Utilização de Técnicas de Mineração de Dados para Identificação de *Personas* em Pacientes com Câncer de Mama

Felipe Silva[1], Isa Pereira[1], Manuelly Victor[1], Erick Trindade[1], Karliane Vale[1], Flavius Gorgônio[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

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

Abstract. *The use of personas is a strategy adopted in various knowledge domains for identifying archetypes that represent groups of individuals based on data, characteristics, or common behavioral patterns within the group. In this work, this strategy is employed by applying clustering and summarizing techniques to a database of breast cancer patients, which includes behavioral data and the respective treatments the patients underwent. Preliminary results indicate potential for identifying patterns and discovering useful information and insights that relate personas to the adopted treatments and the survival time of the patients.*

Resumo. *A utilização de personas é uma estratégia adotada em vários domínios do conhecimento para identificação de arquétipos que representam grupos de indivíduos a partir de dados, características ou padrões de comportamento comuns ao grupo. Neste trabalho, esta estratégia é utilizada aplicando-se técnicas de análise de agrupamentos e sumarização sobre uma base de dados de pacientes diagnosticados com câncer de mama, contendo dados comportamentais e os respectivos tratamentos aos quais os pacientes foram submetidos. Os resultados preliminares indicam potencial para a identificação de padrões e a descoberta de informações úteis e insights que relacionam as personas, os tratamentos adotados e o tempo de sobrevida dos pacientes.*

1. Introdução

Personas são arquétipos, ou seja, padrões de comportamento associados a um personagem, que permitem representar diversos perfis que formam o público estratégico de uma iniciativa, organização, produto ou serviço [Martins and Vanz 2021]. O conceito de *persona* foi desenvolvido por Alan Cooper, na área de design, para modelar e simular usuários no desenvolvimento de projetos de tecnologia [Nielsen et al. 2022], mas vem sendo amplamente utilizado em diversas áreas, inclusive em pesquisas de informática na saúde [Holden et al. 2017, Zhu et al. 2019, Alsaadi and Alahmadi 2021, Haupt et al. 2022].

Nos últimos anos, o crescimento e evolução de pesquisas científicas envolvendo grandes volumes de dados na área da saúde vem tornando necessária a aplicação de tecnologias que auxiliem na análise e extração de informações em bases de dados fornecidas para esses estudos, normalmente compostas por centenas e até milhares de registros de pacientes. Dessa forma, a mineração de dados pode fornecer informações valiosas para diagnóstico e prognóstico por meio da extração de informações úteis, de maneira rápida e eficaz, de grandes quantidades de registros médicos [Tandon et al. 2020].

Considerando-se as dificuldades inerentes ao tratamento individual de grandes volumes de dados, associadas à ideia de que em um grupo de pessoas há, obviamente, indivíduos com características comuns, a criação de *personas* surge como uma técnica para representar um segmento de usuários (no caso, de pacientes) como sendo uma única pessoa que agregue essas características comuns [Jansen et al. 2021]. Esta pessoa fictícia incorpora e representa todos (ou a maioria) dos pacientes em um segmento. Este conceito de *persona* é empregado aqui para se referir à técnica de representação de um conjunto de pacientes a partir dos seus dados em comum (*data-driven persona*).

Nessa perspectiva, o presente artigo apresenta uma proposta de utilização de *personas* para agrupar e descrever pacientes com perfis semelhantes, a partir da aplicação do algoritmo *k-means* sobre uma base de dados de pacientes com câncer de mama, seguida por uma etapa de sumarização, que consiste na identificação e descrição estatística dos dados de cada um dos grupos identificados. O objetivo é identificar relações entre os atributos presentes em cada *persona*, que permitam auxiliar em processos de tomada de decisão, por exemplo, sobre qual tratamento seguir de acordo com as características de cada paciente ou quais características comuns estão associadas à sobrevida do paciente.

2. Referencial Teórico

Persona é uma técnica de modelagem de usuário utilizada em várias áreas do conhecimento para fornecer *insights* sobre o público e comunicar suas necessidades e comportamentos. Em [Alsaadi and Alahmadi 2021], este conceito é utilizado para modelar *personas* baseada na análise de dados quantitativos e em técnicas de aprendizado de máquina, demonstrando que esta estratégia é superior às tradicionais (que usam dados qualitativos) por permitir lidar com grandes volumes de dados a partir do uso de técnicas analíticas.

