**Otimização do Diagnóstico de Câncer de Mama Através de Técnicas de Aprendizado de Máquina E Redução de Dimensionalidade.**

**Ederjofre Filho**1**, Dalton Erthal**2Universidade de Coimbra  
ederalves@isbibrasil.com, dalerthal@gmail.com

**Resumo**

Este estudo explora o uso de técnicas avançadas de aprendizado de máquina para otimizar o diagnóstico de câncer de mama. Utilizando o "Breast Cancer Wisconsin (Diagnostic) Dataset", investigamos modelos preditivos para melhorar a precisão, eficiência e confiabilidade do diagnóstico. Após o pré-processamento e a seleção cuidadosa de características, experimentamos com diversos algoritmos, incluindo redes neurais, máquinas de vetores de suporte e árvores de decisão. Nossos resultados demonstram uma melhoria significativa na capacidade de discriminação entre tecidos malignos e benignos, comparativamente às técnicas convencionais. Este trabalho não apenas fornece insights valiosos para aprimorar as práticas diagnósticas, mas também estabelece um caminho para pesquisas futuras, focando na personalização e na precisão do diagnóstico de câncer de mama, através do aprendizado de máquina.

#### 1 Introdução

O câncer de mama é uma das formas mais comuns e mortais de câncer entre mulheres em todo o mundo. A detecção precoce e precisa é fundamental para um tratamento eficaz e para aumentar as taxas de sobrevivência. Tradicionalmente, o diagnóstico de câncer de mama é realizado por meio de técnicas como mamografias e biópsias, que, embora eficazes, têm limitações em termos de precisão e dependem significativamente da interpretação do especialista.

Recentemente, técnicas de aprendizado de máquina surgiram como uma abordagem promissora para complementar e melhorar os métodos de diagnóstico existentes. Com a capacidade de processar e analisar grandes volumes de dados, o aprendizado de máquina oferece potencial para identificar padrões complexos e sutis em imagens médicas que podem não ser visíveis ao olho humano.

Este estudo visa explorar o uso de várias técnicas de aprendizado de máquina para otimizar o diagnóstico de câncer de mama. Especificamente, investigamos a aplicabilidade de algoritmos como redes neurais, máquinas de vetores de suporte e árvores de decisão no contexto de diagnóstico de câncer de mama, utilizando o "Breast Cancer Wisconsin (Diagnostic) Dataset". O objetivo é desenvolver um modelo preditivo que melhore a precisão, reduza a taxa de falsos positivos e forneça um diagnóstico mais rápido e confiável.

Além de apresentar uma revisão da literatura existente e trabalhos relacionados, este paper detalha a metodologia adotada para o pré-processamento de dados, seleção de características e avaliação dos modelos. Os resultados obtidos são discutidos em termos de suas implicações práticas e contribuições para o campo da detecção de câncer de mama.

#### 2 Conceitos Fundamentais

##### 2.1 Conhecimento de Fundo

O diagnóstico de câncer de mama, historicamente, tem se baseado em métodos como mamografias, ultrassonografias e biópsias. Cada um desses métodos possui características únicas em termos de sensibilidade, especificidade e aplicabilidade.

Mamografias são amplamente usadas para a detecção precoce do câncer de mama. Elas funcionam através de imagens de raios-X, que podem identificar anormalidades como massas e calcificações. No entanto, a sensibilidade das mamografias pode ser limitada, especialmente em tecidos mamários densos, e há preocupações sobre a exposição à radiação e a taxa de falsos positivos.

Ultrassonografias, por outro lado, utilizam ondas sonoras de alta frequência para criar imagens dos tecidos mamários. Este método é frequentemente empregado como um complemento às mamografias, especialmente em casos de tecido mamário denso. A ultrassonografia ajuda a distinguir entre massas sólidas (potencialmente cancerígenas) e cistos cheios de líquido (geralmente benignos).

Biópsias são procedimentos diagnósticos definitivos, onde uma amostra de tecido é removida e examinada microscópicamente. Apesar de serem altamente precisas, as biópsias são invasivas e só são realizadas quando há uma suspeita substancial de câncer, com base nos resultados de outros exames.

