):** Valores altos (vermelho) indicam maior chance de maligno.

Esses resultados estão em total concordância com o conhecimento clínico e com estudos similares que utilizam métodos de interpretabilidade [3], uma vez que características morfológicas como área e pontos côncavos (associadas à irregularidade e crescimento anômalo) são cruciais na distinção entre tumores benignos e malignos [8]. A prevalência de features na categoria "Worst" (pior) reflete que as maiores irregularidades detectadas nas células são os fatores determinantes na predição de malignidade.

5. Conclusão

A detecção precoce do câncer de mama é um desafio global de saúde, e a aplicação de técnicas de *Machine Learning* emerge como uma ferramenta poderosa para auxiliar no diagnóstico. Este trabalho propôs e avaliou a aplicação do algoritmo CatBoost ao conjunto de dados Breast Cancer Wisconsin, com o objetivo de alcançar alta performance e, crucialmente, garantir a interpretabilidade dos resultados para aplicação clínica.

O CatBoost demonstrou um desempenho notável, alcançando uma Acurácia de 98,25% e uma Sensitivity (Revocação) de 100% para casos malignos no conjunto de teste, superando os resultados de trabalhos anteriores com algoritmos como o SVM. A alta Sensitivity é de extrema relevância em um contexto médico, minimizando a taxa de falsos negativos e garantindo que casos críticos sejam identificados.

Além da performance, a aplicação dos métodos de interpretabilidade SHAP foi fundamental. A análise de importância das variáveis validou o modelo, destacando características clinicamente relevantes como Pior Área, Pior Textura e Média de Pontos Côncavos como as maiores influenciadoras na classificação de tumores malignos. Essa descoberta está em consonância com o conhecimento clínico e com estudos similares que utilizam métodos de interpretabilidade no mesmo *dataset* [3], confirmando que o modelo está tomando decisões com base em características morfológicas biologicamente significativas.

Esse nível de transparência é essencial para que profissionais de saúde confiem e adotem sistemas de Inteligência Artificial em sua prática. O modelo CatBoost, quando acompanhado pela sua explicação SHAP, deixa de ser uma "caixa preta" e se torna um sistema de apoio à decisão mais confiável.

Referências

[1] (2020, Feb. 3). WHO. Câncer: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action

[2] (2023, Feb. 9). Femama. Cancer: https://femama.org.br/site/noticias-recentes/por-que-o-cancer-de-mama-se-tornou-a-forma-mais-comum-da-doenca-no-mundo

[3] (2022, Fev). KARATZA, Panagiota; DALAKLEIDI, Kalliopi V.; ATHANASIOU, Maria; NIKITA, Konstantina. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis.

[4] (2018, Mar. 8). SHARMA, Ayush; KULSHRESTHA, Sudhanshu; DANIEL, Sibi B. Machine Learning Approaches for Cancer Detection. *Jaypee Institute of Information Technology*, Noida, Uttar Pradesh, Índia.

[5] MARINHO, Tiago Lima. *Otimização de hiperparâmetros do XGBoost utilizando meta-aprendizagem.* Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas, Maceió, 2021.

[6] DOROGUSH, Aleksei V. et al. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* (NIPS), 2018. p. 6638-6648.

[7] Undefitting and Overfitting. https://didatica.tech/underfitting-e-overfitting/. Accessed: 2025-10-14.

[8] A. Baba, C. Catoi, “Tumor cell morphology,” in Comparative Oncology, The Publishing House of the Romanian Academy, Eds.