Clusterização de Dados Abertos em Oncologia Usando Técnicas de Aprendizado de Máquina: um estudo preliminar sobre recidiva de câncer de mama

Pedro Ferreira Crocco[1,2], Leandro de Souza Junior[1,2], Mariangela Dametto[1], Rodrigo Bonacin[1]

l217739@dac.unicamp.br, pcrocco@cti.gov.br, mdametto@cti.gov.br, rodrigo.bonacin@cti.gov.br

[1] Divisão de Metodologias da Computação - DIMEC, Centro de Tecnologia da Informação Renato Archer – CTI

[2]Univesidade Estadual de Campinas - Unicamp

Abstract. *Breast cancer is the most common among women, so it is of general interest for medicine to study relapse cases. Among the available machine learning techniques, the main focus of this scientific initiation is on the use of clustering algorithms to generate groups of patients in an open database. Our goal is to associate the relapse feature with the other features present in the Fundação Oncocentro de São Paulo database. It is expected to contribute to the study of the effectiveness of clustering techniques in open clinical data on breast cancer. Preliminary results are promising and point to the possibility of continuing studies with more advanced techniques.*

Resumo. *O câncer de mama é o mais comum dentre as mulheres, sendo assim de interesse geral para a medicina estudar os casos de recidiva. Dentre as técnicas de aprendizado de máquina disponíveis o foco principal desta iniciação científica está no uso de algoritmos de clusterização para gerar grupos de pacientes em uma base de dados aberta. Busca-se principalmente associar a feature de recidiva com as demais features presentes na base de dados da Fundação Oncocentro de São Paulo. Espera-se contribuir para o estudo da eficácia de técnicas de clusterização em dados clínicos abertos sobre câncer de mama. Os resultados preliminares são promissores e apontam para possibilidade de continuidade dos estudos com técnicas mais avançadas.*

1. Introdução

De acordo com o Ministério da Saúde do Brasil, o câncer de mama é o mais comum dentre as mulheres, representando cerca de 28% dos novos casos [Ministério da Saúde 2023]. Como consequência, é de interesse geral para a medicina estudar os casos de recidiva de um câncer. A recidiva é definida pela Sociedade Americana de Câncer como a reaparição do tumor após um período em que este não podia ser detectado [American Cancer Society 2016].

Para estudo aprofundado das recidivas de câncer de mama, este artigo foca em utilizar técnicas de aprendizado de máquina em uma base de dados aberta, com dados coletados no Brasil e disponibilizados publicamente pela Fundação Oncocentro de São Paulo (FOSP)[1].

Dentre as ferramentas de aprendizado de máquina disponíveis o foco principal desta iniciação científica são os algoritmos de clusterização. Tais algoritmos foram implementados utilizando biblioteca de código aberto em Python SciKit Learn[2], que contém uma variedade de classes e objetos para a aplicação de diversos algoritmos no contexto da ciência de dados. Para este estudo em específico, foram aplicados 3 algoritmos de clusterização diferentes na mesma base com os dados. Esses dados foram preprocessados para cada caso, sem perda de informação, com o intuito de verificar a presença de grupos específicos que não poderiam ser observados sem o uso dos algoritmos. Busca-se principalmente associar a *feature* de recidiva com as demais *features* presentes na base de dados.

Com este artigo espera-se contribuir para o estudo da eficácia de técnicas de clusterização em dados clínicos abertos sobre câncer de mama. Visa-se em longo prazo apoiar o direcionamento de estudos para certos grupos de pacientes com câncer de mama, bem como apoiar a identificação e diferenciação destes grupos claramente.

O restante deste artigo está organizado da seguinte maneira: a seção 2 apresenta os conceitos adotados neste artigo, a base de dados explorada e os algoritmos de clusterização, a seção 3 apresenta a metodologia empregada, a construção do protótipo e execução dos algoritmos, a seção 4 apresenta os resultados obtidos, por fim, a seção 5 faz as considerações finais, conclui o artigo e apresenta os trabalhos futuros.