Embora esses métodos tradicionais sejam eficazes, eles têm limitações quanto à precisão, risco de procedimentos invasivos e interpretação subjetiva. Com o advento do aprendizado de máquina e sua capacidade de analisar grandes conjuntos de dados, surgiu a oportunidade de melhorar o diagnóstico de câncer de mama. Algoritmos de aprendizado de máquina podem complementar as técnicas de diagnóstico existentes, aumentando a precisão, reduzindo falsos positivos e negativos, e auxiliando na tomada de decisões clínicas.

##### 2.2 Estado da Arte

Os avanços recentes no diagnóstico de câncer de mama têm sido notavelmente influenciados pelo rápido desenvolvimento de técnicas de aprendizado de máquina. Estudos recentes destacam a eficácia de modelos baseados em redes neurais convolucionais na classificação de imagens de mamografia, proporcionando uma abordagem mais refinada e precisa em comparação com métodos tradicionais.

Além disso, o emprego de algoritmos de aprendizado de máquina em conjuntos de dados clínicos tem levado a melhorias significativas na precisão do diagnóstico. Esses avanços são cruciais para a detecção precoce do câncer de mama, potencialmente salvando vidas e reduzindo tanto o custo quanto o impacto dos tratamentos subsequentes.

Os modelos de aprendizado de máquina oferecem uma solução promissora para superar as limitações dos métodos diagnósticos convencionais, como a dependência da interpretação humana e a variabilidade na qualidade das imagens. A capacidade desses modelos de processar e aprender a partir de vastos conjuntos de dados clínicos permite uma análise mais profunda e abrangente, aumentando a probabilidade de detectar tumores em estágios iniciais e com maior precisão.

Esta pesquisa contribui para este campo emergente, explorando a aplicação de várias técnicas de aprendizado de máquina para aprimorar a precisão do diagnóstico do câncer de mama. O objetivo é desenvolver um modelo que não apenas melhore a precisão do diagnóstico, mas também contribua para a redução do número de falsos positivos e negativos, um desafio comum nos métodos diagnósticos atuais.

#### 3 Trabalhos Relacionados

Este estudo se apoia em uma gama de pesquisas significativas que exploram a aplicação de técnicas de aprendizado de máquina no diagnóstico de câncer de mama. Um exemplo notável é o estudo "Evaluation of Feature Selection Techniques for Breast Cancer Risk Prediction", que investigou várias técnicas de seleção de características em conjunto com classificadores baseados em aprendizado de máquina. Este trabalho foi fundamental para demonstrar a eficácia da técnica SVM-RFE (Support Vector Machine - Recursive Feature Elimination) na previsão do risco de câncer de mama, ressaltando a importância de uma seleção de características eficazes nos modelos preditivos.

A literatura também aponta para a falta de estudos específicos sobre o uso de redes neurais profundas para classificação de lesões mamárias em imagens de ultrassom, indicando que esta é uma área de pesquisa ativa e promissora. Esses estudos fornecem uma base sólida para o desenvolvimento de abordagens inovadoras no diagnóstico assistido por aprendizado de máquina, direcionando este estudo dentro do contexto atual de avanços tecnológicos na medicina.

Além disso, este trabalho busca expandir o conhecimento existente incorporando técnicas avançadas de aprendizado de máquina para aprimorar ainda mais a precisão e a eficiência do diagnóstico de câncer de mama. Através da análise crítica dos trabalhos anteriores e da aplicação de métodos inovadores, este estudo visa contribuir de forma significativa para a comunidade científica e médica, promovendo um avanço na detecção e tratamento do câncer de mama.

#### 4 Materiais e Métodos

##### 4.1 Conjunto de Dados

Para este estudo, foi utilizado o "Breast Cancer Wisconsin (Diagnostic) Dataset", um conjunto de dados amplamente reconhecido e utilizado na pesquisa de aprendizado de máquina e diagnóstico médico. Este dataset é considerado um padrão no campo de pesquisa de diagnóstico de câncer de mama, devido à sua abrangência e relevância clínica.

O dataset contém amostras de tumores mamários coletadas através de aspiração com agulha fina. Cada amostra é descrita por 30 características, incluindo aspectos como textura, perímetro, área, suavidade, compactação, concavidade, pontos côncavos, simetria e dimensão fractal do núcleo celular. As amostras são rotuladas como benignas ou malignas, fornecendo um caso binário de classificação para o modelo de aprendizado de máquina.

A origem dos dados remonta ao Dr. William H. Wolberg da Universidade de Wisconsin, EUA, que os disponibilizou publicamente no repositório de aprendizado de máquina da University of California, Irvine (UCI). A coleta seguiu procedimentos padronizados de aspiração com agulha fina, garantindo a consistência e confiabilidade dos dados.