[Haupt et al. 2022] propõem uma abordagem semelhante, baseada em análise de agrupamentos e análise fatorial, para modelar comportamentos e atitudes classificadas como arriscadas em 482 indivíduos americanos em relação a hábitos comportamentais ligados à pandemia de COVID-19. A pesquisa identifica perfis de indivíduos a partir de fatores psicológicos e situacionais, utilizando uma estrutura de *persona* baseada em dados relacionados à área da investigação.

Dois outros trabalhos relacionados abordam a utilização de *personas* a partir da análise de dados de saúde de pacientes idosos dos EUA. Em ambos os casos, os autores utilizam algoritmos de análise de agrupamentos sobre dados quantitativos para criação de *data-driven personas*. No primeiro, [Holden et al. 2017] utilizam uma base de dados contendo registros médicos de 32 idosos hospitalizados com insuficiência cardíaca. Os resultados permitiram identificar 6 *personas* que, apesar de possuírem características semelhantes entre si, algumas características diferentes entre os indivíduos demonstraram potencial importância para a concepção de serviços de saúde. Já em [Zhu et al. 2019],

foi utilizado o algoritmo *k-means* sobre uma base de dados contendo 170.704 pacientes idosos dos EUA, aleatoriamente selecionados de uma pesquisa nacional. A análise de agrupamentos foi seguida por uma etapa de análise qualitativa e, ao final, o estudo apresenta uma metodologia de construção de *personas* para representação de idosos.

3. Metodologia

Neste trabalho, foram utilizadas as técnicas de análise de agrupamentos (algoritmo *k-means*) e sumarização para identificação de *personas* em pacientes com diagnóstico de câncer de mama. Foi utilizada uma base de dados coletada pela Universidade Estadual do Rio Grande do Norte (UERN) contendo registros médicos de 616 pacientes, incluindo informações fenotípicas, do diagnóstico (sistema TNM para classificação de tumores malignos) e do tratamento (procedimentos e fármacos utilizados) de cada paciente. Esta base de dados é originária de uma pesquisa aprovada pelo Comitê de Ética em Pesquisa com Seres Humanos (CEP) da UERN, com o parecer de aprovação nº. 2.445.404. A estratégia utilizada na metodologia é sintetizada na Figura 1.

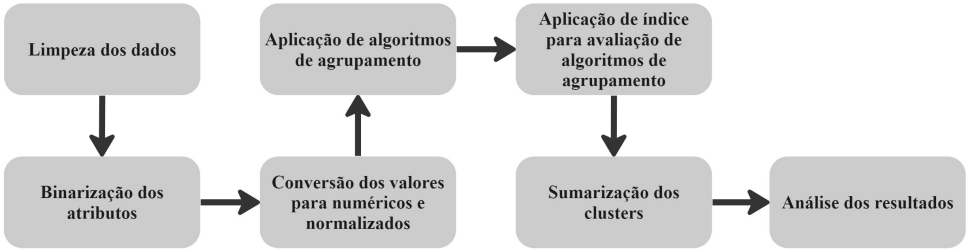

Figura 1. Descrição das etapas da metodologia

4. Resultados Preliminares

Após a execução dos algoritmos de agrupamento e sumarização, foram identificados os seguintes resultados preliminares apresentados na Tabela 1. Cada agrupamento é simbolizado por uma *persona*, que representa os padrões mais comumente identificados no respectivo *cluster*. Os números entre parênteses representam a quantidade de indivíduos com aquela característica mais frequentemente encontrada (moda).

