**CODERHOUSE**

**Data Science**

William Ataide

**Revolucionando a Detecção do Câncer de Mama: Uma Abordagem de Machine Learning** Integrando Ciência de Dados na Luta Contra o Câncer

Betim – Minas Gerais

2023

Internacional Comissária de Despachos Aduaneiros

**Revolucionando a Detecção do Câncer de Mama: Uma Abordagem de Machine Learning** Integrando Ciência de Dados na Luta Contra o Câncer

Artigo apresentado ao curso de Data Science, como parte dos requisitos necessários à obtenção do título de Cientista de Dados.

Orientador: Prof: Wesley Furriel.

Tutor: Fábio Lopes Gemelli.

Betim – Minas Gerais

2023

**RESUMO**

O estudo explora a aplicação de ciência de dados e algoritmos de aprendizado de máquina no diagnóstico ou previsão de câncer de mama. Utilizando técnicas de Machine Learning, o estudo se alinha ao Outubro Rosa, mês de conscientização sobre o câncer de mama. A metodologia inclui análise de dados e modelagem preditiva com ferramentas Python como pandas, numpy, matplotlib, seaborn, pycaret e variações da scikit-learn. Seguindo o princípio de Pareto, 80% dos dados foram usados no treinamento do algoritmo e 20% para teste. O modelo de classificação random forest alcançou 79,5% de sucesso nos dados de teste. Este avanço na inteligência artificial beneficia a saúde, especialmente na detecção de câncer, apoiando os médicos no diagnóstico. Contudo, ressalta-se que modelos confiáveis não substituem o diagnóstico precoce, a principal feature no combate ao câncer de mama.

**1 INTRODUÇÃO**

O câncer de mama permanece como uma das principais causas de mortalidade entre mulheres em todo o mundo, desafiando os sistemas de saúde e a comunidade científica. No entanto, com os avanços tecnológicos, particularmente na ciência de dados e no aprendizado de máquina (machine learning), surgiram novas perspectivas para o diagnóstico precoce e mais preciso dessa doença. O presente artigo explora a intersecção entre a ciência de dados e a oncologia, focalizando como algoritmos de machine learning podem ser empregados para aprimorar a detecção e o diagnóstico do câncer de mama.

Neste contexto, o mês de Outubro Rosa, um movimento internacional de conscientização para o controle do câncer de mama, serve como pano de fundo ideal para esta investigação. O artigo baseia-se em um estudo de caso detalhado, utilizando um conjunto de dados públicos que inclui variáveis chave relacionadas ao núcleo do tumor. Através de uma metodologia que combina teoria e prática, examinamos como diferentes técnicas de processamento de dados e algoritmos de aprendizado de máquina podem ser aplicados para identificar padrões e fazer previsões precisas sobre a presença de células cancerígenas.

Com uma abordagem que integra a análise exploratória de dados, a pré-processamento e a modelagem preditiva, este artigo não só demonstra a aplicabilidade prática da ciência de dados no campo da saúde, mas também oferece insights sobre os desafios e as oportunidades que emergem dessa intersecção. O objetivo é não apenas fornecer uma compreensão técnica das ferramentas e técnicas utilizadas, mas também destacar a importância da inovação tecnológica na luta contra o câncer de mama, ressaltando o potencial de salvar vidas através da detecção precoce e intervenção oportuna.

**3 METODOLOGIA**

A base de dados utilizada neste estudo contém 570 entradas, cada uma representando um caso de diagnóstico de câncer de mama. Os dados são bastante abrangentes, incluindo um total de 33 colunas. Entre elas, encontramos características como o raio médio das células tumorais (radius\_mean), textura (texture\_mean), perímetro (perimeter\_mean), área (area\_mean), suavidade (smoothness\_mean), compactação (compactness\_mean), concavidade (concavity\_mean), pontos côncavos (concave points\_mean), simetria (symmetry\_mean) e dimensão fractal (fractal\_dimension\_mean). Estes atributos são repetidos sob diferentes condições: 'mean' (média), 'se' (erro padrão) e 'worst' (pior caso). Além disso, cada registro é identificado por um 'id' e inclui um diagnóstico ('diagnosis'), que é a variável alvo, indicando se o tumor é maligno ('M') ou benigno ('B').

