Aplicando Clusterização para Identificação de Grupos em Base de Dados do Questionário CUDIT

Clustering Application for Group´s Identification in CUDIT Database

Regis Borges[1]
orcid.org/0000-0002-0738-6765

Hugo Silva[2]
orcid.org/0000-0002-7958-2474

Ludmila Costa[2]
orcid.org/0000-0001-7137-4323

Paulo Rogério Morais[3]
orcid.org/0000-0002-0336-3793

Carmelo Bastos-Filho[1]
orcid.org/0000-0002-0924-5341

Kelsy Areco[3]
orcid.org/ 0000-0002-7801-757X

Dartiu da Silveira[3]
orcid.org/ 0000-0001-9264-904X

Dimitri Daldegan-Bueno[3]
orcid.org/ 0000-0002-9352-2873

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: rb@ecomp.poli.br; hugo.silva@upe.br; ludmila.costa@upe.br; paulo.morais@unir.br;carmelo.filho@upe.br ;kelsi.areco@unifesp.br;dartiu.silveira@unifesp.br; dimitridaldegan@gmail.com

[2]Faculdade de Ciências Médicas, Universidade de Pernambuco, Recife, Brasil.

[3] Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, Brasil

DOI: 10.25286/repa.v7i2.2216

Como citar este artigo pela NBR 6023/2018: Regis Borges; Hugo Silva; Ludmila Costa; Paulo Rogério Morais; Carmelo Bastos-Filho; Kelsy Areco Dartiu da Silveira; Dimitri Daldegan-Bueno. Aplicando Clusterização para Identificação de Grupos em Base de Dados do Questionário CUDIT. Revista de Engenharia e Pesquisa Aplicada, v.7, n. 2, p. 42-47, 2022

RESUMO

Nas últimas décadas, podemos verificar uma tendência em diversos países de regulamentar o uso medicinal e recreativo da *cannabis*, o que pode gerar um enorme mercado global. No entanto, o histórico de quase um século na condição de ilegalidade, em praticamente todo o planeta, contribui para que a literatura científica sobre o tema seja severamente limitada. Partindo dessa problemática, o objetivo deste artigo é apresentar um modelo computacional, utilizando a linguagem Python e empregando técnicas de clusterização, tais como KMeans, Agglomerative Clustering e Spectral Clustering, de modo a encontrar grupos de interesse numa massa de dados. A base de dados utilizada foi obtida da aplicação do questionário do teste CUDIT – Cannabis Use Disorder Identification Test, composto de oito perguntas com cinco opções de resposta, que permitiram graduar o comportamento do indivíduo em cada questão. Os resultados demonstraram que a técnica foi capaz de separar grupos com padrões distintos e que apresentam consistência em análise qualitativa preliminar.

PALAVRAS-CHAVE: Clusterização; Python; *cannabis*; CUDIT.

ABSTRACT

In recent decades, we can see a trend in several countries to regulate the medicinal and recreational use of cannabis, which can generate a huge global market. However, the history of almost a century in the condition of illegality, in practically the entire planet, contributes to the scientific literature on the subject being severely limited. Based on this issue, the aim of this article is to present a computational model, using the Python language and employing clustering techniques, such as KMeans, Agglomerative Clustering and Spectral Clustering, in order to find interest groups in a mass of data. The database used was obtained from the application of the CUDIT test questionnaire – Cannabis Use Disorder Identification Test, composed of eight questions with five answer options, which allowed grading the individual's behavior in each question. The results showed that the technique was able to separate groups with different patterns and that present consistency in a preliminary qualitative analysis.

KEY-WORDS: Clustering; Python; *cannabis*; CUDIT;

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

Há vários estudos que demonstram que o uso de drogas é uma prática desde tempos pré-históricos (Cf. MERLIN, 2003). Segundo o autor, há provas arqueológicas do uso de substâncias psicoativas que remontam 10 mil anos atrás, e evidência histórica do uso cultural há 5 mil anos. **[1]** Isto denota que o uso de substâncias psicoativas remete às origens das primeiras sociedades humanas. A finalidade do uso tende a variar principalmente entre o medicinal, religioso, cultural e recreativo.

Dentre a vasta gama de substâncias classificadas como drogas, uma das mais populares mundialmente é a *cannabis*, mais popularmente conhecida no Brasil como "maconha". Trata-se de uma espécie de planta originária da Ásia cujo cultivo remonta milênios atrás.

