Novo Ensino Suplementar

Predição de tumores malignos utilizando inteligência artificial

Trabalho de Conclusão de Curso

por

João Pedro Pacheco Santos

Orientador: Prof. Eduardo Adame

Maceió - AL, Outubro / 2025

João Pedro Pacheco Santos

Predição de tumores malignos utilizando inteligência artificial

Monografia apresentada como requisito para obtenção da certificação Completa em Inteligência Artificial e Ciência dos Dados do Novo Ensino Suplementar.

Orientador: Prof. Eduardo Adame

Maceió

2025

Agradecimentos

Agradeço a toda a equipe do programa Novo Ensino Suplementar por implementar, acompanhar e impulsionar o ensino de excelência para mim e para os estudantes de Alagoas de forma equitativa. Agradeço, também, ao meu orientador por me ajudar nesse processo de aprendizado.

Qualquer tecnologia suficientemente avançada é indistinguível de magia.

Arthur C. Clarke

RESUMO

Este trabalho tem como objetivo analisar o desempenho de diferentes modelos de aprendizado de máquina na classificação de tumores malignos e benignos de mama. Para isso, utilizou-se a base de dados Breast Cancer Wisconsin, que contém características morfológicas de células diversas, extraídas de massas mamárias. Foram implementados e comparados seis modelos de classificação -Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes e Gradient Boosting- avaliados a partir das métricas de acurácia, precisão, recall e F1-Score. Os resultados obtidos indicaram que o modelo Random Forest apresentou o melhor desempenho geral entre os demais, destacando-se pela sua capacidade de classificação. Portanto, o estudo demonstra o potencial do uso da inteligência artificial como ferramenta complementar ao diagnóstico clínico, contribuindo para a detecção precoce do câncer de mama.

Palavras-chave: Inteligência Artificial, Aprendizado de Máquina, Dados de Câncer de Mama, Classificação, Random Forest.

ABSTRACT

This study aims to analyze the performance of different machine learning models in classifying malignant and benign breast tumors. To this end, we used the Wisconsin Breast Cancer database, which contains morphological characteristics of various cells extracted from breast masses. Six classification models were implemented and compared—Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, and Gradient Boosting—and evaluated based on accuracy, precision, recall, and F1-Score metrics. The results indicated that the Random Forest model performed best overall, standing out for its classification ability. Therefore, the study demonstrates the potential of using artificial intelligence as a complementary tool to clinical diagnosis, contributing to the early detection of breast cancer.

Keywords: Artificial Intelligence, Machine Learning, Breast Cancer Datasets, Classification, Random Forest.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

TP	Verdadeiro Positivo
FP	Falso Positivo
TN	Verdadeiro Negativo
FN	Falso Negativo

SUMÁRIO

1 INTRODUÇÃO

O câncer de mama é um dos tipos de câncer com maior incidência e letalidade no mundo [1]. No Brasil, o Instituto Nacional de Câncer (INCA) estima que, para cada ano dentre 2023/2025, sejam diagnosticados 73.610 novos casos de câncer de mama [2]. No entanto, o câncer de mama apresenta uma alta probabilidade de cura quando descoberto em estágios iniciais. Os métodos tradicionais de diagnóstico clínico, como a mamografia e a análise histopatológica, são eficazes, mas podem gerar falsos positivos e dependem da interpretação humana [3].

Nesse contexto, as técnicas de Inteligência Artificial (IA) e aprendizado de máquina vêm sendo estudadas e aplicadas para melhorar a detecção precoce de tumores malignos, oferecendo análise automatizada de dados clínicos e mais barata de ser feita. O uso da IA no contexto da saúde tem demonstrado enormes impactos para acelerar o diagnóstico e auxiliar os profissionais na decisão correta, melhorando a elaboração de soluções mais precisas [4]. Este trabalho foca no problema da pesquisa: Como a aplicação da inteligência artificial na classificação de dados sobre tumores malignos se comporta em diferentes modelos de aprendizado e como pode contribuir para a melhoria dos métodos tradicionais de diagnóstico?

Para responder a essa questão, esse projeto propõe a interpretação de modelos de aprendizado para a construção de modelos de IA capazes de identificar tumores malignos e benignos de mama. Espera-se demonstrar que certas abordagens de algoritmos podem oferecer maior velocidade e acurácia, auxiliando o diagnóstico clínico.

