57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

ANÁLISE COMPUTACIONAL DE BOLHAS SOCIAIS: UM ESTUDO DE CASO NA UNIVERSIDADE FEDERAL DO CEARÁ - CAMPUS RUSSAS

Ilton C. de Lima Neto, Jailon William B. Oliveira da Silva, Pablo Luiz Braga Soares, Antonio Janderson Nascimento de Azevedo, Carlos Vinicio D. Araújo
Laboratório de Pesquisa & Desenvolvimento do NEMO, Universidade Federal do Ceará
Avenida Felipe Santiago – Nº 411, 62.900-420, Russas/CE, Brasil
{ilton.cavalcante, williambrunocc, carlosvinicio}@alu.ufc.br,
jjandersonazevedo@gmail.com pablo.soarea@ufc.br

RESUMO

Este trabalho investiga a formação de bolhas sociais entre estudantes da Universidade Federal do Ceará – Campus Russas. A metodologia adotada envolveu a aplicação de um questionário com 44 perguntas, resultando na coleta de 196 respostas. Os dados foram pré-processados por meio de codificação, engenharia de atributos e normalização, gerando uma base estruturada. Duas abordagens analíticas foram utilizadas: o algoritmo *k-means*, para agrupar os estudantes com base na similaridade de atributos, e a modelagem do problema como uma instância de corte máximo em grafos. Além disso, um modelo de *Random Forest* foi treinado para identificar os fatores mais relevantes na segmentação, alcançando 92.59% de acurácia. Afinidades culturais e hábitos de convivência se destacaram como os principais vetores de formação das bolhas. Os resultados indicam uma baixa integração entre estudantes de diferentes cursos, destacando a necessidade de políticas institucionais que incentivem interações interdisciplinares e o uso compartilhado de espaços acadêmicos.

PALAVRAS CHAVE. Segmentação social, *k-means*, Corte Máximo.

Tópicos: AD&GP – Pesquisa Operacional na Administração e Gestão da Produção; EST&AM – PO Analytics em Estatística e Aprendizado de Máquina

ABSTRACT

This study investigates the formation of social bubbles among students at the Federal University of Ceará – Russas Campus. The adopted methodology involved administering a questionnaire with 44 questions, resulting in the collection of 196 responses. The data were pre-processed through encoding, feature engineering, and normalization, producing a structured dataset. Two analytical approaches were applied: the k-means algorithm, used to cluster students based on attribute similarity, and the modeling of the problem as a maximum cut instance in graphs. Additionally, a Random Forest model was trained to identify the most influential factors in the segmentation process, achieving an accuracy of 92.59%. Cultural affinities and social habits emerged as the main drivers behind the formation of these bubbles. The findings indicate low integration among students from different academic programs, highlighting the need for institutional policies that promote interdisciplinary interactions and the shared use of academic spaces.

KEYWORDS. Social segmentation. *k-means*. Max-Cut.

Paper topics: AD&GP – OR in Production Management and Administration; EST&AM – OR Analytics in Statistics and Machine Learning

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

1. Introdução

A crescente diversidade dos perfis estudantis nas universidades públicas brasileiras, impulsionada por políticas de inclusão e interiorização do ensino superior, tem provocado mudanças significativas nas dinâmicas sociais dentro dos campi. Ainda assim, apesar da coexistência de múltiplos cursos e contextos culturais, nem sempre há uma integração efetiva entre os estudantes, o que pode resultar em fragmentações no tecido social universitário. Esse fenômeno, frequentemente descrito como "formação de bolhas sociais", é recorrente em diversos contextos, como no ambiente acadêmico, no local de trabalho, nas redes sociais e em comunidades locais e caracteriza-se pela criação de grupos com forte coesão interna e limitada interação com outros, dificultando a construção de redes colaborativas e interdisciplinares o que gera um isolamento social relativo [Pariser, 2011].

No contexto do campus da Universidade Federal do Ceará (UFC) em Russas, a presença de cursos com naturezas técnico-científicas distintas e a estrutura física compartilhada entre eles oferecem uma oportunidade ímpar para analisar como tais bolhas se formam e se mantêm. Estudos anteriores têm evidenciado que fatores como afinidade acadêmica, segmentações espaciais e ausência de políticas de engajamento podem reforçar o isolamento social entre estudantes [Bond e Bergdahl, 2023; Oliveira et al., 2020].

Com base nesse cenário, técnicas de mineração de dados vêm sendo adotadas como instrumentos capazes de revelar padrões latentes de comportamento e agrupamento. Dentre elas, algoritmos de clusterização, como o *k-means*, tem se destacado na literatura por sua aplicabilidade em contextos educacionais e sociais [Maia et al., 2022; Gurusamy, 2017]. Além disso, algoritmos de clusterização partem do pressuposto de que pontos distantes devem pertencer a grupos distintos. Nesse contexto, o problema do corte máximo (Max-Cut) consiste em dividir os vértices de um grafo em dois conjuntos, de modo que a soma dos pesos das arestas que os conectam seja a maior possível. Assim, ao maximizar as distâncias entre vértices de diferentes grupos, também se maximiza o valor do corte, o que, por consequência, leva ao agrupamento de elementos mais semelhantes dentro de cada grupo — caracterizando, portanto, uma forma natural de clusterização [Otterbach et al., 2017]. A colaboração entre diferentes métodos de agrupamento permite não apenas identificar padrões de segmentação, mas também compreender as nuances estruturais e culturais que os sustentam.

