Análise comparativa de desempenho dos algoritmos KNN, SVM e Árvores de Decisão no diagnóstico de câncer de mama

1st Ezequiel Gonçalves
I.C.E.T.
Universidade Federal de Viçosa
Rio Paranaíba, Brasil
ezequiel.goncalves@ufv.br

2nd Jonatan Henrique da Silva
I.C.E.T.
Universidade Federal de Viçosa
Rio Paranaíba, Brasil
jonatan.silva@ufv.br

3rd Thamara Melo
I.C.E.T.
Universidade Federal de Viçosa
Rio Paranaíba, Brasil
thamara.melo@ufv.br

***Resumo*—O câncer de mama é o tipo mais prevalente entre as mulheres no Brasil e o segundo mais comum globalmente, representando 28% dos novos casos diagnosticados no país [1]. Embora os exames clínicos e de imagem sejam métodos investigativos comuns, a confirmação diagnóstica ocorre principalmente por meio da biópsia. Diagnósticos imprecisos podem criar uma falsa sensação de segurança, impactando diretamente a taxa de cura, notavelmente maior quando o câncer é detectado precocemente [1].Este artigo apresenta uma análise comparativa de desempenho de três algoritmos utilizados para o diagnóstico de câncer de mama: KNN (K-nearest neighbors), SVM (Support Vector Machine) e Árvore de Decisão. O método baseia-se na análise de amostras em formato CSV, classificadas como malignas ou benignas, oferecendo uma visão sobre a eficácia desses algoritmos na detecção precoce e precisa do câncer de mama. Após o pré-processamento dos dados para remover colunas e informações irrelevantes à classificação, métricas de avaliação foram aplicadas após o treinamento. Cada algoritmo foi avaliado individualmente usando uma matriz de confusão para calcular acurácia, precisão, revocação e F1-Score. A comparação dos resultados revelou que o método SVM demonstrou ser o mais promissor, alcançando uma acurácia de 98%, superando os outros métodos. Destaca-se sua precisão de 100% na identificação de células malignas. Enquanto isso, tanto o KNN quanto a Árvore de Decisão apresentaram resultados semelhantes, com uma acurácia de 95% e uma precisão de 93% na detecção de células malignas.**

***Palavras-chave*—câncer de mama, KNN, SVM, Árvore de Decisão**

I. Introdução

O câncer engloba uma ampla variedade de condições, incluindo mais de 100 tipos de doenças [5], todas caracterizadas pelo crescimento celular anormal e descontrolado em tecidos onde não são naturalmente encontrados. Esse desvio no comportamento celular, muitas vezes, resulta de alterações genéticas que comprometem o funcionamento adequado das células, afetando seu ciclo de vida e permitindo o surgimento de células cancerosas em várias partes do corpo. Células que apresentam mutações podem originar tumores benignos ou malignos.

Os tumores malignos se destacam por sua capacidade de invadir os tecidos ao redor e se espalhar. Embora o processo de transformação de uma célula normal em uma célula cancerosa seja gradual, a multiplicação das células malignas ocorre de forma extremamente rápida. Essas células têm a habilidade de migrar pelo corpo, estabelecendo novos focos tumorais em diferentes regiões.

Por outro lado, os tumores benignos são marcados pela falta de invasão e disseminação para outros tecidos. Ao contrário dos tumores malignos, quando um tumor benigno é completamente removido, muitas vezes isso significa o fim definitivo do problema naquela região específica, com pouca chance de recorrência no mesmo local.

O câncer de mama permanece como o tipo mais prevalente de câncer global, correspondendo a aproximadamente 2,3 milhões (11,7%) dos casos novos registrados em 2020 [2]. Notavelmente mais incidente em mulheres, excluindo os tumores de pele não melanoma, no Brasil, esse índice atinge 74 mil (10,5%) casos [2]. Embora diversos fatores de risco contribuam para a suscetibilidade ao câncer de mama, é importante destacar que, em muitos casos, ele pode ser atribuído a mutações genéticas hereditárias. A detecção precoce e o conhecimento aprofundado desempenham um papel crucial na eficácia do tratamento e na perspectiva de cura da doença.

A utilização da Inteligência Artificial (IA) no diagnóstico do câncer de mama tem se destacado como uma abordagem inovadora e promissora. A aplicação do aprendizado de máquina desempenha um papel fundamental na análise e interpretação de dados, permitindo a identificação de padrões que podem passar despercebidos pelos métodos tradicionais. A notável capacidade da IA de assimilar e aprender com grandes volumes de dados contribui para uma análise mais precisa e eficiente, ultrapassando as limitações da observação humana.