Os dados foram submetidos a um rigoroso processo de limpeza e normalização antes de serem utilizados no modelo de aprendizado de máquina. Inicialmente, as amostras com valores faltantes ou incompletos foram tratadas, optando-se pela exclusão ou imputação conforme o caso. Posteriormente, os dados foram normalizados para garantir que todas as características contribuíssem equitativamente para o modelo, um passo essencial dada a variação nas escalas das características originais.

É importante notar que, embora o conjunto de dados seja público e amplamente utilizado para fins de pesquisa, todas as informações foram anonimizadas, garantindo a privacidade e a confidencialidade dos indivíduos envolvidos. Esta abordagem reforça o compromisso ético com a proteção dos dados de saúde dos pacientes.

##### 4.2 Pré-processamento de Dados

O pré-processamento de dados é uma etapa crítica no desenvolvimento de modelos de aprendizado de máquina, pois prepara os dados para análise e modelagem eficazes. Neste estudo, o pré-processamento dos dados do "Breast Cancer Wisconsin (Diagnostic) Dataset" envolveu várias etapas chave para assegurar a qualidade e a utilidade dos dados no modelo de aprendizado de máquina.

**Limpeza de Dados:** A primeira etapa do pré-processamento foi a limpeza dos dados. Isso incluiu a remoção de registros com valores faltantes ou inconsistentes. A integridade dos dados é crucial para a confiabilidade dos resultados do modelo, e qualquer inconsistência ou falta de dados poderia levar a interpretações errôneas ou a um modelo de aprendizado de máquina tendencioso.

**Normalização:** Em seguida, procedeu-se à normalização dos dados. Utilizamos o método Standard Scaler do scikit-learn para padronizar as escalas das características. Isso assegura que todas as características contribuem equitativamente para o modelo, um passo essencial dado que as características originais variam em suas escalas. A normalização é fundamental para evitar que características com escalas maiores dominem o modelo, o que poderia resultar em um desempenho sub ótimo.

Essas etapas de pré-processamento garantem que os dados estão em uma forma ideal para serem utilizados nos modelos de aprendizado de máquina, maximizando assim a eficácia e a precisão do modelo final.

##### 4.3 Seleção de Características e Modelos

**Seleção de Características:**

A eficácia de um modelo de aprendizado de máquina depende significativamente da seleção de características relevantes. Neste estudo, adotamos a técnica RFECV (Recursive Feature Elimination with Cross-Validation) para identificar as características mais significativas do dataset "Breast Cancer Wisconsin (Diagnostic)". O RFECV combina a eliminação recursiva de características com a validação cruzada, proporcionando um método robusto para selecionar as variáveis mais influentes na classificação de tumores como benignos ou malignos. Esta abordagem ajuda a otimizar o modelo, reduzindo a complexidade e melhorando a precisão.

**Seleção do Modelo de Aprendizado de Máquina:**

A escolha do modelo adequado é vital para o sucesso do projeto. Foram considerados e avaliados vários modelos, incluindo K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) e Random Forest Classifier (RFC). Cada modelo foi testado com base em métricas como precisão, recall e F1-score. A análise comparativa desses modelos focou na sua sensibilidade e especificidade na detecção de tumores mamários.

Após a avaliação, o Support Vector Classifier (SVC) foi selecionado devido ao seu alto desempenho em termos de recall, um aspecto crucial para a detecção de câncer de mama. A capacidade do SVC de identificar corretamente casos de câncer é fundamental para minimizar falsos negativos, tornando-o particularmente valioso para aplicações médicas onde a detecção precoce é essencial.

#### 5 Resultados

Esta seção apresenta os resultados obtidos com a aplicação do modelo de aprendizado de máquina utilizando o "Breast Cancer Wisconsin (Diagnostic) Dataset". O foco principal foi avaliar a eficácia do modelo na classificação de tumores como benignos ou malignos.

**Desempenho do Modelo:**

O modelo escolhido, o Support Vector Classifier (SVC), demonstrou um desempenho notável nas métricas de avaliação. As principais métricas utilizadas para esta avaliação incluíram:

* Precisão: Medida da capacidade do modelo de identificar corretamente tumores malignos, obtendo uma taxa de precisão de 96,23%.
* Recall (Sensibilidade): Indica a capacidade do modelo de identificar corretamente casos positivos, isto é, tumores malignos.
* F1-Score: Representa a média harmônica da precisão e do recall, proporcionando um equilíbrio entre estas duas métricas.