Este trabalho tem como objetivo investigar a formação de bolhas sociais entre estudantes de graduação e pós-graduação do campus da UFC em Russas, a partir de informações coletadas por meio de questionários, que compõem a base de dados utilizada na análise. Por meio da aplicação do algoritmo *k-means* e da resolução de uma instância do problema do Max-Cut, propõe-se uma análise cooperativa e quantitativa dos padrões de interação interpessoal no campus. Espera-se, com isso, fornecer subsídios empíricos que contribuam para o desenvolvimento de estratégias institucionais voltadas à promoção da inclusão, do engajamento e da cooperação entre diferentes grupos acadêmicos.

O restante deste trabalho está estruturado da seguinte forma: na Seção 2, são apresentados os fundamentos teóricos e metodológicos sobre bolhas sociais, clusterização e os algoritmos utilizados nesta pesquisa, incluindo *k-means* e *Random Forest*, além da explanação sobre a formalização matemática do Max-Cut; a Seção 3 descreve o delineamento metodológico do estudo, abrangendo o processo de coleta dos dados por meio de questionários, bem como os procedimentos de análise e aplicação dos algoritmos de agrupamento; a Seção 4 apresenta os resultados obtidos com a aplicação dos métodos, discutindo os padrões de segmentação social identificados no campus da UFC em Russas; por fim, a Seção 5 expõe as conclusões e contribuições deste estudo.

2. Fundamentação Teórica

Este tópico apresenta os fundamentos teóricos e metodológicos deste trabalho. Na Seção 2.1, discutem-se as bolhas sociais e o isolamento em ambientes acadêmicos, além da aplicação de

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

mineração de dados e clusterização para identificar padrões sociais. A Seção 2.2 detalha o algoritmo *k-means*, amplamente utilizado em agrupamento de dados. Na Seção 2.3, apresenta-se o Problema do Corte Máximo, sua definição formal e aplicações em agrupamento e análise de redes sociais. Por último, a Seção 2.4 descreve o algoritmo *Random Forest*, suas principais características e usos em classificação.

2.1. Bolhas Sociais e Isolamento em Ambientes Acadêmicos

A compreensão das interações sociais em ambientes universitários exige uma abordagem multidimensional, que contemple tanto os aspectos comportamentais quanto os estruturais envolvidos no engajamento estudantil. A literatura contemporânea reconhece o engajamento como um fenômeno complexo, composto por dimensões cognitivas, afetivas e sociais, sendo esta última frequentemente negligenciada nas análises institucionais [Bond e Bergdahl, 2023].

As chamadas "bolhas sociais" configuram-se como agrupamentos fechados de indivíduos que compartilham características similares — sejam elas acadêmicas, culturais ou comportamentais — e que apresentam baixa propensão à interação com grupos distintos. Em contextos universitários, essa segmentação pode ser reforçada por múltiplos fatores: organização curricular por curso, barreiras socioculturais, desigualdade no acesso a espaços de convivência e carência de iniciativas integradoras por parte das instituições [Oliveira et al., 2020].

Esses agrupamentos podem gerar efeitos negativos sobre o clima acadêmico, dificultando a troca interdisciplinar e reduzindo o sentimento de pertencimento de determinados estudantes. Em especial, campi com infraestrutura compartilhada entre cursos distintos, como o da UFC em Russas, tornam-se ambientes propícios para investigar como tais bolhas se configuram.

Além disso, a identificação de padrões de relacionamento interpessoal em contextos educacionais tem sido impulsionada pelo uso de técnicas de mineração de dados. A clusterização — ou agrupamento — é uma dessas técnicas, e visa organizar dados em subconjuntos (clusters) de modo que itens dentro de um mesmo grupo apresentem maior similaridade entre si do que com itens de outros grupos [Gurusamy, 2017].

No campo educacional, a clusterização já foi aplicada com êxito em estudos sobre perfis de engajamento, análise de comportamento discente e identificação de riscos acadêmicos [Martins et al., 2021; Maia et al., 2022]. Tais abordagens permitem mapear agrupamentos latentes com base em respostas a questionários, traçando uma representação empírica das dinâmicas sociais presentes.

2.2. Algoritmo de Agrupamento: *k-means*

O algoritmo *k-means* é uma técnica de aprendizado de máquina não supervisionado utilizada para realizar a tarefa de agrupamento, isto é, seu objetivo é particionar um conjunto de dados em k grupos (*clusters*) de forma que os elementos dentro de cada *cluster* sejam mais semelhantes entre si do que em relação aos de outros *clusters*. Para isso, a técnica tenta minimizar o erro quadrático entre a média empírica (centróide) de um *cluster* e os pontos que pertencem a ele [MacQueen et al., 1967].

Conforme Bishop e Nasrabadi [2006], a função objetivo do algoritmo *k-means* é dada por:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{1}$$

onde J representa a soma das distâncias quadráticas entre cada ponto x_n e o centróide μ_k do *cluster* C_k, sendo $r_{nk} = 1$ se x_n pertence ao *cluster* k, e 0 caso contrário. O objetivo é minimizar J, garantindo que os pontos estejam o mais próximos possível de seus centróides. A cada iteração, J é minimizado primeiro em relação a r_{nk}, atribuindo cada ponto ao centróide mais próximo, e

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

depois em relação a μ_k, atualizando os centróides como a média dos pontos do grupo. O processo se repete até a convergência, assegurando a melhor alocação dos pontos aos respectivos *clusters*.

A Figura 1 ilustra a comparação entre os dados antes e depois da aplicação do *k-means*. Na Figura 1 (a), todos os pontos são exibidos na mesma cor, sem nenhum grupo definidos, representando o estado inicial dos dados. Já na Figura 1 (b), após aplicar o *k-means*, os pontos são agrupados em grupos distintos, sendo representados com cores diferentes, e os centróides são destacados na cor preta.

