Um Estudo Comparativo entre Algoritmos de Classificação de Dados para Prognóstico de Sobrevida em Pacientes com Diagnóstico de Câncer de Mama

Manuelly Victor[1], Isa Pereira[1], Felipe Silva[1], Erick Trindade[1], Flavius Gorgônio[1], Karliane Vale[1], Yasmin Rebeca[2], Maria de Lourdes Morais[2]

[1]Laboratório de Inteligência Computacional Aplicada a Negócios
Departamento de Computação e Tecnologia (DCT)
Universidade Federal do Rio Grande do Norte (UFRN) – Caicó/RN – Brasil
[2] Departamento de Odontologia (DOD)
Universidade Estadual do Rio Grande do Norte (UERN) – Caicó/RN – Brasil

{manuelly.rodrigues.705,isa.soares.709}@ufrn.edu.br
{felipe.benicio.701,erick.trindade.071}@ufrn.edu.br
{flavius.gorgonio,karliane.vale}@ufrn.br
yasminnascimento@alu.uern.br, mariaarruda@uern.br

Abstract. *The use of classification algorithms has been an important ally in activities that require prognostics, including in the healthcare field. An example of this is the use of these algorithms in analyzing the survival time of patients diagnosed with cancer. In this work, a comparative analysis is conducted between the Naive Bayes, J48, and KNN algorithms to verify the effectiveness of the treatment concerning the patient's survival time. To carry out this study, a database of patients diagnosed with breast cancer was used, containing behavioural data and the respective treatments to which the patients were subjected. Preliminary results indicate that the J48 algorithm outperformed the other tested algorithms in all analyzed metrics (accuracy, precision, recall, and F1-Score), demonstrating its potential to assist in research involving the planning of treatments for patients based on survival time.*

Resumo. *A utilização de algoritmos de classificação tem sido um importante aliado em atividades que demandam prognósticos, inclusive na área de saúde. Um exemplo disso é a utilização destes algoritmos na análise do tempo de sobrevida em pacientes com diagnóstico de câncer. Neste trabalho, é realizada uma análise comparativa entre os algoritmos Naive Bayes, J48 e KNN para verificar a eficácia do tratamento utilizado em relação ao tempo de sobrevida do paciente. Para realizar esse estudo, foi usada uma base de dados de pacientes diagnosticados com câncer de mama, contendo dados comportamentais e os respectivos tratamentos aos quais os pacientes foram submetidos. Resultados preliminares indicam que o algoritmo J48 obteve desempenho superior aos demais algoritmos testados em todas as métricas analisadas (acurácia, precisão, recall e F1-Score), demonstrando seu potencial para auxiliar pesquisas que envolvam o planejamento de tratamentos de pacientes com base no tempo de sobrevida.*

1. Introdução

Nos últimos anos, a utilização de algoritmos de mineração de dados tornou-se uma ferramenta importante em diversos estudos direcionados à saúde, proporcionando a capacidade de lidar com cenários que implicam na manipulação e análise de grandes volumes de dados. De acordo com a Organização Mundial de Saúde (OMS), o câncer é considerado a segunda principal causa de mortalidade resultante de enfermidades no mundo, destacando o câncer de mama como o tipo mais frequente entre a população feminina [Silva 2022].

Considerando que os registros hospitalares digitais abrigam dados, qualitativos e quantitativos, aptos a servir como fonte de informação para estudos, a investigação por determinantes associados à estimativa do tempo de sobrevida para pacientes diagnosticados com câncer de mama tem se destacado como um campo de estudo. Para tanto, são utilizados algoritmos de aprendizado de máquina (AM) que possibilitam identificar padrões ocultos nos dados para a predição, que é realizada a partir de casos históricos armazenados em conjuntos de dados [Santos et al. 2023].

Diante deste cenário, o presente trabalho realiza um estudo comparativo entre três algoritmos de AM usados na classificação de dados, a saber: *k-nearest neighbors* (KNN), árvore de decisão (J48) e *Naive Bayes*. Com isso, pretende-se identificar a eficácia do tratamento em relação à sobrevida do paciente. A pesquisa foi desenvolvida a partir de uma base de dados contendo 616 instâncias e 29 atributos, incluindo alguns referentes ao sistema TNM que descreve características anatômicas da doença, considerando o tumor primário (T), os linfonodos (N) e a presença ou ausência de metástases a distância (M).

