UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
CURSO DE ENGENHARIA DE PRODUÇÃO

Diego Aparecido da Silva

Avaliação de índices de validação interna em agrupamentos: uma abordagem experimental com dados sintéticos

Juiz de Fora
2025

Diego Aparecido da Silva

Avaliação de índices de validação interna em agrupamentos: uma abordagem experimental com dados sintéticos

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Engenheiro de Produção.

Orientador: Prof. Dr. Antônio Ângelo Missiaggia Picorone

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Diego Aparecido.
Avaliação de índices de validação interna em agrupamentos: uma abordagem experimental com dados sintéticos / Diego Aparecido da Silva. – 2025.
41 f. : il.

Orientador: Antônio Ângelo Missiaggia Picorone
Trabalho de Conclusão de Curso – Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Curso de Engenharia de Produção, 2025.

1. Palavra-chave. 2. Palavra-chave. 3. Palavra-chave. I. Sobrenome, Nome do orientador, orient. II. Título.

Diego Aparecido da Silva

Avaliação de índices de validação interna em agrupamentos: uma abordagem experimental com dados sintéticos

Trabalho de Conclusão de Curso apresentado a Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Engenheiro de Produção.

Aprovada em 15 de agosto de 2025

BANCA EXAMINADORA

Prof. Dr. Antonio Angelo Missiaggia Picorone - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Fernando Marques de Almeida Nogueira
Universidade Federal de Juiz de Fora

Prof. Dr. Raphael Fortes Marcomini
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Esta jornada exigiu tempo, esforço e superação. Ao chegar ao fim dessa etapa, carrego comigo um profundo sentimento de gratidão por todos que, de alguma forma, contribuíram para que este momento se tornasse possível.

Agradeço aos Orixás que regem meu caminho e aos guias espirituais que me acompanham com sabedoria e firmeza. Foram sua presença constante, seus conselhos silenciosos e sua força invisível que me sustentaram nos momentos de incerteza. Que Oxalá continue iluminando minha jornada com paz, equilíbrio e discernimento.

Ao meu orientador, Prof. Dr. Antônio Angelo Missiaggia Picorone, pela orientação técnica, paciência e disponibilidade, sempre contribuindo com críticas construtivas e direcionamentos que enriqueceram este trabalho.

Aos meus familiares, pelo amor incondicional, incentivo constante e compreensão nos momentos de ausência e dedicação.

E a todos os amigos que, de forma direta ou indireta, estiveram presentes nesta caminhada, oferecendo apoio, palavras de incentivo e parceria nos momentos mais difíceis.

A cada um de vocês, o meu sincero muito obrigado. Axé!

RESUMO

A validação da qualidade de agrupamentos é um desafio recorrente na análise de dados, especialmente em contextos de aprendizado não supervisionado, em que não há rótulos de referência. Nesse cenário, os índices de validação interna se apresentam como alternativas para apoiar a definição do número adequado de agrupamentos. Diante disso, o presente trabalho aborda a avaliação de índices internos aplicados ao algoritmo *K-means*, com o objetivo geral de analisar sua capacidade em identificar corretamente o número de grupos em diferentes cenários estruturais. Os objetivos específicos são: replicar o estudo de Liu et al. (2010), ampliando o número de índices de 11 para 42; aplicar os índices a cinco conjuntos de dados sintéticos com propriedades controladas; e comparar o desempenho das métricas em diferentes contextos de agrupamento. Como metodologia, foram gerados dados sintéticos representando situações de ruído, densidades variadas, subestruturas internas e formas assimétricas. O algoritmo *K-means* foi aplicado a esses conjuntos, e os resultados evidenciaram que apenas dois índices (*gdi33* e *gdi43*) apresentaram desempenho consistente em todos os cenários, alcançando 100% de acerto na identificação do número correto de agrupamentos, enquanto os demais mostraram limitações diante de características específicas dos dados. Os resultados indicam que não há um único índice capaz de se destacar em todos os cenários, evidenciando a importância de selecionar métricas de acordo com as características da base de dados. Recomenda-se a aplicação da metodologia a dados reais e a comparação com outros algoritmos de agrupamento como trabalhos futuros.

Palavras-chave: agrupamento de dados; validação interna; algoritmo *K-means*; métricas de agrupamento; dados sintéticos.