2. Conceitos, Base de Dados e Algoritmos de Clusterização

Esta seção apresenta primeiramente a base de dados da FOSP e conceitos ligados ao câncer de mama utilizados neste trabalho na subseção 2.1, bem como os algoritmos K-Prototypes, K-Means e *Density-Based Spatial clustering of Applications with Noise* (DBSCAN) na subseção 2.2.

2.1. Câncer de Mama e a base de dados

O câncer de mama, identificado pelo CID-O 50X (Figura 1), é um tumor maligno originado do crescimento anormal e descontrolado de células da região. Sua origem pode ser notada nas glândulas, canais, mamilos, tecidos de gordura e vasos sanguíneos/linfáticos [American Cancer Society 2021].

A base de dados tomada como objeto para o estudo foi a base de câncer da FOSP, contando com mais de 1.000.000 de pacientes, tendo cada um cerca de 100 *features* disponíveis para análise. As *features* incluem dados regionais, pessoais (Idade, Sexo) e dados a respeito do tumor (Tratamento realizado, recidiva, classificação TNM,

[1]https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc/
[2] https://scikit-learn.org/stable/index.html

metástases). Conforme ilustra a Figura 1, na base de dados da FOSP, as regiões afetadas pelo câncer são descritas pela *feature* Morfologia, que permite um estudo mais segmentado. Tais *features* são cruciais na análise dos clusters e devem ser levados em consideração na análise de resultados.

C50	Mama
C500	Mama, mamilo
C501	Mama, porção central da
C502	Mama, quadrante superior interno da
C503	Mama, quadrante inferior interno da
C504	Mama, quadrante superior externo da
C505	Mama, quadrante inferior externo da
C506	Mama, porção axilar da
C508	Mama, lesão sobreposta da
C509	Mama, SOE (exclui pele da mama c44.5)

Figura 1. CID-O para o câncer de mama (Topologia) fonte; https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/cid-o/

2.3. Algoritmos de Clusterização

Esta seção apresenta os principais algoritmos utilizados nesta iniciação científica, ou seja, o K-Prototypes (subseção 2.3.1), K-Means (subseção 2.3.2) e DBSCAN (subseção 2.3.3).

2.3.1 K-Prototypes

Tendo em vista o grande volume de dados e a distribuição entre dados categóricos e numéricos, o primeiro algoritmo escolhido foi o K-Prototypes [Huang 1998], uma união dos algoritmos K-means com o K-Modes, criada especificamente para o caso em que o tipo de dados das *features* não é homogêneo.

O K-Prototypes utiliza como métrica de distância entre dois pontos distintos a soma da distância euclidiana para *features* numéricas, junto com o coeficiente de dissimilaridade para *features* categóricas. Tudo isso ponderado por um coeficiente γ, evitando que certos tipos de *features* influenciem mais que outras, conforme ilustra a Equação 1.

$$d_2(X, Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \quad (1)$$

Em que o primeiro termo da equação se refere à distância euclidiana e o segundo termo à dissimilaridade discutida em [Kaufman e Rousseeuw 1990] apresentada na Equação 2.

$$d_1(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (2)$$

Uma das limitações do algoritmo é que, para que funcione, o número *K* de clusters precisa ser escolhido previamente. Tal escolha será discutida na seção 3 deste artigo. O algoritmo, no geral, pode ser dividido em 4 passos distintos:

1) *K* clusters aleatórios são selecionados da base de dados.
2) A distância de cada ponto para os *K* centroides é computada e cada ponto é associado ao cluster com menor distância do centroide.
3) A média/moda é calculada para cada cluster e os centroides são redefinidos.
4) O passo 2 e 3 são repetidos até que o deslocamento entre centroides seja mínimo/nulo (convergência do algoritmo).

2.3.2 K-Means

Conforme apresentado por Sinaga e Yang (2020), o algoritmo K-means baseia-se na distância euclidiana para computar dissimilaridades entre pontos, representada pela Equação 3.

$$: \sum_{j=1}^{p} (x_j - y_j)^2 \quad (3)$$

Ou seja, os dados precisam essencialmente ser numéricos, o que significa que a distância entre pontos precisa ter um significado físico/prático, o que torna difícil a aplicação quando a base de dados possui uma diversidade considerável de *features* categóricas.