2. Referencial Teórico

Com a crescente complexidade dos problemas a serem tratados computacionalmente e do volume de dados gerados por diferentes setores, tornou-se clara a necessidade de ferramentas computacionais sofisticadas que fossem mais autônomas, reduzindo a necessidade de intervenção humana e dependência de especialistas. Neste contexto, o aprendizado de máquina emergiu como uma área de notável importância, caracterizada por sua habilidade de criar por si própria, a partir de experiências passadas, uma hipótese ou função capaz de resolver o problema que se deseja tratar [Faceli 2011].

Algoritmos de AM tem sido utilizados com sucesso em problemas do mundo real, incluindo aqueles relacionados com a área de saúde. Em [Santos et al. 2023], foi apresentado um estudo comparativo entre algoritmos de AM que realizam a predição da sobrevida do pacientes com Câncer de Mama. Neste trabalho, os algoritmos *Naïve Bayes*, *Random Forest*, *Multilayer Perceptron* e *AdaBoost* foram aplicados a uma base de dados pública contendo 1570 pacientes. Os resultados mostram que o algoritmo *Random Forest* apresentou maior acurácia e especificidade, porém o *AdaBoost* foi mais sensível.

Com o propósito de avaliar o potencial da mineração de dados na classificação dos tumores de mama, [Wickeinecki and Viegas 2020] realizou investigações utilizando o algoritmo KNN. Com isso, observou-se uma possível substituição de algumas técnicas tradicionais por AM, com o intuito de otimizar o tempo de análise, reduzir custos e elevar a precisão dos diagnósticos. Nesta pesquisa comparativa, contemplou-se que o algoritmo trabalhado obteve uma média de acerto de 75% no diagnóstico dos tumores.

Em [Fonseca et al. 2021], foi realizado um estudo, a partir do uso dos algorit-

mos de mineração de dados, para a detecção de problemas relacionados à prescrição e administração de medicamentos. Como resultado da pesquisa, foram identificados equívocos, por exemplo, na prescrição de protetores na UTI Neonatal, vitaminas e repositores.

3. Metodologia

Esta pesquisa analisa uma base de dados disponibilizada pela Universidade Estadual do Rio Grande do Norte (UERN) incluindo registros médicos de 616 pacientes. Tal base de dados possui informações fenotípicas, do diagnóstico e do tratamento de cada paciente, além de outros parâmetros. Ressalta-se que esta pesquisa foi submetida e aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos (CEP) da UERN, com o parecer de aprovação nº 2.445.404.

A Figura 1 apresenta a metodologia deste trabalho que inicia com a limpeza dos dados e segue para a etapa do pré-processamento, em que foram geradas duas cópias da base de dados. Na primeira, foi identificado o número de vezes que o paciente realizou o tratamento (processo definido como *quantificação*). Na segunda, foi realizada a *binarização*, que consiste na confirmação ou negação da ocorrência do tratamento, isto é, o 1 representava a realização do tratamento e o 0 a ausência dele. O terceiro passo da metodologia foi a aplicação dos algoritmos de classificação *Naive Bayes*, J48 e KNN, usando-se subconjuntos de treinamento e teste. Por fim, foi realizada a análise comparativa dos resultados obtidos por cada algoritmo.