Com base neste conjunto de dados, o estudo seguiu a seguinte metodologia:

Preparação e Análise Exploratória de Dados: Inicialmente, conduzimos uma análise exploratória para familiarizar-nos com o conjunto de dados. Isso incluiu a verificação de valores faltantes, a análise de distribuições de variáveis e a exploração de correlações entre diferentes atributos. Essencialmente, esta fase visava entender melhor as características dos dados e prepará-los para análises subsequentes.

Pré-processamento de Dados: Após a análise exploratória, procedemos com o pré-processamento dos dados. Isso envolveu tarefas como normalização das variáveis para garantir que todas tivessem o mesmo peso durante o treinamento do modelo de machine learning e o tratamento de quaisquer inconsistências ou valores ausentes encontrados.

Divisão do Conjunto de Dados: Dividimos os dados em dois subconjuntos: treinamento e teste. O conjunto de treinamento foi usado para treinar o modelo de machine learning, enquanto o conjunto de teste serviu para avaliar o desempenho do modelo em dados não vistos. A divisão típica adotada foi de 80% para treinamento e 20% para teste.

Seleção e Treinamento do Modelo de Machine Learning: O modelo selecionado e treinado, Random Forest. O modelo foi ajustado e otimizado com base no conjunto de treinamento.

Avaliação de Modelos: Os modelos foram avaliados com base em métricas como a área sob a curva ROC. Isso nos permitiu não apenas determinar a eficácia de cada modelo, mas também garantir sua capacidade de distinguir corretamente entre diagnósticos malignos e benignos.

Aplicação em Dados de Teste: Finalmente, o modelo treinado foi aplicado ao conjunto de teste para avaliar seu desempenho em condições reais, proporcionando uma estimativa de como eles se comportariam na prática clínica.

Esta abordagem metodológica assegurou uma análise robusta e detalhada do conjunto de dados, destacando a capacidade do algoritmo de machine learning de contribuir significativamente para o diagnóstico do câncer de mama.

**4 ANÁLISE DOS RESULTADOS**

Desempenho do Modelo: O modelo de classificação Random Forest, que foi o algoritmo de machine learning aplicado neste estudo, alcançou um sucesso significativo. O modelo obteve uma taxa de acerto de 79,5% nos dados de teste, que correspondem a 20% do conjunto total de dados. Este resultado é indicativo de uma boa performance, especialmente considerando a complexidade e a variabilidade inerente aos dados biomédicos.

Importância da Inteligência Artificial na Saúde: O estudo reforça a importância crescente das técnicas de inteligência artificial, particularmente os algoritmos de machine learning, no campo da saúde. A aplicação dessas técnicas na detecção de câncer tem se mostrado uma ferramenta valiosa, auxiliando os médicos no diagnóstico preciso e, consequentemente, no tratamento mais eficaz dos pacientes.

Diagnóstico Precoce como Chave: Embora os modelos de machine learning, como o Random Forest, tenham demonstrado eficácia na detecção de câncer de mama, o estudo enfatiza que eles não substituem a importância do diagnóstico precoce. O diagnóstico precoce continua sendo a "principal feature", ou seja, o aspecto mais crucial no tratamento bem-sucedido do câncer de mama. Isso destaca a necessidade de um equilíbrio entre as inovações tecnológicas e as práticas clínicas tradicionais.

Análise Geral

A análise sugere que a integração de técnicas avançadas de ciência de dados e machine learning na área da saúde tem um potencial significativo. O sucesso do modelo Random Forest no estudo é um testemunho disso. No entanto, a ênfase no diagnóstico precoce serve como um lembrete vital de que, enquanto a tecnologia pode aumentar significativamente a eficácia do diagnóstico e do tratamento, ela deve ser usada em conjunto com práticas clínicas estabelecidas e não como um substituto para elas. ​