1.1 Metodologia

Existem vários métodos disponíveis para a extração de uma base de dados. Uns são mais invasivos e outros não. O método utilizado se baseia na extração de células do tumor encontrado, sendo ele benigno ou maligno, e a partir de uma série de métricas, são extraídos dados acerca das suas características. Esses dados são computados e organizados em uma tabela com seus seguintes diagnósticos. Essa base de dados serve para treinar modelos de aprendizado para identificar distinções entre os tumores.

2 ANÁLISE EXPLORATÓRIA DOS DADOS

A base de dados utilizada, conhecida como Breast Cancer Wisconsin (Diagnostic) Dataset [5], apresenta 32 colunas com características de cada célula estudada, extraídas de imagens digitalizadas de massas mamárias. Dentre estas, a coluna "id" deve ser descartada por ser um identificador único para cada paciente e não ter valor preditivo. As 31 colunas restantes representam dez características do núcleo celular, calculadas em termos de média, erro padrão e o "pior" valor (média dos três maiores valores), além da coluna de diagnóstico, contendo dois tipos de dados: M (tumor maligno) e B (tumor benigno).

Tabela 1: DataFrame da base de dados

id	diagnosis	radius_mean	texture_mean	perimeter_mean	...	area_mean	fractal_dimension_worst
842302	M	17.99	10.38	122.80	...	1001.0	0.11890
842517	M	20.57	17.77	132.90	...	1326.0	0.08902
84300903	M	19.69	21.25	130.00	...	1203.0	0.08758
84348301	M	11.42	20.38	77.58	...	386.1	0.17300
84358402	M	20.29	14.34	135.10	...	1297.0	0.07678
				...			
926424	M	21.56	22.39	142.00	...	1479.0	0.07115
926682	M	20.13	28.25	131.20	...	1261.0	0.06637
926954	M	16.60	28.08	108.30	...	858.1	0.07820
927241	M	20.60	29.33	140.10	...	1265.0	0.12400
92751	B	7.76	24.54	47.92	...	181.0	0.07039

569 rows × 32 columns

A seguir, a média, desvio padrão, mínimo e máximo das colunas da base de dados sobre a média normal de cada coluna.

Tabela 2: Estatísticas Descritivas das Características (Mean)

Característica	**Média**	**Desv. Padrão**	**Mínimo**	**Máximo**
radius_mean	14.13	3.52	6.98	28.11
texture_mean	19.29	4.30	9.71	39.28
perimeter_mean	91.97	24.30	43.79	188.50
area_mean	654.89	351.91	143.50	2501.00
smoothness_mean	0.10	0.01	0.05	0.16
compactness_mean	0.10	0.05	0.02	0.35
concavity_mean	0.09	0.08	0.00	0.43
concave_points_mean	0.05	0.04	0.00	0.20
symmetry_mean	0.18	0.03	0.11	0.30
fractal_dimension_mean	0.06	0.01	0.05	0.10

Depois da base de dados posicionada no algoritmo, é essencial observar se há dados nulos ou duplicados a partir das funções `isnull` e `duplicated`. Ao executar as funções, a base de dados não apresentou dados nulos ou duplicados, retirando a necessidade do tratamento dos dados.

Um método para observar quais colunas estão fortemente relacionadas é o mapa de calor de correlações. Esse método apresenta uma matriz de cores na qual a interseção das linhas e colunas em um ponto representa o quão relacionados os respectivos dados estão.