A criminalização da *cannabis* teve origem em diversos países, após a Convenção Internacional do Ópio realizada em 1912. No Brasil, desde então, foram publicados alguns Decretos referentes à criminalização de entorpecentes. O Decreto-Lei 891, de 25/11/1938, por exemplo, proibiu o "plantio, a cultura, a colheita e a exploração" em território nacional da "*Cannabis* sativa" e sua variedade "índica", exceto "para fins terapêuticos", "desde que haja parecer favorável da Comissão Nacional de Fiscalização do Entorpecentes". **[2]**

Nas últimas décadas, percebe-se uma tendência em descriminalizar e regulamentar o uso da *cannabis*. Em 2013, o Uruguai foi a primeira nação a legalizar a produção, venda e consumo de *cannabis* de modo regulamentado; nos Estados Unidos, apesar da criminalização em nível federal, diversos estados regulamentaram o uso medicinal e outros, tais com Califórnia, Colorado e Nevada regulamentaram inclusive o uso recreativo.

Apesar do histórico milenar, a condição de ilegalidade imposta no início do século XX resulta em baixo grau de conhecimento científico a respeito do uso e efeitos relacionados a *cannabis*. Tanto o uso como medicamento na indústria farmacêutica, bem como o uso recreativo, que se manteve popular em ambiente de ilegalidade, indicam o alto potencial comercial da *cannabis*.

Desenvolver conhecimento científico sobre a *cannabis* e seus efeitos parece ser primordial para conhecer os riscos e buscar estabelecer condições seguras de uso, bem como tratamentos adequados para casos de abuso.

Alinhado a esse contexto, Adamson e Sellman (2003) **[3]** conduziram o estudo que aplicou pela primeira vez o teste CUDIT (*Cannabis* Use Disorder Identification Test), que consiste em dez perguntas com respostas graduadas, direcionadas a mapear, dentre outros, a frequência de consumo, o grau de possíveis efeitos relacionados ao uso e a sugestão de alguém para cortar o consumo. O questionário CUDIT foi formulado como uma modificação do teste AUDIT, concebido para mapear transtornos no uso de álcool.

Ao longo do tempo, o questionário CUDIT foi alvo de propostas de reformulação, incluindo as apresentadas por Adamson *et al* (2010) **[4]** e Loflin *et al* (2017) **[5].**

A aplicação da tradução do questionário proposto por Loflin *et al* (2017) **[5]**, em uma pesquisa realizada pela internet, resultou em uma base de dados que reúne 7876 amostras de pessoas que admitiram o consumo mínimo de *cannabis* mapeado pelo questionário.

A aplicação de métodos estatísticos convencionais não permitiu a identificação de diferentes grupos de usuários nas amostras da base de dados.

2 OBJETIVOS

O objetivo geral desta pesquisa é construir um modelo computacional, empregando técnicas de clusterização capazes de identificar grupos de interesse na massa de dados obtidas com a aplicação do questionário CUDIT.

3 MATERIAIS E MÉTODOS

A partir do arquivo *Comma Separated Values* (CSV) que contém a base de dados, todo o processamento foi efetuado por algoritmos escritos em Python. As bibliotecas do scikitLearn foram utilizadas nos processos de normalização, redução de características, clusterização e avaliação dos resultados.

3.1 Pré-processamento

O pré-processamento dos dados consistiu em excluir todas as colunas que não fossem atreladas às respostas do questionário CUDIT. O conjunto resultante foi normalizado de forma que os dados de entrada do algoritmo de clusterização consistem em uma tabela com 7876 linhas referentes a cada usuário participante e 8 colunas contendo valores

DOI: 10.25286/repa.v7i2.2216

numéricos normalizados referentes às respostas para cada uma das perguntas do questionário.

Após observar resultados insatisfatórios quando aplicados os processos de clusterização diretamente na base de dados original, buscamos aplicar técnicas de redução de "features" ou colunas da base dados. Verificamos que a aplicação da técnica de Features Agglomeration, reduzindo de 8 para 3 colunas de atributos, melhorou significativamente o resultado das avaliações. Cabe esclarecer que a função Features Agglomeration aplica técnica de clusterização partindo de um cluster para cada coluna de dados original e vai reunindo as colunas mais semelhantes até que sobre apenas a quantidade de colunas desejada.

No conjunto de dados de três colunas resultantes, verificamos que a pergunta 1 do questionário CUDIT referente à frequência de uso é integralmente transposta para a saída. O mesmo ocorre com a pergunta 8 referente à intenção de reduzir ou parar o consumo, enquanto as perguntas de 2 a 7 foram aglutinadas, matematicamente, em um único atributo.