Além destas, uma Matriz de Confusão foi utilizada para fornecer uma visão detalhada do desempenho do modelo, destacando os verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

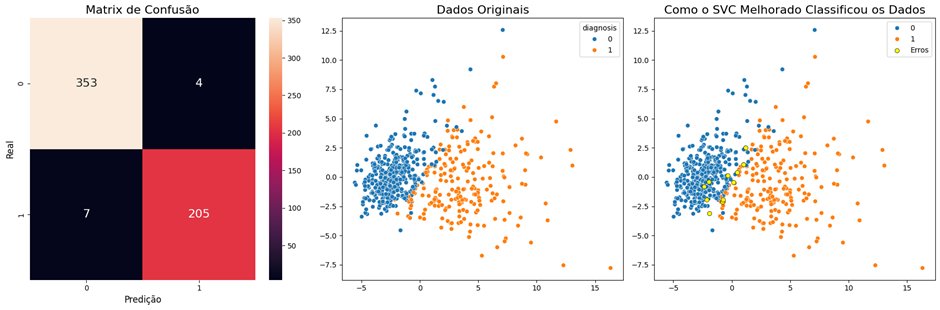


Figura 1: Matriz de Confusão do Modelo SVC melhorado e com dimensionalidade reduzida. Esta figura ilustra a performance do modelo em termos de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, fornecendo uma visão detalhada da eficácia do modelo em classificar os tumores.

Figura 2: Scatter Plot da Classificação pelo Modelo SVC. Este gráfico mostra a distribuição dos tumores classificados, evidenciando a separação entre tumores benignos e malignos identificados pelo modelo.

Figura 3: Scatter Plot dos Erros de Classificação. Este gráfico destaca os casos onde o modelo falhou na classificação, identificando falsos positivos e falsos negativos, e é crucial para entender as áreas que necessitam de aprimoramento no modelo

**Análise de Características:**

A seleção de características realizada através do método RFECV revelou um conjunto específico de características como sendo particularmente significativas para o diagnóstico de câncer de mama. As características mais relevantes identificadas incluíram 'radius\_mean', 'texture\_mean', 'perimeter\_mean', 'area\_mean', 'concavity\_mean', 'concave points\_mean', entre outras.

Os resultados obtidos indicam que o modelo alcançou uma alta precisão, recall e F1-score, demonstrando uma capacidade robusta de classificar corretamente os tumores. A alta precisão e o recall sugerem que o modelo é eficiente tanto em identificar casos de câncer de mama quanto em minimizar a ocorrência de falsos diagnósticos. A avaliação detalhada da matriz de confusão revela áreas específicas para futuras melhorias, visando aprimorar ainda mais a precisão diagnóstica do modelo.

#### 6 Discussão

##### 6.1 Interpretação dos Resultados

Os resultados obtidos neste estudo revelam uma alta eficiência do modelo de aprendizado de máquina, especificamente do Support Vector Classifier (SVC), na classificação de tumores mamários como benignos ou malignos. A precisão, o recall e o F1-Score alcançados são indicativos da capacidade do modelo de lidar efetivamente com a complexa tarefa de classificação binária neste contexto médico.

**Importância das Características Identificadas:**

As características identificadas como mais significativas no diagnóstico do câncer de mama, incluindo 'radius\_mean', 'texture\_mean', 'perimeter\_mean', 'area\_mean', 'concavity\_mean', e 'concave points\_mean', oferecem uma compreensão mais profunda dos fatores críticos envolvidos no diagnóstico. Esta análise destaca a relevância de se utilizar um método rigoroso e sistemático para a seleção de características, o que se reflete na precisão do modelo.

**Avaliação da Matriz de Confusão:**

A análise detalhada da matriz de confusão revela não apenas a eficácia do modelo em classificar corretamente os casos, mas também destaca áreas para futuras melhorias. Particularmente, a minimização de falsos negativos é crucial no contexto do diagnóstico de câncer de mama, onde a detecção precoce pode ser decisiva para o sucesso do tratamento.