2.1 Mapa de Calor de Correlação

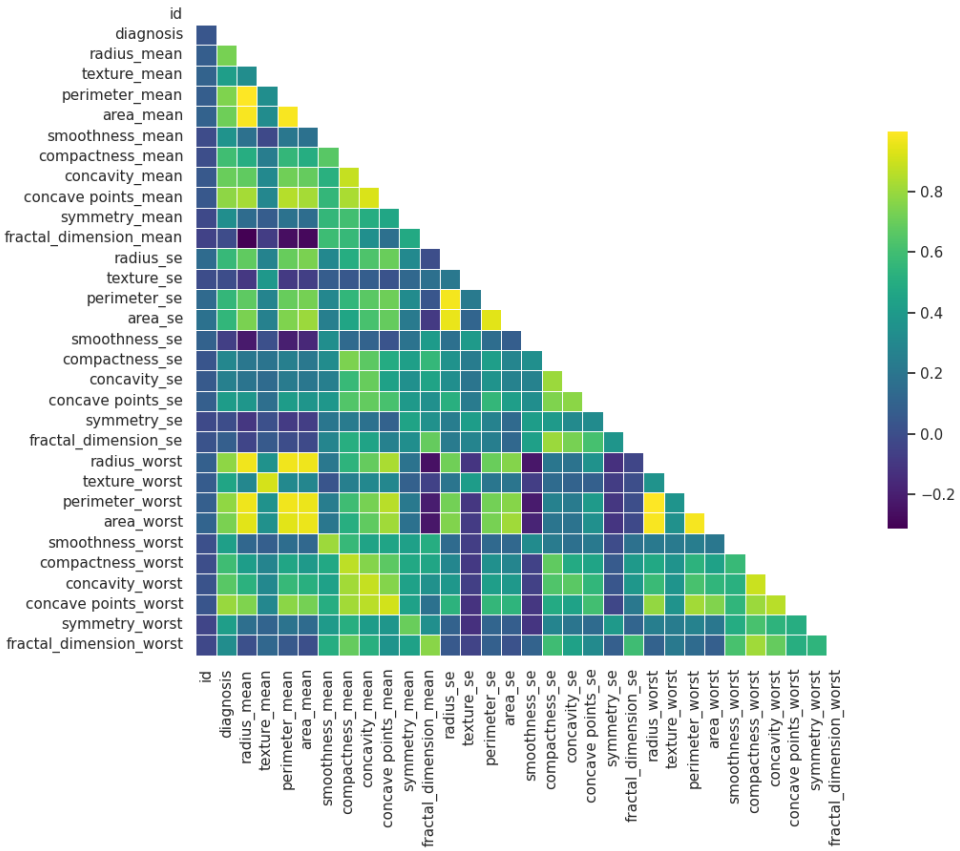

Figura 1: Matriz de Correlação

A partir dos dados observados, constata-se que os dados fortemente relacionados são aqueles que apresentam uma coloração em sua interseção mais amarela em relação aos demais. Isso pode ser observado em 21 casos. Nesses casos, temos que uma variável pode ser obtida por meio de equações com a outra variável como base. Exemplo disso é o perímetro, que depende do valor do raio da célula pela seguinte fórmula:

$$\text{Perímetro} = 2 \cdot \pi r. \tag{2.1}$$

Essas dependências entre equações, no entanto, não representam uma maioria em relação ao total de correlações. Por isso, os modelos de aprendizado, como Random Forest ou Gradient Boosting, podem ter métricas parecidas.

2.2 **Features Importantes**

Outro aspecto importante é o peso que cada coluna de dados tem em relação ao treino de dados. O método utilizado é o Mean Decrease in Impurity (MDI), com base no modelo de decisão Random Forest, no qual utiliza as variáveis para dividir os nós da árvore de decisão e, assim, as variáveis mais importantes chegam até as camadas mais profundas da árvore de decisão, sendo essas essenciais para o treino de dados.

A seguir, a tabela apresentando os 10 features mais importantes para a classificação de tumores:

Tabela 3: Importância Relativa das 10 Principais Características (Gini)

Característica (Feature)	**Importância Relativa**
area_worst	0.1514
concave_points_worst	0.1265
radius_worst	0.0935
perimeter_worst	0.0836
concave_points_mean	0.0811
perimeter_mean	0.0771
radius_mean	0.0620
concavity_mean	0.0508
area_mean	0.0459
concavity_worst	0.0300

A partir dessa métrica, vê-se que a área e a concavidade são os fatores mais decisivos para a classificação do tumor entre maligno e benigno, o que é, também, decisivo para a análise clínica, pois a célula maligna se espalha irregularmente na mama, criando uma textura falha em relação ao tumor benigno [1]. Portanto, a análise computacional apresenta concordância em relação à análise clínica.

Para grandes volumes de dados, a extração dessas features se torna importante para economizar o custo computacional sem perder a eficiência dos modelos de aprendizado de máquina.

3 AJUSTE DOS MODELOS

Para que um modelo possa funcionar adequadamente aos dados inseridos, primeiro é necessário ajustá-lo conforme o modo como a base de dados está organizada.

Como a base de dados utilizada se trata de informações numéricas, à exceção da coluna de classificação, e as métricas mais decisivas estão relacionadas a medidas proporcionais, quanto mais irregular for a célula, maior é a probabilidade dela ser maligna. Os ajustes serão mais simples de serem feitos, pois dispensam a junção de dados ou manipulações complexas.

Para implementar os dados, primeiro é necessário dividir o conjunto de dados. Para isso, é necessário criar dois conjuntos de dados, chamados X e Y. Os dados X contêm as métricas de cada tipo de câncer, retirando a coluna ID por ser trivial e a coluna Diagnóstico para o treino dos dados. Já os dados Y contêm apenas a coluna de diagnóstico, para treino e verificação de predições.