3.2 Métodos de clusterização

Na sequência, foram aplicados os seguintes métodos de clusterização por meio de funções disponíveis na biblioteca do scikit-Learn:

3.1.1 Kmeans

Tem por objetivo minimizar a inércia. A partir de uma quantidade estabelecida de clusters K, estabelece K amostras da base de dados como os centróides iniciais. Cada amostra da base de dados é associada ao centróide mais próximo, calcula-se a média de cada grupo e atribui-se um novo conjunto de centróides. Este processo é repetido até que se verifique deslocamento dos centróides inferior ao critério de parada.

3.1.2 Agglomerative Clustering

Aplica clusterização hierárquica de baixo para cima, inicia com cada amostra em seu próprio cluster e, sucessivamente, vão se aglutinando de acordo com a métrica parametrizada como critério de ligação.

3.1.3 Spectral Clustering

Aplica processo de incorporação de baixa dimensão da matriz de afinidade entre as amostras e, na sequência, aplica um método de clusterização tal como o KMeans.

3.1.4 Mean-shift, DBSCAN e Affinity Propagation

Estes métodos foram implementados no algoritmo, porém não apresentaram resultados satisfatórios.

3.3 Métricas de avaliação

A métrica de avaliação utilizada foi o Coeficiente de Silhueta, que consiste em um método que utiliza os próprios dados da base para avaliação dos clusters. O coeficiente é determinado pela equação (1):

$$s = \frac{b-a}{max(a,b)} \quad (1)$$

Onde:

s: Coeficiente de Silhueta

a: A distância média entre cada amostra e a média das demais amostras do mesmo grupo.

b: A distância média entre cada amostra e a média de todas as amostras do grupo mais próximo.

Os métodos Calinski-Harabasz, proposto em 1974 **[6]** e Davies-Bouldin, proposto em 1979 **[7]**, também foram implementados no algoritmo para verificar a consistência dos grupos encontrados.

4 RESULTADOS

Além de obter o modelo com o maior Coeficiente de Silhueta, buscamos limitar o número de clusters, já que um grande número de clusters significaria grande número de grupos muito específicos de usuários, o que dificultaria a interpretação e associação a possíveis propostas de ação clínica. Considerando que a base de dados inclui uma classificação original de três categorias e que os modelos retornaram bons resultados quando formados três grupos, a busca final foi encontrar as melhores soluções formando três grupos.

Segue a avaliação dos melhores resultados utilizando 3 modelos diferentes:

4.1 Kmeans

Obteve-se o coeficiente de Silhueta de 0,512.

O cluster 0 concentrou 3591 (52,2%) dos indivíduos, o cluster 1 828 (12,1%) e o cluster 2 2457 (35,7%).

Os valores mais frequentes para cada resposta, em cada grupo formado pela clusterização, apresentados na Tabela 1 ilustra de modo mais direto as principais características dos grupos:

- Cluster 0: Usuários muito frequentes (CUDIT_1=4) com intenção de reduzir ou parar o consumo (CUDIT_8=4)
- Cluster 1: Usuários pouco frequentes (CUDIT_1=2)
- Cluster 2: Usuários muito frequentes (CUDIT_1=4) sem intenção de reduzir ou parar o consumo (CUDIT_8=0)

Tabela 1- Valores Mais Frequentes por Cluster – Método KMeans

cluster	sexo	idade	CUDIT_1	CUDIT_2	CUDIT_3	CUDIT_4	CUDIT_5	CUDIT_6	CUDIT_7	CUDIT_8
0	1	18	4	2	0	0	0	1	0	4
1	1	18	2	1	0	0	0	0	0	0
2	1	18	4	2	0	0	0	0	0	0

Fonte: Os Autores.

A tabela 2 permite visualizar, dentro de cada cluster, a quantidade de casos para cada uma das combinações de respostas 1 e 8 possíveis. Observando o cluster 2, por exemplo, fica muito evidente que há um agrupamento de todos os usuários que responderam 3 ou 4 na pergunta 1 e 0 na pergunta 8, sustentando, dessa forma, a descrição de usuários frequentes que não manifestam intenção de parar.

Tabela 2- Avaliação das respostas CUDIT_1 e CUDIT_8 por cluster – Método KMeans

CUDIT_1	1			2			3			4		
CUDIT_8	0	2	4	0	2	4	0	2	4	0	2	4
cluster 0	0	1	82	0	14	163	0	250	356	0	1093	1632
cluster 1	276	62	0	394	96	0	0	0	0	0	0	0
cluster 2	0	0	0	0	0	0	456	0	0	2001	0	0

Fonte: Os Autores.