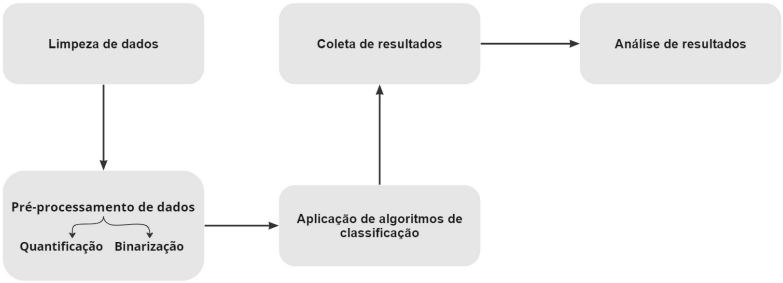

Figura 1. Descrição das etapas da metodologia

4. Resultados Preliminares

A Tabela 1 representa os resultados obtidos a partir da aplicação dos algoritmos sobre as duas bases de dados (quantificação e binarização). A referida tabela está organizada da seguinte forma: (1) a primeira coluna apresenta os algoritmos utilizados nos experimentos e (2) as demais colunas exibem, em valores percentuais, os resultados das métricas consideradas no estudo: i) acurácia, ii) precisão, iii) *recall* e iv) *F1-score*. É importante destacar que: i) a acurácia representa a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões; ii) a precisão mede a quantidade de vezes que o modelo acerta em relação ao total de vezes que ele tenta acertar; iii) a *Recall* mede a quantidade de vezes que o seu modelo acerta em relação ao total de vezes que ele deveria ter acertado; e iv) a *F1-Score* combina precisão e recall de maneira equilibrada.

Os valores destacados em negrito são os melhores resultados para cada métrica. Analisando a referida tabela observa-se que o J48 alcançou melhor desempenho em todas

as métricas quando comparado aos demais algoritmos. Explorando cada base de dados separadamente, quantificação e binarização, percebe-se que, em ambas, os melhores resultados também foram alcançados pelo J48 em 100% dos casos (4 das 4 métricas). Avaliando comparativamente as bases de dados, verifica-se que a base de dados quantificada (quantificação) obteve melhores resultados na maioria dos casos (3 das 4 métricas).

Tabela 1. Resultados das métricas: Acurácia, Precisão, *Recall* e *F1-Score*

Algoritmos	**Acurácia**	**Precisão**	***Recall***	***F1-Score***
Quantificação				
KNN	76.46%	85.07%	84.17%	84.62%
J48	**89.61**%	97.01%	**90.09**%	**93.42**%
Naive Bayes	78.89%	81.87%	89.51%	85.52%
Binarização				
KNN	75.32%	84.43%	83.36%	83.89%
J48	89.28%	**97.22**%	89.58%	93.25%
Naive Bayes	78.08%	81.02%	89.20%	84.91%

5. Considerações Finais

Este trabalho apresentou uma proposta de análise comparativa de algoritmos de AM para classificação de dados sobre uma base de dados clínicos, com o intuito de realizar prognósticos sobre o tempo de sobrevida em pacientes diagnosticados com câncer de mama. Para tanto, foram geradas duas cópias diferentes da base de dados (quantificação e binarização) e sobre as quais foram aplicados os algoritmos KNN, J48 e *Naive Bayes*, analisando-se os resultados a partir das métricas: acurácia, precisão, *recall* e *F1-score*.

Como resultados preliminares, foi possível concluir que o algoritmo J48 se destacou como o melhor para o prognóstico do tempo de sobrevida dos pacientes diagnosticados com câncer de mama. Esta análise é reforçada pelo fato do algoritmo ter obtido o maior desempenho nas duas cópias da base de dados, em todas as métricas. Trabalhos futuros poderão averiguar a performance do J48 em termos estatísticos e compará-lo com outros algoritmos de classificação. Além disso, para o enriquecimento científico da pesquisa, podem ser investigadas a influência da ordem de realização de cada tratamento e de fatores sócio-econômicos relacionados ao tempo de sobrevida desses pacientes.

Referências

Faceli, K. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC.

Fonseca, A. S. M. et al. (2021). Mineração de dados de problemas relacionados a medicamentos registrados pela farmácia clínica de um hospital universitário. *Arquivos Catarinenses de Medicina*, 50(2):142–155.

Santos, P. D. d., Yahata, E., Piheiro, T. S., Oliveira, F. S. d., and Simões, P. W. (2023). Algoritmos de machine learning para predição da sobrevida do câncer de mama. *Journal of Health Informatics*, 15(Especial):1–12.

Silva, M. E. A. (2022). Proposta e avaliação de um modelo híbrido de seleção de características para o prognóstico do câncer de mama. Master's thesis, Universidade Federal de Alagoas, Maceió, AL.

Wickeinecki, A. S. Z. and Viegas, S. C. (2020). A tecnologia e saúde uma simbiose. *ReFAQI-Revista de Gestão Educação e Tecnologia*, 8(2):1–18.