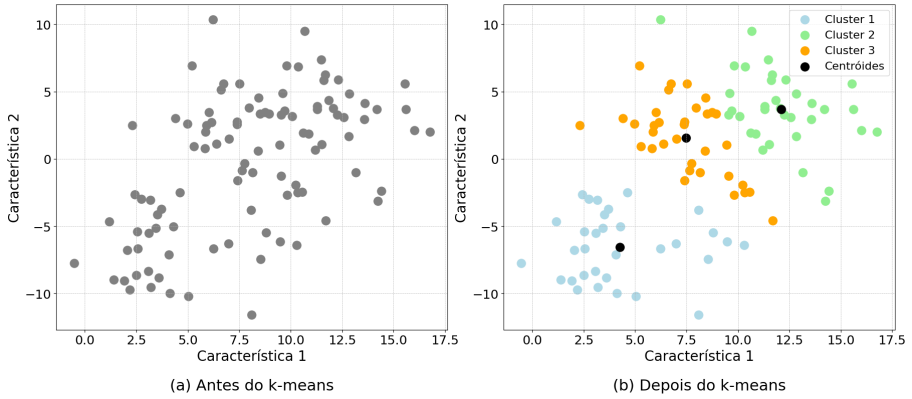


Figura 1: Aplicação do método *k*-means para o particionamento dos dados.

Durante a execução do *k-means*, a escolha do número de grupos k é essencial e pode ser feita pelo método do cotovelo (*Elbow Method*) ou a análise de silhueta. O método do cotovelo baseia-se no coeficiente de aglomeração, traçado em função de k, sendo o ponto de inflexão do gráfico o valor ideal [Ketchen e Shook, 1996]. Já a análise de silhueta avalia o quão próximos os pontos estão de seus próprios grupos em relação aos vizinhos [Rousseeuw, 1987].

2.3. Problema do Corte Máximo

O Problema do Corte Máximo (Max-Cut) é um problema de otimização combinatória, classificado como NP-difícil [Karp, 1972]. Ele pode ser definido da seguinte forma: dado um grafo não direcionado $G = (V, E)$, com pesos não negativos W_i nas arestas $(i, j) \in E$, o objetivo é encontrar o subconjunto de vértices S que maximize o peso das arestas no corte $(S, \overline{S})$, ou seja, o peso das arestas com um extremo em S e o outro em $\overline{S}$ [Goemans e Williamson, 1995].

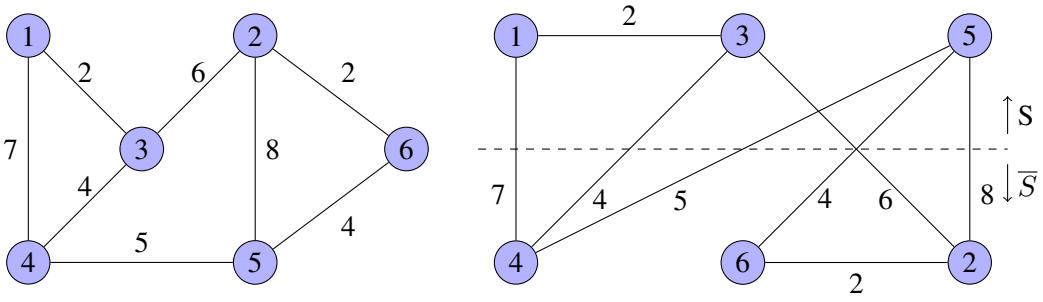
A Figura 2 ilustra uma instância do Problema do Corte Máximo. Na Figura 2 (a), o grafo $G = (V, E)$ possui $|V| = 6$ vértices e $|E| = 8$ arestas ponderadas. A Figura 2 (b) mostra a solução ótima, onde o agrupamento dos vértices é $S = \{1, 3, 5\}$ e $\overline{S} = \{2, 4, 6\}$. As arestas de corte, conectando vértices de S e $\overline{S}$, são $\{\{1, 4\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{5, 2\}, \{5, 6\}\}$, com valor de corte igual a 34.

Esse problema possui aplicações práticas significativas, especialmente por fazer parte da classe de problemas de otimização, pois muitos problemas surgem diretamente em seu formato, enquanto outros podem ser reduzidos a ele, assim como outros problemas discutidos por Garey e Johnson [1979] e Jonathan L. Gross [2013]. Em áreas como agrupamentos de dados, por exemplo, cada vértice pode representar um dado, e o peso das arestas refletem a dissimilaridade entre eles. Desta forma, maximizar o valor de corte equivale a agrupar elementos mais semelhantes, representando um algoritmo de agrupamento natural [Otterbach et al., 2017].

Outro exemplo esta na análise de redes sociais, onde vértices representam autores de uma postagem e suas arestas, representam interações em discussões. Note que desta forma, o corte máximo pode separar usuários em grupos com opiniões opostas sobre um determinado tema, como

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

Figura 2: Exemplo de uma instância do Problema do Corte Máximo



(a) Grafo não direcionado com arestas ponderadas

(b) Solução ótima do Corte Máximo

descrito por [Agrawal et al., 2003]. Assim, consequentemente, agrupando indivíduos com opiniões semelhantes.