O algoritmo KNN é reconhecido como um método de aprendizado supervisionado, notável por sua abordagem de aprendizado lento. Este método se baseia na utilização dos valores de seus vizinhos mais próximos para realizar classificações. A definição dos rótulos ocorre por meio da maioria dos vizinhos, onde a classe mais prevalente é escolhida. A distância entre os pontos de dados é fundamental para determinar a proximidade e, consequentemente, estabelecer os limites de decisão. O parâmetro 'k' desempenha um papel crucial ao definir

a quantidade de vizinhos a serem considerados, geralmente preferindo-se números ímpares para evitar empates durante o processo de decisão [3].

Por outro lado, o Support Vector Machine (SVM) também é um algoritmo de aprendizado supervisionado, capaz de executar tarefas de classificação e regressão. O SVM utiliza vetores para realizar suas classificações, empregando hiperplanos, ótimos para separar e rotular os dados. Esses hiperplanos atuam como fronteiras onde os dados de diferentes classes são segregados, permitindo que os dados de uma classe (rotulados como A, por exemplo) sejam separados dos dados da classe oposta (rotulados como B, por exemplo) [8].

Referente às Árvores de Decisões, sua peculiaridade está na organização hierárquica, onde cada nó representa uma decisão baseada em determinadas características dos dados. Esses nós bifurcam-se em duas ramificações, simbolizando possíveis respostas ou rótulos do algoritmo. Os caminhos formados nessa estrutura representam as regras de decisão, estabelecendo condições que, se satisfeitas pelo nó pai, direcionam para os rótulos correspondentes nas folhas [6].

Uma análise dos algoritmos para fins de comparação pode ser intuitiva na orientação e escolha do melhor algoritmo para desenvolvimento de uma IA para detecção do câncer de mama. Portanto, espera-se que um dos algoritmos se demonstre mais promissor, servindo de modelo para futuras implementações e/ou comparações.

II. Trabalhos Relacionados

[7] realizou uma análise comparativa dos métodos computacionais de classificação de estoque empresarial baseados em: Redes Neurais, KNN e SVM. O [7] efetuou procedimentos relacionados à administração empresarial como: planejamento, controle de compras e estoques, com o intuito de reduzir a variedade demasiada de produtos à lotes menores para que seja feita uma melhor gestão de estoque, reduzindo assim, erros simples como a classificação destes produtos, entendendo-se também que com o auxílio de uma ferramenta computacional "inteligente" pode ser útil até para tomadas de decisões dentro das organizações. Com isso, [7] comparou as performances dos métodos com uma classificação de referência, realizando uma série de etapas para que sejam tratado os dados, analisando características e a seleção de atributos, além do treinamento, validação e avaliação dos modelos para que seja feita uma comparação e este demonstre um resultado favorável à algum dos modelos analisados.

Já [4] utiliza um método de seleção de variáveis provindas de exames clínicos, para efetuar a identificação do câncer de mama por meio da mineração de dados, onde utiliza uma técnica multivariada de Análise de Componentes Principais (PCA) que se aplica ao banco de dados que se produziu por meio da extração deste dos exames. Ele utiliza os métodos K-Vizinhos Mais Próximos (KNN) e Análise Discriminante (AD) para, além de separar e classificar os dados como benignos ou malignos, treinar a IA para efetuar os diagnósticos por meio dos exames.

Para a realização deste presente trabalho, foram considerados a separação e classificação de dados por meio dos algoritmos KNN e SVM e análise propostos por [7], além da utilização de um destes algoritmos para treinamento de diagnósticos com câncer de mama na qual [4] fez uso.

III. Método

A metodologia utilizada para avaliar o desempenho dos algoritmos KNN, SVM e Árvores de Decisão no diagnóstico do câncer de mama foi a aplicação destes algoritmos sob um *dataset* de diagnósticos. O *dataset* escolhido foi retirado do *website* ***Kaggle***, uma plataforma de ciência de dados que oferece uma vasta variedade de conjuntos de dados em várias áreas, sendo este amplamente utilizado pela comunidade. O *dataset* utilizado possui 569 linhas, correspondentes a diferentes diagnósticos, além de 31 colunas, estas por sua vez correnpondendo a diversas características das células presentes no local analisado. A plataforma responsável por executar esta implementação é o *Google Colab*, um ambiente gratuito que possibilita o desenvolvimento de um *Python Notebook*. Com o *dataset* sendo acessado via *Google Drive* compartilhado, procedeu-se a implementação dos algoritmos, fazendo o carregamento dos dados e logo após uma preparação dos mesmos para aplicação da lógica. Feito isso, os dados foram treinados e testados, seguindo as definições e parâmetros necessários. A implementação[1] baseou-se no uso de bibliotecas, principalmente a *scikit-learn*, uma ferramenta robusta na ciência de dados, a qual dispõe das implementações dos três algoritmos utilizados, facilitando o desenvolvimento. Após a aplicação dos algoritmos sob o conjunto de dados, com o auxílio desta mesma ferramenta, foram realizadas análises sobre os resultados, de forma a comparar o desempenho de cada algoritmo no auxílio ao diagnóstico do câncer de mama.