Na coluna Y, como os dados não são numéricos, será necessário usar a função LabelEncoder, do SKLearn, que substitui o diagnóstico benigno por 0 e o diagnóstico maligno por 1. Isso melhora a capacidade de classificação do modelo.

Uma vez que os dados estão separados corretamente, para que eles se adequem ao limite entre 0 e 1 da coluna Y, uma normalização dos dados X é necessária. A normalização põe os dados em um limite entre 0 e 1, ao invés de números grandes, melhorando a capacidade de leitura e processamento dos dados. O método utilizado é o StandardScaler do SKLearn.

Os dados, uma vez que estão tratados, serão subdivididos em treino e teste pela função `Train_Test_Split`, com métricas que separam 80% dos dados em treino, que são alocados para a função `X_Train` e para `Y_Train`, e os 20% dos dados restantes vão para a função teste, a fim de monitorar como o modelo está se comportando.

3.1 **Cross Validation**

Uma vez que os dados estão organizados, é necessário testar qual modelo se adequará melhor aos dados. O cross-validation faz exatamente isso: treina diferentes modelos e analisa, no final, qual se saiu melhor em desempenho.

O cross-validation utiliza o StratifiedKFold para separar os dados dentro do modelo. Essa função permite separar os dados em várias amostras com a mesma proporção da base de dados, evitando o treino acidental de apenas uma classe de dados. A partir disso, os parâmetros de teste serão imputados. Os parâmetros foram limitados da seguinte forma:

- n_splits = 50: o validação cruzada ocorrerá 50 vezes.
- shuffle = True: Os dados serão embaralhados antes de serem treinados para evitar que padrões de dados na base original comprometam na classificação do modelo.
- random_state = 42: fixa o padrão de embaralhamento dos dados

A métrica de performance para a classificação dos parâmetros será feita pelo F1-score, acurácia, precisão e Recall.

A acurácia, a precisão, o recall e o F1-score são métricas essenciais para avaliar o desempenho de modelos de classificação. A acurácia mede a proporção de previsões corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de amostras, oferecendo uma visão geral do desempenho. No entanto, ela pode ser enganosa em bases de dados desbalanceadas. A precisão foca nos resultados positivos previstos, indicando quantos deles estavam de fato corretos, sendo crucial quando o custo de um falso positivo é alto. O recall (ou sensibilidade) mede a capacidade do modelo de identificar todos os resultados positivos relevantes, sendo importante quando o custo de um falso negativo é significativo. Por fim, o F1-score é a média harmônica entre precisão e recall, fornecendo um único valor que equilibra ambas as métricas, tornando-se especialmente útil quando há um desequilíbrio entre as classes ou quando tanto falsos positivos quanto falsos negativos são igualmente indesejáveis [6].

Os modelos de classificação a serem treinados são:

- **Logistic Regression**: Modelo estatístico baseado em uma função logística (sigmoide), que estima a probabilidade de uma amostra pertencer a uma classe. É eficiente em problemas linearmente separáveis e oferece boa interpretabilidade [7].
- **Random Forest**: Conjunto de múltiplas árvores de decisão geradas de forma aleatória, cujo resultado final é obtido por votação das previsões individuais. É robusto contra overfitting e captura bem relações não lineares [7].
- **Support Vector Machine (SVM)**: Modelo que busca um hiperplano ótimo que maximize a margem entre as classes. Utiliza funções de kernel para lidar com dados não linearmente separáveis, sendo eficaz em espaços de alta dimensionalidade [7].
- **K-Nearest Neighbors (KNN)**: Classificador baseado em instâncias que atribui a classe de uma nova amostra de acordo com as classes dos seus k vizinhos mais próximos. É simples e eficaz, mas sensível à escala dos dados e ao valor de k [4].
- **Naive Bayes**: Modelo probabilístico baseado no Teorema de Bayes, assumindo independência condicional entre as variáveis preditoras. É rápido, eficiente e adequado para dados textuais e classificações em tempo real [4].
- **Gradient Boosting**: Técnica de ensemble que constrói sequencialmente modelos fracos (geralmente árvores de decisão), onde cada novo modelo corrige os erros dos anteriores. Apresenta alto desempenho e boa capacidade de generalização [4].