O cluster 0 agrupa todos os demais que responderam 3 ou 4 à pergunta 1, mas que indicaram algum grau de intenção em reduzir ou parar. Além desses, agrupa todos os demais que responderam 4 à pergunta 8 e uma pequena parte dos que responderam 2 na pergunta 8 e 1 ou 2 à pergunta 1. A manifestação de intenção de reduzir ou parar é a principal característica do grupo e se acentua quanto maior for a frequência de uso.

O cluster 1 não contempla nenhum caso de resposta 3 ou 4 à pergunta 1. A falta ou pouca intenção de reduzir ou parar pode ser consequência do baixo nível de consumo de modo que a caracterização deste grupo como usuários pouco frequentes parece mais adequada que incluir a observação de que manifestam nenhuma ou moderada intenção de reduzir ou parar.

4.2 Spectral clustering

Obteve-se o coeficiente de Silhueta de 0,516.

O cluster 0 concentrou 893 (13,0%) dos indivíduos, o cluster 1 3526 (51,3%) e o cluster 2 2457 (35,7%).

Os valores mais frequentes para cada resposta, em cada grupo formado pela clusterização, apresentados na Tabela 3, ilustra de modo mais direto as principais características dos grupos:

- Cluster 0: Usuários pouco frequentes (CUDIT_1=2)
- Cluster 1: Usuários muito frequentes (CUDIT_1=4) com intenção de reduzir ou parar o consumo (CUDIT_8=4)
- Cluster 2: Usuários muito frequentes (CUDIT_1=4) sem intenção de reduzir ou parar o consumo (CUDIT_8=0)

Tabela 3- Valores Mais Frequentes por Cluster – Método Spectral Clustering

cluster	sexo	idade	CUDIT_1	CUDIT_2	CUDIT_3	CUDIT_4	CUDIT_5	CUDIT_6	CUDIT_7	CUDIT_8
0	1	18	2	1	0	0	0	0	0	0
1	1	18	4	2	0	0	0	1	0	4
2	1	18	4	2	0	0	0	0	0	0

Fonte: Os Autores.

A tabela 4 permite visualizar, dentro de cada cluster, a quantidade de casos para cada uma das combinações de respostas 1 e 8 possíveis. Observando o cluster 2, por exemplo, fica muito evidente que há um agrupamento de todos os usuários que responderam 3 ou 4 na pergunta 1 e 0 na pergunta 8, sustentando, dessa forma, a descrição de usuários frequentes que não manifestam intenção de parar.

Tabela 4- Avaliação das respostas CUDIT_1 e CUDIT_8 por cluster – Método Spectral Clustering

CUDIT_1	1			2			3			4		
CUDIT_8	0	2	4	0	2	4	0	2	4	0	2	4
cluster 0	276	63	69	394	91	0	0	0	0	0	0	0
cluster 1	0	0	13	0	19	163	0	250	356	0	1093	1632
cluster 2	0	0	0	0	0	0	456	0	0	2001	0	0

Fonte: Os Autores.

O cluster 1 agrupa todos os demais que responderam 3 ou 4 à pergunta 1, mas que indicaram algum grau de intenção em reduzir ou parar. Além desses, agrupa parte dos usuários de menor frequência que manifestam intenção de reduzir ou parar. A manifestação de intenção de reduzir ou parar é a principal característica do grupo e se acentua quanto maior for a frequência de uso.

DOI: 10.25286/repa.v7i2.2216

O cluster 0 não contempla nenhum caso de resposta 3 ou 4 à pergunta 1, levando a caracterização desse grupo como casos de usuários pouco frequentes. A intenção de parar pouco afeta quando CUDIT_1 = 1, mas, quando CUDIT_1 = 2, a intenção de parar atua mais fortemente no sentido de deslocar usuários de consumo menos frequentes para o grupo de usuários mais frequentes que manifestam intenção de reduzir ou parar.

4.3 Agglomerative clustering

Obteve-se o coeficiente de Silhueta de 0,496.

Os valores mais frequentes para cada resposta, em cada grupo formado pela clusterização, apresentados na Tabela 5, ilustra de modo mais direto as principais características dos grupos:

- Cluster 0: Usuários pouco frequentes (CUDIT_1=2)
- Cluster 1: Usuários muito frequentes (CUDIT_1=4) sem intenção de reduzir ou parar o consumo (CUDIT_8=0)
- Cluster 2: Usuários muito frequentes (CUDIT_1=4) com intenção de reduzir ou parar o consumo (CUDIT_8=4)

Tabela 5- Valores Mais Frequentes por Cluster – Método Agglomerative Clustering

cluster	sexo	idade	CUDIT_1	CUDIT_2	CUDIT_3	CUDIT_4	CUDIT_5	CUDIT_6	CUDIT_7	CUDIT_8
0	1	18	2	1	0	0	0	0	0	0
1	1	18	4	2	0	0	0	0	0	0
2	1	18	4	2	0	0	0	1	0	4

Fonte: Os Autores.