Ao escolher *K* centroides, o algoritmo seleciona *K* pontos aleatórios da base de dados e as 3 etapas são seguidas, em ordem:

1) A distância euclidiana de cada ponto é computada e o cluster respectivo é associado.
2) É calculada a média de cada cluster e os centroides são atualizados. (O novo centroide não necessariamente é um ponto existente no cluster, e sim um derivado das médias).

3) O passo 1 e 2 são repetidos até que o algoritmo apresente a convergência. Isto é, não haja mais movimentações entre clusters.

O algoritmo tem como principal fator limitante a necessidade de escolha dos *K clusters*. Porém, sua vantagem é a aplicação para bases de dados com alto volume, visto que sua complexidade de tempo computacional pode ser denotada por O(knt) [Xu e Tian 2015].

2.3.3 DBSCAN

Diferentemente dos demais algoritmos apresentados, o DBSCAN é um algoritmo de clusterização baseado na densidade de pontos da amostra. Essencialmente, um ponto p é um ponto central se uma quantidade mínima de pontos está a uma distância mínima ε deste ponto. Um ponto q é diretamente ligado a p, se existe uma distância ε que une ambos. Portanto, o conceito do algoritmo é encontrar clusters na amostra de acordo com o número de pontos principais existentes, baseando-se na densidade de pontos ao redor destes. Como hiper parâmetros, o modelo recebe o número mínimo de pontos para formar pontos principais e uma distância mínima ε. Schubert e outros (2017) apresenta detalhes sobre o funcionamento do algoritmo DBSCAN.

3. Metodologia, Protótipo e Execução

O primeiro passo foi o pré-processamento da base de dados. Para tanto, foi necessário filtrar todos os pacientes que apresentavam câncer de mama, analisar as *features* impactantes e assim aplicar os algoritmos K-Prototypes, K-Means e DBSCAN e, por fim, comparar os resultados. Esta comparação é necessária para verificar se realmente existem clusters que podem ser identificados pelos 3 algoritmos, avaliando também a sua performance para verificar qual método é o mais eficaz, interpretando logo após os resultados individuais.

3.1 Escolha de *features* de impacto e tratamento de dados

Ao analisar a base de dados, com apoio de uma especialista do domínio, as seguintes *features* foram escolhidas após serem julgadas como mais impactantes:

- **IDADE:** Idade do paciente, em anos.
- **CONSDIAG:** Diferença de dias entre a primeira consulta e o diagnóstico da doença.
- **TRATCONS:** Diferença de dias entre a consulta e o tratamento.
- **DIAGTRAT:** Diferença de dias entre o diagnóstico e o tratamento.
- **ESCOLARI:** Escolaridade do paciente.
- **UFRESID:** Unidade Federativa de residência do paciente.
- **CATEATEND:** Categoria em que foi atendido.
- **CLINICA:** Tipo de clínica do diagnóstico.
- **DIAGPREV:** Diagnóstico anterior.
- **EC:** Estadiamento clínico.
- **META01:** Local da metástase.
- **BASEDIAG:** Base do diagnóstico.

- **MORFO:** Morfologia do tumor.
- **T:** Tamanho do tumor primário, de acordo com a classificação TNM. [https://www.cancer.gov/about-cancer/diagnosis-staging/staging]
- **N:** Número de linfonodos com câncer próximos ao tumor, de acordo com a classificação TNM.
- **M:** Se o câncer sofreu metástase, de acordo com a classificação TNM.
- **TRATHOSP, TRATFANTES, TRATFAPOS:** Combinação de tratamentos realizados para tratamento da doença.
- **RECNENHUM, RECLOCAL, RECREGIO, RECDIST:** Presença ou não de recidivas locais, regionais ou à distância.

Dentre todos os pacientes, foram retirados da base todos aqueles em que o campo **ULTINFO** indicava óbito e o campo **RECNENHUM** indicava a ausência de recidiva simultaneamente, afinal o objetivo principal do estudo é buscar relações entre a recidiva e as demais *features*. A inclusão dessa *feature* geraria resultados enviesados pela presença de óbitos pelo primeiro câncer, sem chance de manifestação de uma recidiva do tumor. O número total de pacientes n análise restante foi de *110.553*.