4 AVALIAÇÃO DOS RESULTADOS

A partir das métricas estabelecidas anteriormente, o modelo está pronto para ser executado. Ao fazer o cross-validation, os resultados das métricas de performance dos modelos são:

Tabela 4: Métricas de desempenho médio dos modelos de classificação

Modelo	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	94.6	94.8	90.6	92.5
Random Forest	95.4	94.6	93.4	93.8
SVM	91.4	96.3	80.2	87.2
KNN	93.5	93.4	89.2	91.0
Naive Bayes	93.9	94.9	88.7	91.4
Gradient Boosting	95.1	95.7	91.1	93.1

A partir desses dados, o modelo Random Forest se mostrou mais eficiente em três métricas diferentes. Os outros modelos aparentemente têm performances semelhantes, mas quando comparado o melhor modelo ao pior em questão de eficiência, o resultado se atenua: 4% em acurácia, 2,9% em precisão, 13,2% em Recall e 6,6% em F1-Score.

4.1 Curva ROC

Outra medida importante para mostrar a frequência de acertos ao longo do treinamento é a curva ROC. Essa curva é determinada em um plano cartesiano no qual o eixo x corresponde à taxa de falsos positivos, enquanto o eixo y corresponde à taxa de verdadeiros positivos. Os resultados dos modelos estão na imagem a seguir:

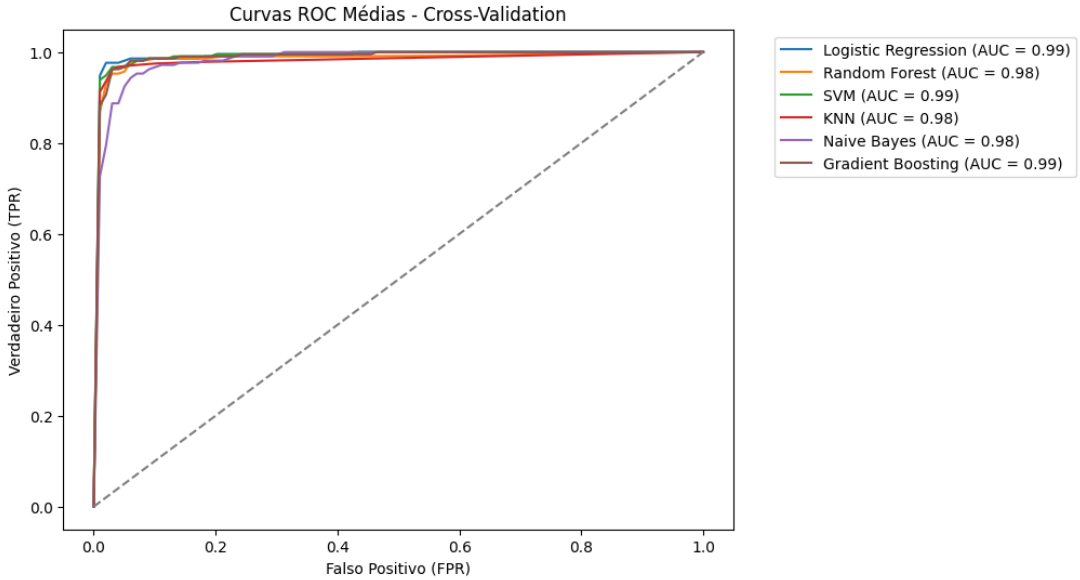

Figura 2: Curva ROC dos Modelos

A partir da curva ROC, é possível ver que os modelos se comportam de maneira muito semelhante, distinguindo-se em poucos dados. Todos os métodos de classificação produzem resultados muito parecidos com essa base de dados.

4.2 **Matriz de Confusão**

A métrica por trás da curva ROC se constitui na matriz de confusão. Ela contabiliza o número de dados verdadeiramente e falsamente classificados. Esse número vem em um gráfico no qual o número dos quadrados superior esquerdo e inferior direito são, respectivamente, verdadeiro negativo e verdadeiro positivo.