2.4. Random Forest

Conforme Breiman [2001], o *Random Forest* é um modelo de aprendizado supervisionado composto por uma coleção de classificadores em estrutura de árvores, onde a predição final é obtida pelo voto unitário de cada árvore a respeito de um registro. Cada árvore é construída a partir de amostragem com reposição (bootstrap) da base de treinamento e de seleção aleatória de atributos em cada nó, buscando o melhor ponto de divisão e promovendo tanto diversidade nas amostras quanto nos critérios de divisão. Dentre suas aplicações destacam-se: suporte na prevenção de doenças cardíacas [Pal e Parija, 2021], avaliação de risco de desastres de inundações [Zhu e Zhang, 2022] e na predição de vulnerabilidade social em aplicações web [Koushik e Rama, 2024].

Além da sua capacidade de predição, ele também fornece medidas de relevância global das variáveis, calculadas a partir da média da redução de impureza (MDI) em todas as divisões das árvores, podendo ser consideradas como um subproduto de baixo custo que ocorre na fase de treinamento. Portanto, é possível avaliar a contribuição de cada atributo em relação ao desempenho do modelo [Sutera et al., 2021].

3. Metodologia

Este tópico está estruturado da seguinte forma: a Seção 3.1 apresenta a coleta de 196 questionários; a Seção 3.2 detalha limpeza, codificação, engenharia e normalização dos dados até sua conversão em uma instância do problema Max-Cut; a Seção 3.3 expõe a definição de k, aplicação de *k-means*, Max-Cut e análise de importância via *Random Forest*; e a Seção 3.4 lista as bibliotecas Python utilizadas em Jupyter Notebook.

A Figura 3 apresenta um fluxograma simplificado com as principais etapas da metodologia adotada neste estudo. O diagrama destaca o fluxo entre a coleta de dados, o pré-processamento, a aplicação dos algoritmos de agrupamento (*k-means e Max-Cut*) e a análise das variáveis por meio do *Random Forest*. As etapas serão detalhadas nas subseções seguintes.

3.1. Coleta dos dados

A coleta de dados foi realizada no Campus da Universidade Federal do Ceará em Russas, por meio de um questionário online disponibilizado via *Google Forms*. Para incentivar a participação dos estudantes, ofereceu-se 1 crédito, equivalente a uma refeição, no Restaurante Universitário (RU) para àqueles estudantes que respondessem ao questionário completo. A divulgação do questionário ocorreu por meio de grupos de aplicativo de troca de mensagens instantâneas, postagens em redes sociais e envio de e-mails para todos os estudantes por meio do Sistema Integrado de Gestão de Atividades Acadêmicas - SIGAA. Ao todo, foram obtidas 196 respostas e destas, 179

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

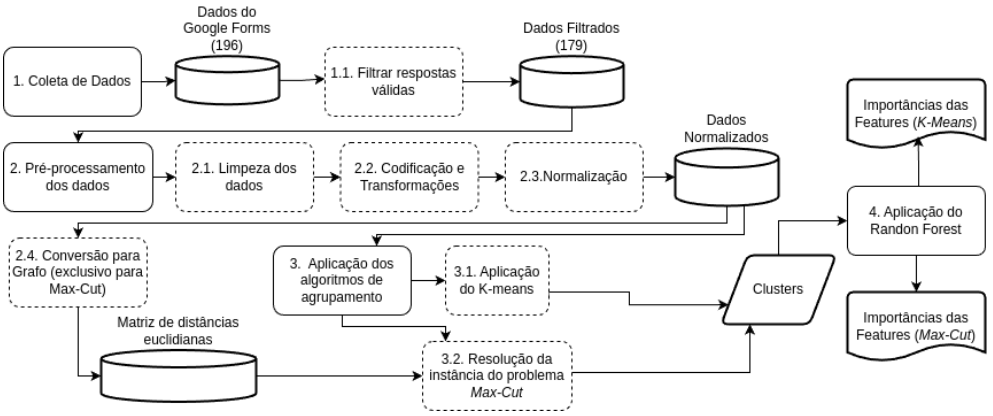


Figura 3: Fluxograma simplificado da metodologia.

foram respostas válidas, abrangendo perfis diversos de alunos de graduação e pós-graduação, o que conferiu representatividade às diferentes turmas e períodos de ingresso.

O instrumento de coleta combinou questões demográficas e acadêmicas, como curso, semestre de ingresso, idade e gênero, com perguntas específicas sobre uso dos espaços comuns (biblioteca, sala de estudos, laboratórios, RU) e padrão de interação entre discentes de cursos distintos. Além disso, foram incluídos itens direcionados ao fenômeno das "bolhas sociais", indagando diretamente se o aluno percebia divisão entre grupos, sua participação em atividades intercurso e o grau de convivência com colegas de outras áreas. Essa estrutura permitiu capturar tanto variáveis objetivas (por exemplo, frequência de uso de transporte público e de espaços de convivência) quanto percepções subjetivas sobre isolamento e integração no ambiente acadêmico. As perguntas utilizadas no questionário podem ser acessadas neste link: `https://shre.ink/eZ4D`.

3.2. Pré-processamento

Inicialmente, foi realizada a deduplicação dos registros, realizando a remoção de respostas que continham duplicidade de matrícula e preservando apenas uma resposta por discente. Em seguida, constatou-se a ocorrência de valores ausentes em uma única coluna categórica, em três registros distintos; nesses casos, optou-se por rotular tais células como "desconhecido", preservando a consistência sem introduzir viés numérico. As colunas irrelevantes para a análise, como número de matrícula, cartão do RU e campos de *feedback* para pesquisas futuras, foram então descartadas, reduzindo o conjunto de atributos aos estritamente necessários. Quanto ao tratamento das variáveis, distinguiram-se dois tipos de colunas:

- **Uma alternativa:** aquelas de resposta única (por exemplo, "Você interage com discentes de cursos diferentes do seu?"), submetidas a *Label Encoding* para converter cada categoria em chave inteira;
- **Múltiplas alternativas:** as de múltipla escolha (por exemplo, "Quais locais do Campus você costuma frequentar com mais frequência fora do horário de aula?"), cujas opções viraram variáveis dummy independentes, sinalizando presença (1) ou ausência (0) de cada item.