A. Tratamento dos Dados

Para o tratamento dos dados foi escolhido o *dataset* mencionado. Portanto, foram feitas a exclusão de dados nulos e irrelevantes e a modificação de diagnósticos a valores numéricos binários (0 e 1), sendo, 1 para M (maligno) e 0 para B (benigno), podendo assim, facilitar a separação dos dados em váriaveis de teste e treino para serem projetadas e validadas no método a ser aplicado.

B. Aplicação do método KNN

Foram feitas classificações, onde, por parâmetro, é escolhido a quantidade de vizinhos delimitando o escopo do dado. Com isso, foi atribuído um número ímpar(3), pois oferece uma melhor precisão para a separação colocando os dados em uma "grande classe" [7]. Após a configuração do método, este foi treinado e submetido a um teste de previsão para que auxilie na geração de pontos de acurácia. Também foi realizada a aplicação da matriz de confusão, esta que traz valores de Falso Positivos (FP), Falso Negativo (FN), Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), servindo como

[1]Disponível em: https://github.com/EzequielBurg/projeto-ia

base para os cálculos de precisão, *recall*, *f1-score* e acurácia, que reafirmam o valor obtido nos pontos de acurácia.

C. Aplicação do método SVM

Para o método SVM, a classificação foi feita por meio da função de base radial (RBF), que consiste em calcular a distância entre dois vetores no espaço, sendo estes compostos pelas características das células, com o auxílio do parâmetro de regularização definido em 1.0, para que se obtenha uma força de regularização maior dos dados. Com isso, o algoritmo foi treinado e submetido a um teste de previsão, além da aplicação da matriz de confusão e cálculos de precisão, *recall*, *f1-score* e acurácia, assim como na aplicação do método KNN.

D. Aplicação do método Árvore de Decisão

O método de Árvore de Decisão definiu regiões da classe, fazendo com que um ponto médio seja a fronteira entre uma classe e outra. Assim, os atributos se tornam pontos que pertencem a uma certa região, o que consequentemente classifica o dado por meio da região pertencente. Na aplicação deste, foram efetuadas as mesmas formas de treinamento e testes utilizados nos modelos anteriores, oferecendo assim como os outros uma base para a análise a ser vista a seguir.

IV. Resultados e Discussão

Os resultados obtidos após treino, teste e análise dos dados foram de certa forma semelhantes, com diferenças sutis de acurácia e desempenho.

Para todos os métodos abordados, é efetuado individualmente o cálculo de acurácia, matriz de confusão e um relatório de classificação. Implementações estas também fornecidas pela biblioteca *scikit-learn*.

Após aplicação do método KNN, conforme metodologia, é obtida uma acurácia de 0.95 (ou 95%). Na avaliação do desempenho, a precisão manteve uma taxa elevada tanto para classificação como benigno, quanto para o maligno, sendo a figura 1, um exemplo como o algoritmo se desempenhou no diagnóstico do câncer.

Na aplicação da Árvore de Decisão, demonstrou manter o mesmo desempenho que o algoritmo KNN, possuindo a mesma acurácia de 0.95 (ou 95%), além de possuir taxa de desempenho similar ao algoritmo mencionado. O que demonstra tanto na figura 1 quanto na figura 2 que ambos os algoritmos podem oferecer uma precisão bastante aceitável na predição dos dados e consequente diagnóstico.

Já a aplicação do método SVM mostrou ser mais promissora em relação aos outros algoritmos analisados, devido ao alcance de uma acurácia de 0.98 (ou 98%), performando melhor nos parâmetros de validação do algoritmo e sendo bastante preciso no reconhecimento das células malignas, com destaque para as células benignas, onde trouxe uma precisão de 1.0 (ou 100%) na identificação, conforme a figura 3 exemplifica.