Em seguida, todas as *features* cujo valor era nulo (NaN) foram substituídos por *0,* pois são majoritariamente metástases não encontradas, o que justifica a substituição. Também, a *feature* RECNENHUM foi substituída por REC (negação de RECNENHUM) devido à falta de clareza e interpretação enganosa que poderia gerar.

Para comparação, foi computado o total de recidivas apresentadas perante o total de pacientes analisados, resultando no histograma apresentado na Figura 2.

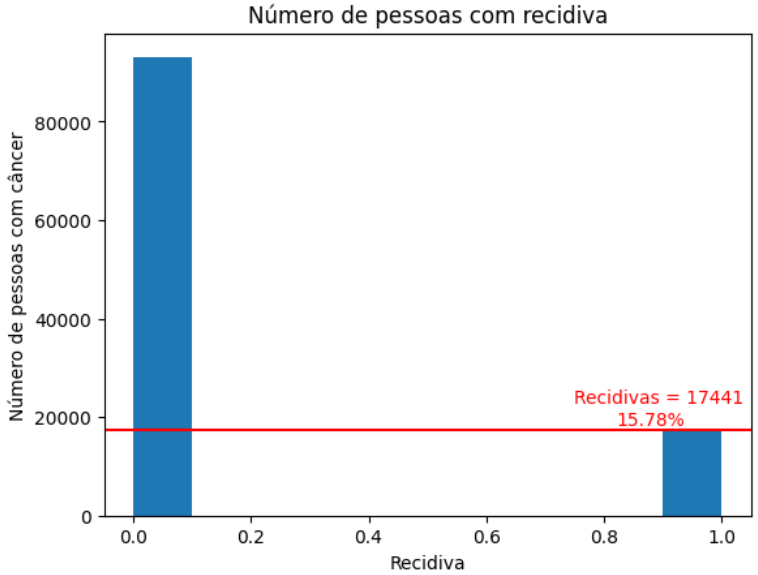

Figura 2. Histograma comparando valores de recidivas e não-recidivas

O histograma apresentado na Figura 3 apresenta o quantitativo de pacientes de acordo com a idade. É possível visualizar a maior incidência de pacientes na faixa etária de 40-60 anos.

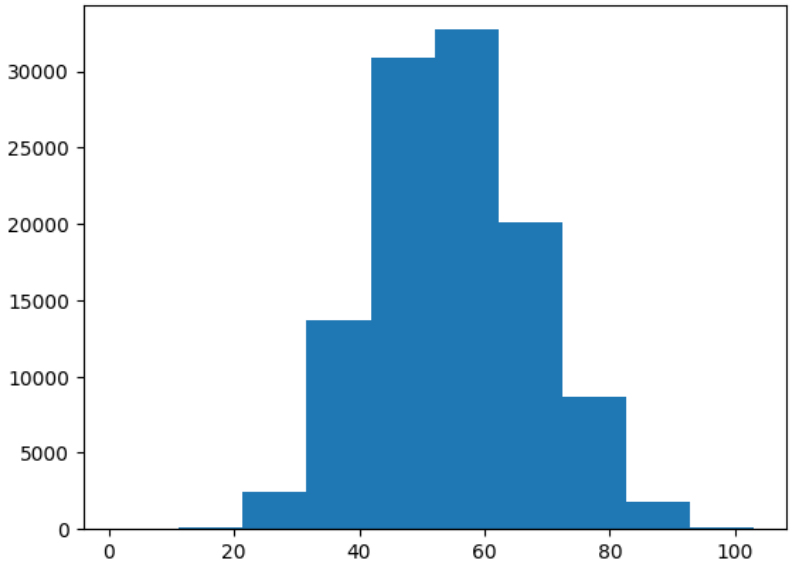

Figura 3. Distribuição de pacientes por faixa etária

Conforme a Figura 4, para evitar enviesamento das *features* numéricas sobre as *features* categóricas, foi utilizada a função de pré-processamento MinMaxScaler[3] da biblioteca SciKit Learn para limitar as features numéricas dentro do intervalo (0:1).