Em seguida, procedeu-se ao agrupamento de colunas semanticamente próximas, unindo, por exemplo, "livros acadêmicos/científicos" e "livros técnicos" via operador lógico OR; e condensando modalidades de esporte em categorias como "individual" ou "coletivo", bem como agrupando locais de uso (biblioteca, laboratório) em "ambientes acadêmicos" e opções de lazer (RU, cantina)

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

em "alimentação". Após esse processamento, a base de dados totalizou 79 colunas, possuindo informações distintas entre si. Adicionalmente, todos os atributos foram normalizados por escalonamento, garantindo que diferenças de magnitude não influenciassem indevidamente os algoritmos de *k-means*, *Random Forest* e a modelagem do grafo para aplicação do corte máximo.

Por fim, para aplicar o agrupamento através da resolução de uma instância do Problema do Corte Máximo, a base escalonada é convertida em um grafo completo, em que cada aresta representa a distância euclidiana entre dois registros, enquanto os vértices representa o registro em si. Conforme descrito na Seção 2.3, a seleção do subconjunto de vértices S que maximiza o valor do corte no grafo funciona como um agrupador natural: vértices mais distantes tendem a ficar em *clusters* diferentes, enquanto vértices mais próximos acabam agrupados juntos.

3.3. Análise da base de dados

A identificação das "bolhas sociais" partiu da determinação prévia do número ideal de grupos na base de dados. Para isso, aplicaram-se os métodos do cotovelo e da silhueta sobre as variáveis que indicam frequência de interação e uso de espaços comuns. Com o número k de bolhas definido, foi executado o agrupamento através do *k-means*, resultando em uma nova coluna indica a bolha de cada registro.

A partir do grafo construído, o peso das arestas com base na distância euclidiana, obteve-se uma instância do problema do corte máximo. Essa instância foi resolvida com o *solver* Biq Mac, que emprega o método exato *Branch & Bound* aliado à Programação Semi-Definida (SDP) Rendl et al. [2010]. O corte resultante define dois subconjuntos: S, representando uma bolha social, e seu complementar, representando outra. Essa divisão foi incorporada à base original por meio de uma nova coluna, indicando a bolha associada a cada registro.

Para explorar os atributos que mais influenciam a formação dessas bolhas, calculou-se a correlação entre a coluna-alvo (rótula da bolhas) e as demais variáveis, destacando aquelas de maior associação estatística. Além disso, treinou-se um modelo de *Random Forest* para prever o *cluster* de cada aluno conforme cada abordagem (*k-means* e Max-Cut). A importância das variáveis, obtida a partir do ganho de informação em cada nó de decisão, permitiu ordenar os fatores que mais contribuíram para a segmentação social. Caso a acurácia do modelo ultrapasse 90%, considerou-se que a predição é confiável e que as variáveis mais relevantes refletem efetivamente os fatores determinantes das bolhas no campus.

Entre as abordagens adotadas, o uso conjunto de *k-means* e já a resolução da base de dados como uma instância do Max-Cut justifica-se por oferecer perspectivas complementares sobre a segmentação social. O *k-means* proporciona uma análise baseada na similaridade de atributos dos discentes, agrupando-os de acordo com padrões de interação e perfil. Já o agrupamento através do Max-Cut atua na estrutura do grafo criado a parte da base de dados escalonada, identificando divisões que maximizam as desconexões entre grupos, revelando potenciais zonas de tensão ou isolamento. A combinação dessas abordagens permite não apenas analisar a coesão interna dos grupos, mas também compreender as fronteiras sociais que caracterizam as bolhas no ambiente universitário.

3.4. Ferramentas utilizadas

Para a análise das bolhas sociais, utilizou-se Python 3 em ambiente Jupyter Notebook. A biblioteca *pandas* foi empregada no carregamento e manipulação dos dados, enquanto as bibliotecas *seaborn* e *matplotlib.pyplot* apoiaram a visualização gráfica. No pré-processamento, aplicaram-se *StandardScaler* para normalização, *LabelEncoder* para variáveis de resposta única e *MultiLabelBinarizer* para colunas de múltipla escolha. O agrupamento exploratório utilizou o *k*-means, responsável também por gerar o gráfico do cotovelo, enquanto a análise da silhueta foi feita com o *silhouette_score*. Para modelar o grafo de distâncias, calcularam-se pares de distâncias euclidianas

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

com *pairwise_distances*. Por fim, na fase de classificação, empregou-se o *RandomForestClassifier* com divisão de dados via *train_test_split* do *scikit-learn*.

4. Resultados

Esta seção apresenta os resultados obtidos com a aplicação dos algoritmos de agrupamento *k-means* e Max-Cut, acompanhados da análise de importância das variáveis por meio do classificador *Random Forest*. O objetivo foi investigar a formação de bolhas sociais entre estudantes da UFC – Campus Russas, com base em dados coletados por questionário.