Para ressaltar as diferenças entre os algoritmos, foram plotados gráficos para a Curva ROC e a Curva de Precisão-Revocação. Esses gráficos mostram o ponto comum em que cada algoritmo se encontra quando a taxa de FP e a taxa

```
Acurácia de KNN: 0.95
Matriz de Confusão:
VN  FP
FN  VP
  ||
[[68  3]
 [ 3 40]]
VP = Verdadeiro Positivo
VN = Verdadeiro Negativo

FP = Falso Positivo
FN = Falso Negativo

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        71
           1       0.93      0.93      0.93        43

    accuracy                           0.95       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

Figura 1. Desempenho do modelo KNN
Fonte: Próprio autor

```
Acurácia de Árvore de Decisão: 0.95
Matriz de Confusão:
VN  FP
FN  VP
  ||
[[68  3]
 [ 3 40]]
VP = Verdadeiro Positivo
VN = Verdadeiro Negativo

FP = Falso Positivo
FN = Falso Negativo

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        71
           1       0.93      0.93      0.93        43

    accuracy                           0.95       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```

Figura 2. Desempenho do modelo Árvore de Decisão
Fonte: Próprio Autor

de VP são consideradas na Curva ROC. Por outro lado, na Curva de Precisão-Revocação, destaca-se a relevância das instâncias recuperadas, tratadas pela Precisão, e a recuperação de instâncias relevantes, tratadas pela Revocação, convergindo para um ponto em comum que atenda a ambas as métricas. Com isso, o gráfico 4 demonstra ainda mais o quão preciso é o desempenho na classificação de novos dados que cada modelo possui, em especial o SVM. O gráfico 5 também mostra para o método SVM a sua alta capacidade de haver dados relevantes e o quanto destes são recuperados para orientação do algoritmo baseado no modelo em questão.

V. Conclusão

A aplicação de modelos de classificação de dados em um cenário onde o rápido diagnóstico auxilia no andamento dos procedimentos seguintes, acaba se tornando uma ferramenta necessária, quando esta é adequada à importância do rápido procedimento que o tratamento, não somente do câncer de

```
Acurácia de SVM: 0.98
Matriz de Confusão:
VN  FP
FN  VP
  ||
[[71  0]
 [ 2 41]]
VP = Verdadeiro Positivo
VN = Verdadeiro Negativo

FP = Falso Positivo
FN = Falso Negativo

Relatório de Classificação:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        71
           1       1.00      0.95      0.98        43

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114
```