A tabela 6 permite visualizar, dentro de cada cluster, a quantidade de casos para cada uma das combinações de respostas 1 e 8 possíveis. Observando o cluster 2, por exemplo, fica muito evidente que há o agrupamento de todos os usuários que responderam 3 ou 4 na pergunta 1 e 4 na pergunta 8, sustentando, dessa forma, a descrição de usuários frequentes que manifestam clara intenção de parar.

Tabela 6- Avaliação das respostas CUDIT_1 e CUDIT_8 por cluster – Método Agglomerative Clustering

CUDIT_1	1			2			3			4		
CUDIT_8	0	2	4	0	2	4	0	2	4	0	2	4
Cluster 0	276	63	82	394	110	163	0	0	0	0	0	0
Cluster 1	0	0	0	0	0	0	456	250	0	2001	1093	0
Cluster 2	0	0	0	0	0	0	0	0	356	0	0	1632

Fonte: Os Autores.

O cluster 1 agrupa todos os demais que responderam 3 ou 4 à pergunta 1 e 0 ou 2 à pergunta 8, caracterizando usuários frequentes com moderada ou nenhuma intenção de reduzir ou parar.

O cluster 0 não contempla todos os casos de resposta 1 ou 2 à pergunta 1, caracterizando esse grupo como usuários pouco frequentes.

5 CONCLUSÕES

Numa comparação final, verificamos resultados próximos entre os 3 métodos aplicados, sendo que o aglomerativo apresentou desempenho levemente inferior. Há basicamente um empate técnico entre os métodos Spectral e Kmeans, com sutil vantagem para o primeiro, considerando o objetivo do estudo.

Alcançamos, portanto, o objetivo de encontrar grupos de interesse entre os dados analisados. A obtenção de variações nos resultados possibilita que, futuramente, se aplique uma avaliação qualitativa por profissionais da área de saúde, de modo que, além de bom resultado do modelo computacional, os grupos formados tenham maior relevância para a identificação de casos problemáticos, caracterização dos grupos e aplicação de abordagens de informação e tratamento.

REFERÊNCIAS

[1] MERLIN, M. D. Archaeological evidence for the tradition of psychoactive plant use in the old World. **Economic Botany.** New York, v. 57, n. 3, 2003, p. 295–323. Disponível em: http://www.jstor.org/stable/4256701. Acesso em: 08 dez. 2021.

[2] BRASIL. Decreto-lei nº 891, de 25 de novembro de 1938. Dispõe sobre a fiscalização de entorpecentes. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto-lei/1937-1946/del0891.htm. Acesso em: 08 dez. 2021.

[3] ADAMSON, S. J.; SELLMAN, J. D. A prototype screening instrument for cannabis use disorder: the cannabis use disorders identification test (CUDIT) in an alcohol-dependent clinical sample. **Drug and Alcohol Review.** V. 22, n. 3, 2003, p. 309-315. Disponível em: https://doi.org/10.1080/09595230310001544 54. Acesso em: 08 dez. 2021.

[4] ADAMSON, S. J. *et al.* An improved brief measure of cannabis misuse: the Cannabis use disorders identification test-revised (CUDIT-R). **Drug and Alcohol Dependence.** V. 110, n. 1-2, Jul. 2010, p. 137 – 143. Disponível em: https://doi.org/10.1016/j.drugalcdep.2010.02.017. Acesso em: 08 dez. 2021.

[5] LOFLIN, M. *et al*. Assessment of the validity of the CUDIT-R in a subpopulation of cannabis users. **The American Journal of Drug and Alcohol Abuse**. V. 44, n. 1, p. 19-23, 2018. Disponível em: http://dx.doi.org/10.1080/00952990.2017.1376677. Acesso em: 08 dez. 2021.

[6] CALINSKI, T.; J HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics. V. 3, N. 1, p. 1-27. DOI: 10.1080/03610927408827101. Acesso em: 08 dez. 2021.

[7] DAVIES, D.L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence.** Vol. PAMI-1, n. 2, p. 224-227, April 1979. Doi: 10.1109/TPAMI.1979.4766909. Acesso em: 08 dez. 2021.

DOI: 10.25286/repa.v7i2.2216