A fim de determinar a quantidade ideal de bolhas sociais nesta base de dados, aplicaram-se as técnicas do cotovelo e da silhueta, ambas indicando a formação de dois grupos distintos. A Figura 4 ilustra o resultado. Observa-se que, embora o gráfico do cotovelo não apresente uma inflexão acentuada, o método da silhueta evidencia com clareza a existência de dois *clusters*. Assim, adota-se $k = 2$ como a segmentação mais adequada.

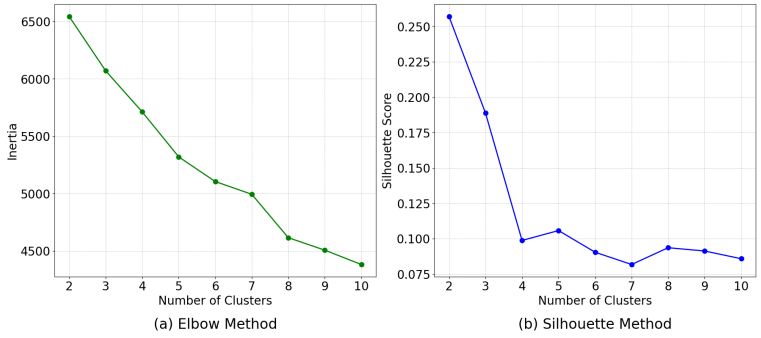


Figura 4: Quantidade de bolhas indicado pelos métodos do cotovelo e silhueta.

O fato da análises do cotovelo e da silhueta indicarem a existência de dois grupos bem definidos implica que, dentro do contexto estudado, há duas bolhas sociais principais no campus. Isso sugere que os discentes tendem a se agrupar em dois grandes grupos com características ou comportamentos semelhantes, seja em termos de uso dos espaços, interação entre cursos ou percepção de integração.

4.1. Agrupamento pelo *k-means*

Ao analisar as variáveis com maior correlação com a coluna de cluster criado pelo *k-means*, notou-se que a formação das bolhas sociais foram majoritariamente relacionadas às preferências culturais e esportivas, bem como ao uso de espaços acadêmicos. A variável "Artes Visuais" apresentou a maior correlação no valor de (0.31), indicando que a afinidade ou envolvimento com atividades ligadas às artes visuais está fortemente associada à segmentação entre os grupos. Em seguida, destacam-se "Esportes individuais" (0.28), "Música Brasileira e Popular" (0.28) e "Esportes de expressão corporal" (0.27), sugerindo que o tipo de prática esportiva e o gosto musical exercem papel relevante na diferenciação social entre os discentes. Entretanto, é válido citar que essas correlação vão de -1 a 1, indicando que essas colunas possuem um papel leve a moderado na formação das bolhas.

Já com base na análise de importância das variáveis utilizando o modelo de *Random Forest*, observou-se que as características mais determinantes para a classificação das bolhas sociais estão relacionadas à frequência e diversidade de interações interpessoais no ambiente universitário. Para a avaliação do modelo, a base de dados foi dividida em 70% para treinamento e 30% para teste, resultando em um desempenho robusto, com acurácia de 92.59%, *F1-score* de 0.92 para a classe 0 e *F1-score* de 0.93 para a classe 1.

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

Conforme ilustrado na Figura 5, a variável com maior importância apontada pelo modelo foi “Com que frequência você interage com colegas de curso fora das aulas?”, indicando que a socialização extraclasse é um fator relevante para a segmentação dos grupos. Em seguida, a variável “Com que frequência você interage com pessoas de outros cursos?” reforça que a expansão das redes sociais para além do próprio curso também contribui para essa segmentação.

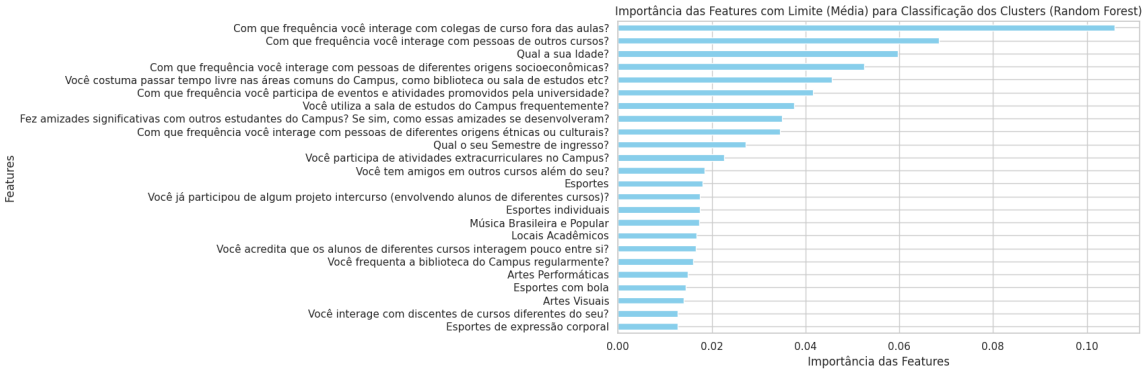


Figura 5: Importância de cada feature para classificação com *k*-means.

Além disso, questões como “Qual a sua idade?” e “Com que frequência você interage com pessoas de diferentes origens socioeconômicas?” indicam que fatores demográficos e a diversidade social possuem papel significativo na formação dessas bolhas.

Outros pontos importantes dizem respeito ao uso dos espaços comuns do campus, à participação em eventos universitários e à frequência de utilização de ambientes acadêmicos, sugerindo que o nível de envolvimento com a vida universitária também influencia na estruturação das bolhas. Nota-se, portanto, que elas não são formadas apenas com base nos cursos de origem, mas também a partir de hábitos de convivência, abertura à diversidade e engajamento nas atividades do campus. Portanto, o gráfico evidencia que, quanto maior o nível de interação e a participação em atividades extracurriculares, menor é a probabilidade de isolamento em uma única bolha social.