Figura 3. Desempenho do modelo SVM
Fonte: Próprio Autor

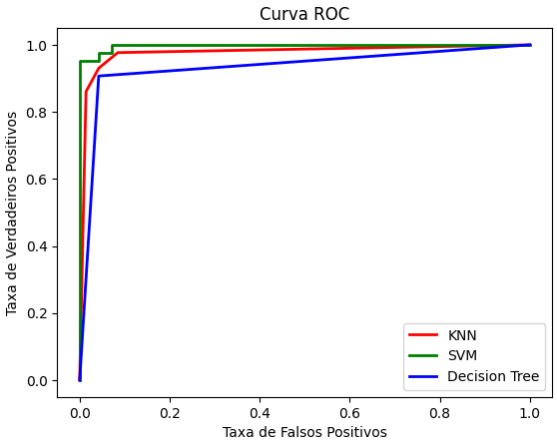

Figura 4. Curva ROC
Fonte: Próprio Autor

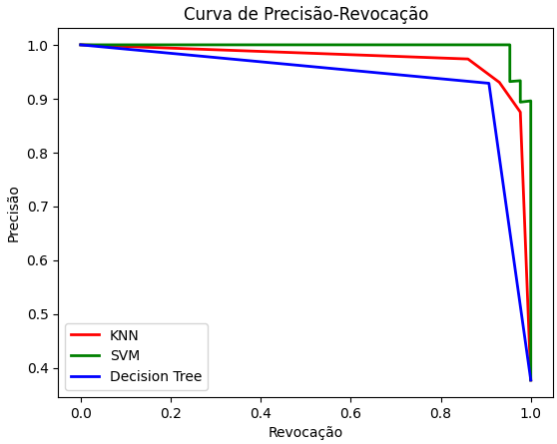

Figura 5. Curva de Precisão-Revocação
Fonte: Próprio Autor

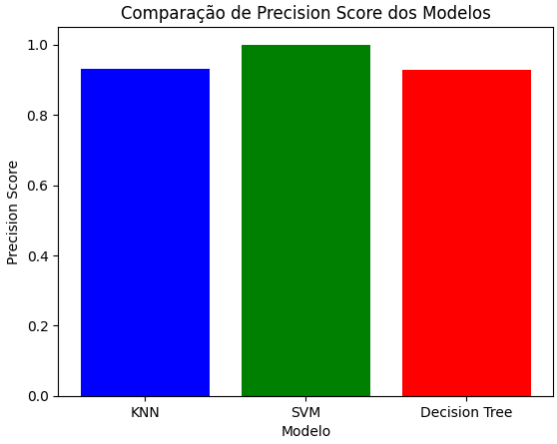

Figura 6. Comparação de Precisão
Fonte: Próprio Autor

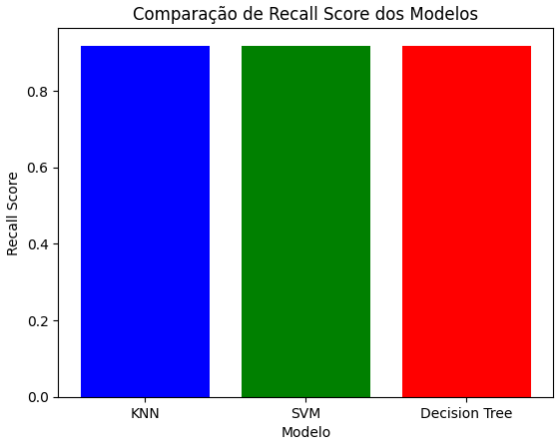

Figura 7. Comparação de Recall
Fonte: Próprio Autor

mama, mas também com o surgimento de possíveis tumores têm a demandar. Em aspectos gerais, os modelos desenvolvidos se desempenharam de forma totalmente satisfatória, o que traz uma visão bastante otimista no uso das IAs em cenários e áreas diferentes do que as funcionalidades tecnológicas atuam comumente. Porém, uma abordagem rica e que seja menos passiva de erros, pode ser mais atrativa, principalmente quando sua aplicação envolve situações na qual são necessárias confirmações mais concretas a respeito dos aspectos celulares para um tratamento mais efetivo de tumores.

Contudo, olhando detalhadamente os aspectos de desempenho de cada um dos métodos, é possível perceber que todos possuem altas taxas de precisão, *recall* e *f1-score*, assim como nos gráficos 6, 7 e 8 respectivamente, sendo estes modelos obtendo desempenho próximos de 1.0 nestes aspectos. Contudo, o método SVM apresentou uma precisão bem mais próxima de 1.0 que os demais, além do *f1-score* poder frisar esta diferença entre os outros modelos, concluindo-se que o modelo

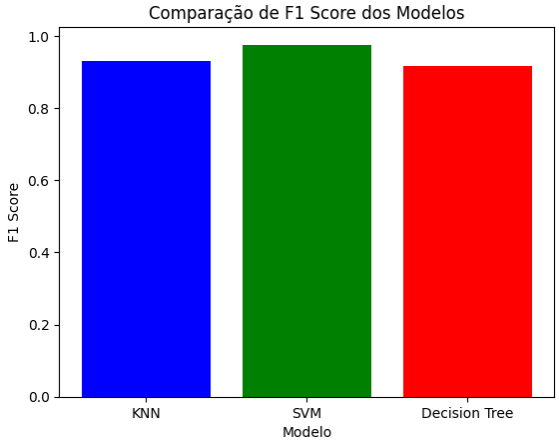

Figura 8. Comparação de F1-score
Fonte: Próprio Autor

em questão performa melhor na detecção do câncer de mama.

Para trabalhos futuros, deseja-se analisar os modelos com a inclusão e tratamento de imagens para tornar mais completa a comparação dos algoritmos, servindo de inspiração para produções de IAs mais robustas na detecção de câncer de mama.

REFERÊNCIAS

[1] Conecte SUS. Câncer de mama. Disponível em: https://conectesus-paciente.saude.gov.br/publico/conteudo/artigo/63fe478e78745c001e8fd397.Acessoem:02deDezembro2023, year=”2023”.

[2] L. M. d. C. L. F. L. M. M. d. O. S. M. d. C. C. Fernanda Cristina da Silva de Lima, Julio Fernando Pinto Oliveira. *Estimativa de Incidência de Câncer no Brasil*. Rev. Bras. Cancerol, 01 2022.

[3] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer. Knn model-based approach in classification. 01 2003.

[4] N. Holsbach, F. S. Fogliatto, and M. J. Anzanello. A data mining method for breast cancer identification based on a selection of variables. *Ciência & Saúde Coletiva*, 19(4):1295, 2014.

[5] Instituto Nacional do Câncer - INCA. O que é câncer. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/o-que-e-cancer.Acessoem: 02deDezembro2023, year=”2023”.

[6] M. Lauretto. Árvores de decisão. 11 2010.

[7] L. A. Nissila. Análise comparativa de métodos computacionais para a classificação de estoque: redes neurais, knn e svm. 2023.

[8] W. Noble. O que é uma máquina de vetores de suporte? 12 2006.