Tabela 1. Resultados estatísticos dos atributos de cada cluster

Característica avaliada	Cluster 1 (469 instâncias)	Cluster 2 (87 instâncias)	Cluster 3 (60 instâncias)
Idade média	53 anos	43 anos	68 anos
Anos de sobrevida	Todos vivos	1.7 anos	2.3 anos
Tratamento mais utilizado	Cirurgia (348)	Radioterapia (62)	Radioterapia (31)
Tamanho do tumor	T9 (176)	T4 (41)	T4 (24)
Característica dos linfonodos	N9 (176)	N1 (43)	N1 (20)
Nível de metástase	M9 (176)	M1 (36)	M1 (26)
Fármaco arredia	380	84	58
Fármaco zometa	84	3	2
Fármacos arredia e zometa	5	0	0
Lateralidade afetada pelo tumor	Esquerda (244)	Direita (50)	Esquerda (31)
Tabagismo	Não (222)	Não (44)	Sim (24)
Alcoolismo	Não (260)	Não (51)	Sim (30)
Características fenotípicas	Parda (268)	Parda (55)	Parda (34)
Nível de escolaridade	Fund incompleto (119)	Nível médio (20)	Fund incompleto (18)
Histórico familiar	Sim (207)	Sim (41)	Sim (26)
Estado da doença	Doença sem evidência (227)	Progressão da doença (55)	Progressão da doença (30)
Estadiamento	3A (96)	4 (44)	4 (30)
Tipo histológico	Carcinoma ductal infiltrante (433)	Carcinoma ductal infiltrante (81)	Carcinoma ductal infiltrante (55)

Ademais, observa-se que o estadiamento, uma das características avaliadas, informa o grau da doença, em especial o estágio e a extensão em que o câncer se encontra.

Logo, quanto mais avançado, maior o número retornado. Além desse atributo, o estado da doença apresenta o grau que se encontra a doença após a realização do tratamento, portanto, os resultados mais recorrentes demonstram que no *cluster* 1 os tratamentos tiveram resultados positivos nos pacientes analisados. Por fim, é importante salientar que o histórico familiar representa a ocorrência do tipo histológico no leito familiar e, mediante a ele, é possível notar a influência desse atributo na manifestação da doença.

5. Considerações Finais

A construção de *personas* baseadas em dados, a partir da utilização de técnicas de aprendizado de máquina, é uma ferramenta importante para profissionais de informática em saúde que necessitam compreender grandes volumes de dados. O artigo apresentou uma proposta de análise de uma base de dados de pacientes com câncer de mama a partir do uso de técnicas de análise de agrupamento e sumarização para identificação de *personas*, que representam perfis existentes dentro da base de dados analisada.

Apesar dos resultados ainda estarem em uma etapa inicial, já foi possível identificar os perfis encontrados com maior frequência em cada *cluster* analisado e verificar relações existentes entre os mesmos e as variáveis existentes na base. Trabalhos futuros poderão investigar relações entre os fatores sócio-ambientais, assim como, a influência da ordem de realização de cada tratamento no tempo de sobrevida desses pacientes, abrindo espaço para novas análises. A utilização de outros algoritmos de agrupamento também pode ser investigada, visto que o *k-means* é bastante sensível a *outliers*.

Referências

Alsaadi, B. and Alahmadi, D. (2021). The use of persona towards human-centered design in health field: Review of types and technologies. In *2021 International Conference on e-Health and Bioengineering (EHB)*, pages 1–4.

Haupt, M. R., Weiss, S. M., Chiu, M., Cuomo, R., Chein, J. M., and Mackey, T. (2022). Psychological and situational profiles of social distance compliance during covid-19. *Journal of Communication in Healthcare*, 15(1):44–53.

Holden, R. J., Kulanthaivel, A., Purkayastha, S., Goggins, K. M., and Kripalani, S. (2017). Know thy ehealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *Int. Journal of Medical Informatics*, 108:158–167.

Jansen, B. J., Salminen, J., Jung, S.-g., and Guan, K. (2021). *Creating Data-Driven Personas*, pages 93–118. Springer International Publishing, Cham.

Martins, M. R. and Vanz, S. A. d. S. (2021). *Construção de personas: mapeamento de estudos e métodos*, pages 225–238. Pimenta Cultural, São Paulo, SP.

Nielsen, L., Jansen, B. J., Salminen, J., Abdelnour Nocera, J., and Jung, S.-G. (2022). Personas: New data, new trends. In *Extended Abstracts of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. ACM.

Tandon, D., Rajawat, J., and Banerjee, M. (2020). Present and future of artificial intelligence in dentistry. *Journal of oral biology and craniofacial research*, 10(4):391–396.

Zhu, H., Wang, H., and Carroll, J. M. (2019). Creating persona skeletons from imbalanced datasets - a case study using u.s. older adults' health data. In *Proc. of the 2019 on Designing Interactive Systems Conference*, page 61–70, New York, NY, USA. ACM.