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
```

Figura 4. Distribuição de pacientes por faixa etária

Por fim, a base se encontrava pronta para ser analisada pelo algoritmo K-Prototypes.

3.2 Implementação do algoritmo K-Prototypes

Para aplicação do algoritmo, foi utilizada a biblioteca K-Modes[4] programada inteiramente em Python.

Como o único hiper parâmetro a ser definido no modelo é o valor K de clusters, foi utilizado o método do cotovelo [Shi e outros 2021], que consiste principalmente em encontrar o ponto em que o custo a ser otimizado apresenta a menor variação dentre uma certa banda de *K* escolhida (2 a 9, nesse caso). Ao analisar a Figura 5, pode-se concluir que o valor com menor variação é para $K = 5$, logo foi utilizado esse valor como base para treinar o modelo.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[4] https://github.com/nicodv/kmodes

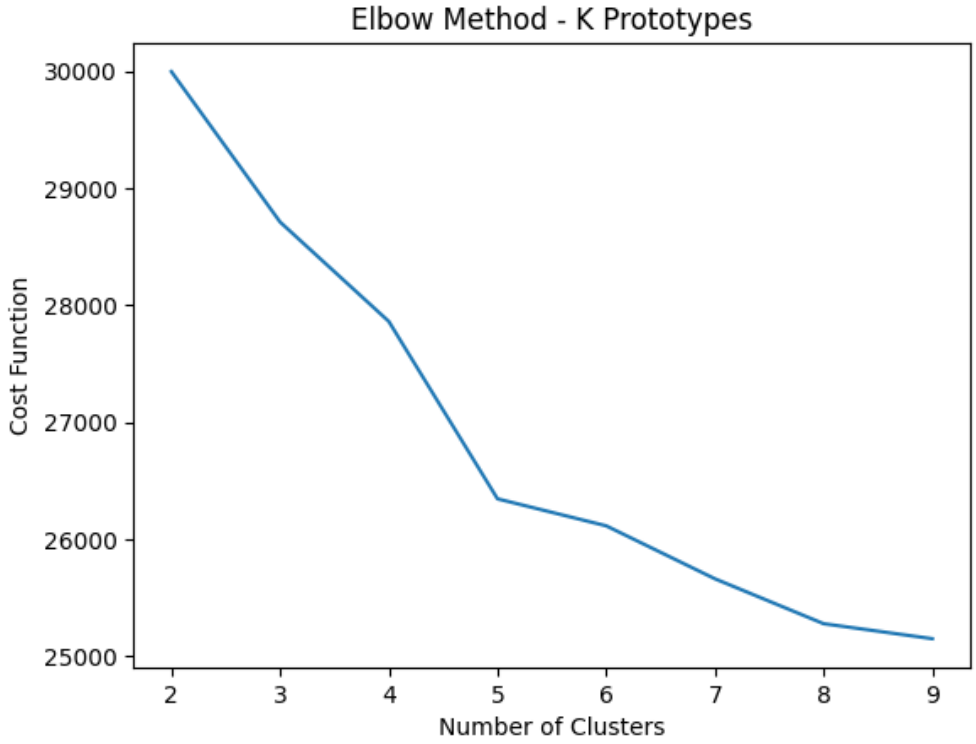

Figura 5. Função de custo versus número de clusters

3.3 Implementação do algoritmo K-Means

Para aplicação do algoritmo K-Means, foi utilizado a classe KMeans[5] da biblioteca SciKit Learn. Como o algoritmo K-Means depende da presença de *features* numéricas, os dados foram tratados para que cada categoria das *features* não numéricas se tornassem um parâmetro independente, denotado por 0 ou 1, num processo denominado One-Hot Encoding [Garavaglia e Sharma 1998].

Apesar do método aumentar a complexidade espacial do algoritmo, foi possível computar os clusters para o valor de *K=5* já calculado previamente, sem perda ou ganho de informação para o último algoritmo.

3.4 Implementação do algoritmo DBSCAN

Para implementar o algoritmo DBSCAN, foi utilizado da biblioteca Sci-kit learn, onde o modelo já estava pronto para uso. Como parâmetros, após testes de diversas combinações, foi utilizado como distância mínima 0.75 e como número mínimo de pontos para formar um cluster o número 235. (Obtido após testes)

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

4. Resultados

Conforme apresenta a Figura 6, ao aplicar o algoritmo K-Prototypes, obtemos 5 clusters, com as seguintes porcentagens de recidiva: *9,28%*, *25,6%*, *18,11%*, *13,01%* e *13.55%*.