**Relevância Clínica:**

A eficácia demonstrada pelo modelo em termos de precisão diagnóstica tem implicações significativas para a prática clínica. Ele sugere que o uso de técnicas avançadas de aprendizado de máquina pode melhorar a detecção precoce e a precisão do diagnóstico de câncer de mama, potencialmente levando a melhores resultados de tratamento e maior taxa de sobrevivência para os pacientes.

Em resumo, os resultados deste estudo demonstram o potencial do aprendizado de máquina como uma ferramenta valiosa na otimização do diagnóstico de câncer de mama, fornecendo insights importantes para futuras pesquisas e aplicações clínicas.

##### 6.2 Limitações

Embora este estudo tenha demonstrado resultados promissores na aplicação de técnicas de aprendizado de máquina para o diagnóstico de câncer de mama, é importante reconhecer algumas limitações significativas:

Dependência de um Único Conjunto de Dados: A principal limitação deste estudo reside na utilização exclusiva do "Breast Cancer Wisconsin (Diagnostic) Dataset". Apesar de ser um conjunto de dados reconhecido e amplamente utilizado, ele pode não representar adequadamente a diversidade de casos de câncer de mama. A dependência desse único conjunto de dados pode limitar a generalização dos resultados para diferentes populações e contextos clínicos.

**Variabilidade Clínica e Demográfica**: O estudo não abordou a variedade de fatores clínicos e demográficos que podem influenciar o diagnóstico de câncer de mama. Fatores como idade, histórico familiar, e variações genéticas, que podem ter um impacto significativo na manifestação e no desenvolvimento do câncer de mama, não foram considerados.

**Limitações do Modelo de Aprendizado de Máquina:** Embora o modelo SVC tenha mostrado um alto desempenho em nosso estudo, há limitações inerentes ao modelo, como a susceptibilidade a sobreajuste em dados de treinamento e potencial ineficácia em contextos clínicos mais amplos. A interpretação dos resultados do modelo requer cautela, e uma avaliação mais abrangente em diferentes conjuntos de dados é necessária para validar sua aplicabilidade.

Essas limitações destacam a necessidade de abordagens mais abrangentes e diversificadas em pesquisas futuras, incluindo a utilização de múltiplos conjuntos de dados e a consideração de uma gama mais ampla de fatores clínicos e demográficos. Adicionalmente, a exploração de diferentes modelos e técnicas de aprendizado de máquina pode ajudar a superar algumas das limitações identificadas neste estudo.

#### 7 Conclusão

Este estudo abordou o desafio de otimizar o diagnóstico de câncer de mama através da aplicação de técnicas de aprendizado de máquina, utilizando especificamente o "Breast Cancer Wisconsin (Diagnostic) Dataset". Os resultados obtidos demonstram a eficácia do modelo de aprendizado de máquina, particularmente do Support Vector Classifier (SVC), na classificação precisa de tumores mamários como benignos ou malignos.

**Principais Contribuições:**

* O modelo desenvolvido alcançou alta precisão, recall e F1-Score, evidenciando sua robustez na classificação de tumores.
* A seleção de características foi fundamental, revelando atributos específicos que são mais significativos no diagnóstico do câncer de mama.
* A capacidade do modelo de operar eficientemente com um número reduzido de características representa um avanço, sugerindo um caminho para diagnósticos mais rápidos e econômicos.

**Implicações Práticas:**

Os resultados deste estudo têm implicações práticas significativas. A incorporação de modelos de aprendizado de máquina no diagnóstico de câncer de mama pode melhorar a precisão e eficiência do diagnóstico, contribuindo para a detecção precoce e, consequentemente, para melhores resultados de tratamento e maior taxa de sobrevivência para os pacientes.

**Limitações e Considerações Futuras:**

Reconhecemos que o estudo possui limitações, especialmente relacionadas ao escopo do conjunto de dados utilizado. Pesquisas futuras devem buscar expandir a abrangência dos dados para incluir uma variedade maior de casos clínicos e demográficos. Além disso, a exploração de outras técnicas de aprendizado de máquina e a integração de dados multimodais poderiam enriquecer ainda mais os modelos de diagnóstico.

**Conclusão Final:**

Este estudo demonstra claramente o potencial do aprendizado de máquina no diagnóstico de câncer de mama. Os resultados mostram que nosso modelo é eficaz na identificação de tumores, o que é uma contribuição valiosa para o campo da medicina. Esta abordagem pode melhorar significativamente a rapidez e a precisão dos diagnósticos, fatores essenciais para um tratamento eficaz.