Já os dados falsos positivos e falsos negativos estão, nessa ordem, nos quadrados dos cantos superior direito e inferior esquerdo. A seguir, um exemplo de matriz de confusão do modelo Logistic Regression:

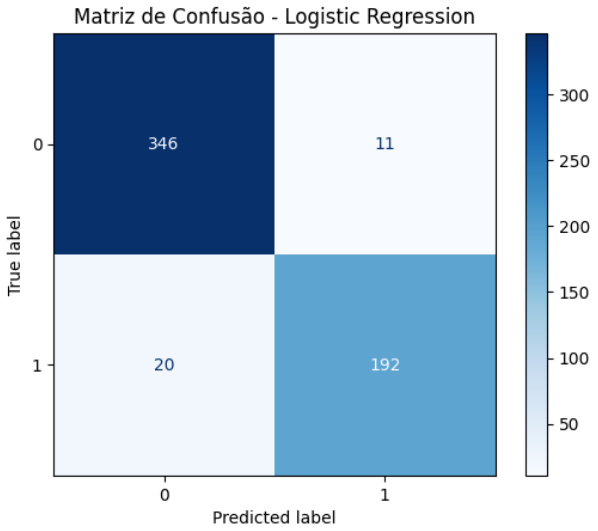

Figura 3: Matriz de Confusão do Modelo Logistic Regression

Observando as outras matrizes dos outros modelos, os resultados dos dados classificados são:

Tabela 5: Matrizes de confusão dos modelos

Modelo	**TN**	**FP**	**FN**	**TP**
Logistic Regression	346	11	20	192
Random Forest	345	12	14	198
SVM	350	7	42	170
KNN	343	14	23	189
Naive Bayes	346	11	24	188
Gradient Boosting	348	9	19	193

A partir dessa tabela, os modelos que apresentaram menos dados falsamente classificados são: Random Forest, com 26 dados, e Gradient Boosting, com 28 dados. Os modelos Logistic Regression, Naive Bayes, KNN e SVM apresentam, respectivamente, 31, 35, 37 e 49 dados falsamente preditos. Os dois melhores modelos são construídos a partir de árvores de decisão, por isso contêm desempenho parecido quando postos em uma mesma base de dados.

5 CONCLUSÃO E PASSOS FUTUROS

A partir da construção dos dados e dos modelos, é possível notar que a base de dados é muito bem tratada e fácil de trabalhar. No entanto, considerando a dimensão e a precisão da finalidade desses classificadores, é necessário trabalhar com uma quantidade muito maior de informações, visto que o câncer adquire múltiplas formas em diversos locais do corpo [1], nas quais as bases de dados pequenas não podem substituir. Além disso, colocar informações de casos de tumores mais excepcionais aumenta ainda mais a acurácia dos modelos de aprendizado de máquina, permitindo que os classificadores possam delimitar melhor as características dos tumores malignos e benignos. Além disso, colocar informações de casos de tumores mais excepcionais aumenta ainda mais a acurácia dos modelos de aprendizado de máquina, permitindo que os classificadores possam delimitar melhor as características dos tumores malignos e benignos.

Outro ponto a ser considerado é o aprofundamento do pré-processamento de dados. Trabalhar em um algoritmo que exclua dados não muito decisivos no início do treinamento dos modelos pode melhorar o modelo trabalhado em grandes volumes de dados, sem comprometer o projeto.

Por fim, o Random Forest se mostrou o melhor modelo para a classificação de tumores, obtendo a melhor performance em várias métricas. Trabalhando os seus parâmetros internos para encontrar o número de árvores, profundidade máxima e amostras, ele se destacará ainda mais entre os outros modelos, melhorando as análises clínicas computadorizadas sobre tumores.

REFERÊNCIAS

[1] SARTORI, A. C. N.; BASSO, C. S. CÂNCER DE MAMA: UMA BREVE REVISÃO DE LITERATURA. *Perspectiva*, p. 7–13, Fev 2019.

[2] (INCA), I. N. de C. *Câncer de mama*. 2025. Publicado em 04/06/2022; atualizado em 29/04/2025. Available at: <https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/mama>.

[3] ELMORE, J. G. et al. Variability in interpretive performance at screening mammography and radiologists' judgments. *New England Journal of Medicine*, v. 352, n. 23, p. 2349–2357, 2005.

[4] TAHMOORESI, M. et al. Early detection of breast cancer using machine learning techniques. *Journal of Telecommunication, Electronic and Computer Engineering*, Asia Pacific University of Technology and Innovation (APU), v. 10, n. 3-2, p. 21–27, 2018.

[5] ALASWAD, N. *Analysis Breast Cancer Prediction Dataset*. 2023. `https://www.kaggle.com/code/nancyalaswad90/analysis-breast-cancer-prediction-dataset`.

[6] scikit-learn. *3.3. Metrics and scoring: quantifying the quality of predictions*. 2025. `https://scikit-learn.org/stable/modules/model_evaluation.html`.

[7] JANA, M. *Exploring Machine Learning Models: A Comprehensive Comparison of Logistic Regression, Decision Trees, SVM, Random Forest, and XGBoost*. 2023.