```
Porcentagem de recidiva: 9.28%
Porcentagem de recidiva: 25.6%
Porcentagem de recidiva: 18.11%
Porcentagem de recidiva: 13.01%
Porcentagem de recidiva: 13.55%
```

Figura 6. Percentagem de recidivas nos clusters criados pelo algoritmo K-Prototypes

Já a Figura 7 apresenta as percentagens ao aplicar o algoritmo K-Means, na quais obtemos para os 5 clusters, respectivamente: *29,25%, 5,72%, 10,75%, 21,21%* e *3,23%*.

```
Porcentagem de recidiva: 29.25%
Porcentagem de recidiva: 5.72%
Porcentagem de recidiva: 10.75%
Porcentagem de recidiva: 21.21%
Porcentagem de recidiva: 3.23%
```

Figura 7. Percentagem de recidivas nos clusters criados pelo algoritmo K-Means

Como os clusters são formados aleatoriamente se baseando no ponto inicial escolhido, não seria possível fazer uma associação de 1:1 com relação aos clusters, porém, podemos esboçar um mapa de calor e checar a similaridade. A Figura 8 exibe este mapa de calor.

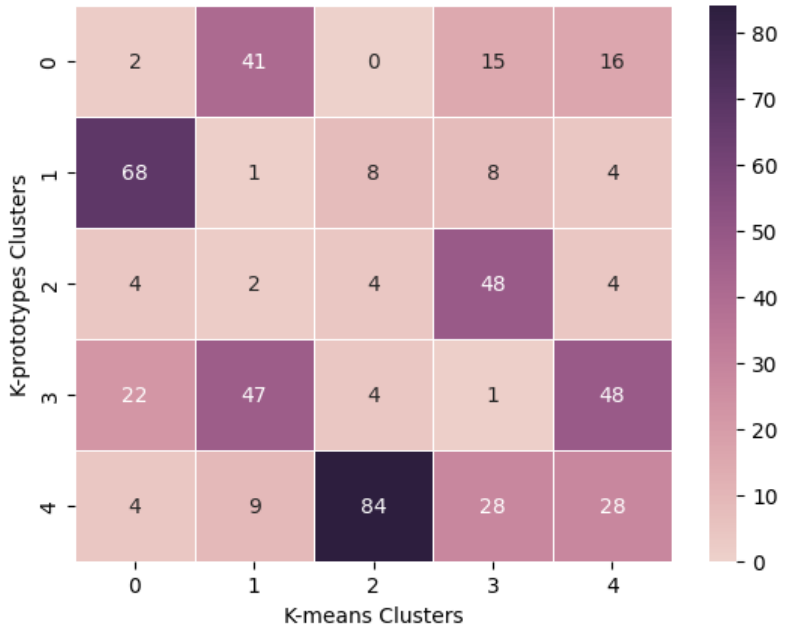

Figura 8. Mapa de calor de similaridade entres os clusters gerados pelos algoritmos K-prototypes e K-means

A Figura 9 apresenta as percentagens dos nove clusters originados do algoritmo DBSCAN, com variações de recidiva entre *5,51%* (cluster com menor índice de recidiva) e *23,03%* (cluster com maior índice de recidiva).

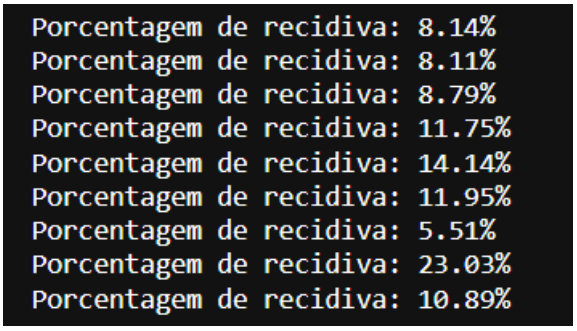

Figura 9. Percentagem de recidivas nos clusters criados pelo algoritmo DBSCAN

Ao analisar os clusters com o maior índice de recidiva, encontramos o cluster 2, 1 e 8 para o K-Prototypes, K-Means e DBSCAN, respectivamente. E, ao analisar as *features* de cada cluster, encontramos uma relação bem alta entre os 3, indicando que os três algoritmos convergiram para um cluster similar em que a recidiva foi presente em maior porcentagem dos casos. Os clusters encontrados pelo DBSCAN foram mais específicos na seleção de pacientes, sendo também menores que os demais.

Como características em destaque nos clusters me maior recidiva, foram encontrados que a idade média desses pacientes foi de 53 anos aproximadamente, sem diagnóstico anterior, tumor nos estágios IIB e IIIA, tumor de morfologia no Carcinoma ductal infiltrante (CID-O 85003) ou Carcinoma lobular, tumor entre 2 a 5 cm (T), metástase em linfonodos regionais e sem metástase a distância.

Tais características estão presentes em todos os 3 clusters com índice de recidiva maior que 23%, o que pode indicar uma forte relação entre o tipo de paciente e a ocorrência de uma recidiva, pelo menos em relação a outros tipos de clusters com dados diferentes.

5. Conclusão e Trabalhos Futuros