4.2. Agrupamento pelo Max-Cut

Ao analisar a correlação da coluna que indica as bolhas feitas ao obter o valor de corte máximo, notou-se que, diferentemente do método do *k-means*, as interações interpessoais no ambiente universitário são os principais fatores associados à formação das bolhas sociais. A variável “Você costuma passar tempo livre nas áreas comuns do Campus, como biblioteca ou sala de estudos etc?” apresentou a maior importância (0.41), seguido de “Você utiliza a sala de estudos do Campus frequentemente?” (0.39), sinalizando que o uso compartilhado de espaços acadêmicos está fortemente relacionado à segmentação entre os grupos. Em seguida, destacam-se , “Você tem amigos em outros cursos além do seu?” (0.38) e “Fez amizades significativas com outros estudantes do Campus?” (0.37), evidenciando que a frequência em ambientes comuns e a ampliação das redes de amizade para além do próprio curso contribuem diretamente na configuração das bolhas sociais. Esses valores, possuem uma correlação significativa, com relevância moderada.

A Figura 6 apresenta as importâncias das variáveis para a classificação das bolhas obtidas a partir da resolução do grafo como uma instância do problema do Max-Cut. Diferentemente do *k-means*, que parte de uma estrutura tabular, essa abordagem considerou a construção de um grafo com base na distância euclidiana entre os registros, buscando particionar os vértices de forma a maximizar a desconexão (valor do corte) entre os grupos. Em outras palavras, quanto maior for essa desconexão entre os grupos, mais distinto serão os elementos de uma bolha em relação a outra.

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

O conjunto S e seu complementar, gerados pelo corte máximo, foram então interpretados como bolhas sociais.

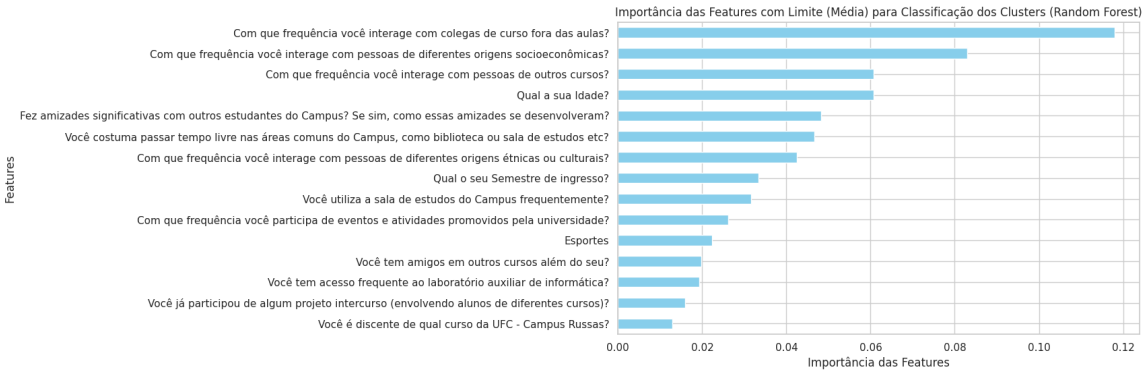


Figura 6: Importância de cada feature para classificação com *M*ax-Cut.

Nesse contexto, o modelo de *Random Forest* foi utilizado para avaliar quais variáveis mais contribuíam para distinguir os grupos formados, performando com 92.59%, com F1-score de 0.92 para a classe 0 e 0.93 para a classe 1, diferenciando apenas o *Recall* e *Precison* se comparado com as bolhas do *k-means*. Observa-se que a variável mais relevante foi novamente "Com que frequência você interage com colegas de curso fora das aulas?", indicando que a socialização extraclasse segue sendo um aspecto central na formação das bolhas, mesmo quando analisado a partir da estrutura topológica da rede. Em segundo lugar, destaca-se "Com que frequência você interage com pessoas de diferentes origens socioeconômicas?", que esteve em quarto lugar pelo método *k-means*, sugerindo que a diversidade de interações sociais exerce papel de destaque na desconexão entre os grupos.

Outras variáveis importantes indicado pelo modelo foram "Com que frequência você interage com pessoas de outros cursos?" e "Qual a sua idade?", o que reforça a ideia de que tanto o alcance das redes sociais quanto fatores demográficos influenciam na segmentação observada. Curiosamente, aspectos ligados ao uso de espaços acadêmicos, como biblioteca e sala de estudos, surgem novamente, apontando que o envolvimento com a vida universitária tem impacto na topologia das interações.

Um aspecto interessante a ser destacado é a alta similaridade entre os agrupamentos gerados pelos dois métodos. Ao comparar as classificações obtidas, observou-se uma taxa de correspondência de 92.7%, indicando que, apesar das abordagens distintas, os grupos formados apresentam grande convergência. No entanto, é importante ressaltar que os métodos partem de pressupostos conceituais diferentes: enquanto o *k-means* realiza a segmentação com base em características tabulares, agrupando registros por proximidade em um espaço vetorial, o Max-Cut atua sobre a estrutura de um grafo, buscando maximizar a dissimilaridade entre os grupos por meio do corte de conexões.

Em termos de interpretação, o *k-means* tende a identificar padrões de afinidade entre os discentes, agrupando-os por semelhanças em preferências e comportamentos. Por outro lado, o Max-Cut revela desconexões estruturais nas interações, evidenciando divisões mais latentes entre os grupos. Assim, embora os agrupamentos resultantes sejam semelhantes, cada método oferece uma perspectiva complementar.