Os algoritmos apontaram que possivelmente existam uma correlação entre um certo grupo e o índice de recidiva. Entretanto, para concluir tais afirmações, seria necessário um estudo aprofundado de profissionais do domínio, bem como o uso de técnicas estatísticas, para checar se realmente essa correlação é plausível ou não existe uma certa tendência dos algoritmos.

Ao comparar os 3 algoritmos, obtemos clusters menores e mais concentrados ao utilizar o DBSCAN e clusters maiores e mais abrangentes com os algoritmos K-Prototypes e K-Means. Dentre eles, o que apresentou resultados mais consistentes (pouca variação) foi o K-Means, porém, a confirmação da existência de seus clusters por meio de outros algoritmos foi essencial na validação do modelo.

Como trabalhos futuros, pretende-se aplicar técnicas mais sofisticadas de clusterização tais como técnicas de clusterização profunda. A elaboração de um modelo próprio, com critérios de distância bem definidos e personalizados também seria uma alternativa viável. Além disso, planeja-se buscar uma base de dados com *features* mais específicas e detalhadas sobre o tumor/paciente.

6. Agradecimentos

Este trabalho foi financiado em parte pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) processo nº 405352/2021-2 e pelo programa PIBIC do CTI Renato Archer. Agradecemos a Fundação Oncocentro de São Paulo por disponibilizar publicamente os dados utilizados neste projeto.

7. Referência

American Cancer Society (2016) What Is Cancer Recurrence? Cancer.org report number 1.800.227.2345. Disponível em: https://www.cancer.org/content/dam/CRC/PDF/Public/8422.00.pdf acessado em 22 de agosto de 2023

American Cancer Society (2021) What Is Breast Cancer? Disponível em: https://www.cancer.org/cancer/types/breast-cancer/about/what-is-breast-cancer.html acessado em 22 de agosto de 2023

Garavaglia, S. e Sharma, A. (1998). A smart guide to dummy variables: Four applications and a macro. In Proceedings of the northeast SAS users group conference (Vol. 43).

Ministério da Saúde (2023) Câncer de Mama: sintomas, tratamentos, causas e prevenção. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-mama acessado em 22 de agosto de 2023

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

Kaufman, L. e Rousseeuw,P.J. (1990). FindingGroupsinData. Wiley,NewYork.

Schubert, E., Sander, J., Ester, M., Peter Kriegel, H.P., e Xiaowei Xu. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst. 42, 3, Article 19 (September 2017), 21 pages. https://doi.org/10.1145/3068335

Sinaga K. P. e Yang, M. S. (2020) Unsupervised K-Means Clustering Algorithm, in IEEE Access, vol. 8, pp. 80716-80727, doi: 10.1109/ACCESS.2020.2988796.

Shi, C., Wei, B., Wei, S. et al. (2021) A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. J Wireless Com Network 2021, 31. https://doi.org/10.1186/s13638-021-01910-w

Xu, D. e Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2, 165-193.