5. Conclusão

Este trabalho investigou a formação de bolhas sociais entre estudantes de graduação e pós-graduação do Campus da Universidade Federal do Ceará em Russas, por meio da aplicação de algoritmos de agrupamento *k-means* e a resolução da instância do problema do *Max-Cut* sobre

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

os dados coletados via questionário. A análise foi complementada pelo uso do classificador *Random Forest*, que permitiu identificar as variáveis mais influentes na segmentação dos estudantes em grupos sociais distintos.

Os resultados obtidos revelaram a presença de dois agrupamentos principais, apontando para a existência de bolhas sociais bem definidas no campus. As principais variáveis associadas à formação desses grupos envolveram padrões de interação interpessoal fora das aulas, afinidades culturais e práticas esportivas. Tanto a abordagem pelo *k-means* quanto a pelo *Max-Cut* apresentaram desempenho semelhante em termos de acurácia (92.59%), mas forneceram perspectivas distintas sobre a estrutura das relações sociais: enquanto o *k-means* agrupou com base na semelhança global de atributos, o *Max-Cut* destacou rupturas e distanciamentos entre os estudantes.

A análise das importâncias das variáveis evidenciou que a frequência de interação com colegas de curso e de outros cursos, além da participação em atividades extracurriculares, são fatores cruciais para a compreensão da segmentação social no ambiente universitário. Esses achados reforçam a necessidade de políticas institucionais que promovam maior integração entre cursos e a valorização de espaços comuns como instrumentos de convivência.

Como trabalhos futuros, planeja-se expandir a base de dados para outros campi da UFC, analisando a generalização dos padrões de bolhas sociais. Propõe-se o uso de Rede Adversária Generativa para gerar registros sintéticos de bolhas específicas, simulando como atributos individuais (ex: hábitos culturais, interações) influenciam a pertença a um grupo, buscando estratégias que podem explicar fatores críticos de segregação. Complementarmente, planeja-se aplicar métodos como o Problema da Diversidade Máxima, visando selecionar subconjuntos de pessoas que são as mais diferentes entre si, revelando a estrutura das bolhas sociais e facilitando ações mais inclusivas e abrangentes.

Referências

Agrawal, R., Rajagopalan, S., Srikant, R., e Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, p. 529–535.

Bishop, C. M. e Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Bond, M. e Bergdahl, N. (2023). *Student Engagement in Open, Distance, and Digital Education*, p. 1309–1324. ISBN 978-981-19-2079-0.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Garey, M. R. e Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in the Mathematical Sciences. W. H. Freeman, first edition edition. ISBN 0716710455,9780716710455.

Goemans, M. X. e Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6): 1115–1145.

Gurusamy, V. (2017). Mining the attitude of social network users using k-means clustering. *International Journal of Advance Research in Computer Science and Software Engineering*, 7:226–230.

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br

Jonathan L. Gross, P. Z., Jay Yellen (2013). *Handbook of Graph Theory, Second Edition*. Discrete Mathematics and Its Applications. Chapman and Hall/CRC, 2 edition. ISBN 1439880182,9781439880180.

Karp, R. M. (1972). Reducibility among combinatorial problems. complexity of computer computations (the ibm research symposia series).

Ketchen, D. J. e Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458.

Koushik, P. e Rama, A. (2024). Comparison of novel reinforcement learning with random forest algorithm to improve prediction rate of social vulnerability in web application. In *AIP Conference Proceedings*, volume 2853. AIP Publishing.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, p. 281–297. Oakland, CA, USA.

Maia, M. M., Andrade, L. H. F. d., e Fernandes, S. (2022). K-means na análise de características socioeconômicas de candidatos ao ensino superior. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, 2(5). URL `https://periodicos.ufersa.edu.br/ecop/article/view/11168`.

Martins, D., Lucena, L., Damasceno, A., Siqueira, H., Tadano, Y., e Caldas, I. (2021). Clusterização do perfil de adolescentes escolares com predisposição ao uso de substância psicoativas. *Research, Society and Development*, 10:e37510212528.

Oliveira, P., Rodrigues, R., Ramos, J., e Silva, J. (2020). Uma análise de algoritmos de clusterização para descoberta de perfis de engajamento. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, p. 1012–1021, Porto Alegre, RS, Brasil. SBC. URL `https://sol.sbc.org.br/index.php/sbie/article/view/12857`.

Otterbach, J. S., Manenti, R., Alidoust, N., Bestwick, A., Block, M., Bloom, B., Caldwell, S., Didier, N., Fried, E. S., Hong, S., et al. (2017). Unsupervised machine learning on a hybrid quantum computer. *arXiv preprint arXiv:1712.05771*.

Pal, M. e Parija, S. (2021). Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series*, volume 1817, p. 012009. IOP Publishing.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Rendl, F., Rinaldi, G., e Wiegele, A. (2010). Solving Max-Cut to optimality by intersecting semidefinite and polyhedral relaxations. *Math. Programming*, 121(2):307.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sutera, A., Louppe, G., Huynh-Thu, V. A., Wehenkel, L., e Geurts, P. (2021). From global to local mdi variable importances for random forests and when they are shapley values. *Advances in Neural Information Processing Systems*, 34:3533–3543.

Zhu, Z. e Zhang, Y. (2022). Flood disaster risk assessment based on random forest algorithm. *Neural Computing and Applications*, p. 1–13.

DOI: 10.59254/sbpo-2025-212178
https://proceedings.science/p/212178?lang=pt-br