ABSTRACT

The validation of clustering quality is a recurring challenge in data analysis, especially in unsupervised learning contexts where no reference labels are available. In this scenario, internal validation indices emerge as alternatives to support the definition of the appropriate number of clusters. Therefore, this study addresses the evaluation of internal indices applied to the *K-means* algorithm, with the general objective of analyzing their ability to correctly identify the number of groups in different structural scenarios. The specific objectives are: to replicate the study of Liu et al. (2010), expanding the number of indices from 11 to 42; to apply the indices to five synthetic datasets with controlled properties; and to compare the performance of the metrics in different clustering contexts. As methodology, synthetic datasets were generated to represent situations of noise, varying densities, internal substructures, and asymmetric shapes. The *K-means* algorithm was applied to these datasets, and the results showed that only two indices (*gdi33* and *gdi43*) achieved consistent performance across all scenarios, reaching 100% accuracy in identifying the correct number of clusters, while the others showed limitations when facing specific data characteristics. The results indicate that there is no single index capable of standing out in all scenarios, highlighting the importance of selecting metrics according to the characteristics of the dataset. It is recommended that the methodology be applied to real datasets and compared with other clustering algorithms in future studies.

Keywords: data clustering; internal validation; *K-means* algorithm; clustering metrics; synthetic data.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O agrupamento de dados, também conhecido como *clusterização* (*clustering*), é uma técnica de análise exploratória que busca dividir um conjunto de dados em grupos ou *clusters*, de modo que os elementos de um mesmo grupo sejam mais semelhantes entre si do que em relação aos de outros grupos. Essa abordagem é amplamente utilizada em contextos de aprendizado de máquina não supervisionado, em que não há rótulos de referência disponíveis para guiar o processo de classificação.

Do ponto de vista matemático, a clusterização pode ser formulada como um problema de otimização. No caso do algoritmo *K-means*, busca-se minimizar a soma das distâncias quadráticas entre os pontos e os centróides dos clusters. Essa função objetivo é representada por:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \left\| x_i - c_j \right\|^2 \tag{1.1}$$

em que n representa o número total de pontos, k o número de clusters, x_i o vetor de atributos do ponto i, c_j o centróide do j-ésimo cluster e z_{ij} uma variável binária que assume valor 1 quando o ponto x_i pertence ao cluster j e 0 caso contrário. A minimização dessa função tem como objetivo encontrar partições que maximizem a homogeneidade interna dos grupos e a separação entre eles.

O agrupamento de dados tem sido utilizado há décadas no processamento de imagens e no reconhecimento de padrões, e consolidou-se como uma ferramenta aplicada em diversas áreas, como segmentação de mercado, diagnóstico de sistemas, detecção de fraudes e análise exploratória, por possibilitar a extração de informações relevantes a partir de dados brutos (NAQA; MURPHY, 2015). Para além dos domínios de negócios e tecnologia, o agrupamento também encontra aplicações na área da saúde, especialmente na organização de imagens médicas — como radiografias, tomografias e ressonâncias magnéticas — com o objetivo de identificar padrões que auxiliem na detecção de tumores e no suporte a diagnósticos (SHUKLA; SHARMA, 2020).

Diversos algoritmos têm sido propostos na literatura com o objetivo de realizar o agrupamento de dados. Entre os mais conhecidos, destaca-se o *K-means*, que particiona os dados em um número pré-definido de agrupamentos, e os métodos hierárquicos, que constroem uma estrutura hierárquica de grupos (SAXENA et al., 2017). Há também algoritmos baseados em densidade, como o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), capazes de identificar agrupamentos com diferentes formas e densidades, além de distinguir ruídos presentes nos dados. Outro exemplo é o algoritmo *Mean Shift*, que localiza regiões de alta densidade no espaço amostral por meio da atualização iterativa de centros de massa (HU et al., 2021).

Embora os métodos de agrupamento sejam consolidados em diversas aplicações, ainda persistem desafios relevantes quanto à sua utilização, como a definição do número ideal de agrupamentos, a escolha do algoritmo mais adequado para cada tipo de dado e a interpretação dos resultados gerados. Conjuntos de dados muito extensos podem demandar alto custo computacional, enquanto bases com elevada dimensionalidade dificultam a análise e a visualização das partições. Além disso, a sensibilidade aos parâmetros do modelo representa uma limitação importante, uma vez que pequenas variações podem impactar significativamente os resultados obtidos.

A escolha do algoritmo de agrupamento influencia diretamente os resultados obtidos, uma vez que diferentes métodos podem produzir partições distintas a partir do mesmo conjunto de dados. Isso ocorre porque cada técnica adota pressupostos próprios sobre o que caracteriza um agrupamento, o que impacta sua capacidade de identificar estruturas relevantes nos dados e as propriedades dos grupos formados (HENNIG, 2021).

Além da escolha do algoritmo, a variação e a sensibilidade aos parâmetros resultam em diferentes partições mesmo para um mesmo conjunto de dados, o que torna a comparação entre métodos um desafio. A avaliação da qualidade dos agrupamentos constitui uma etapa crítica, uma vez que não existe uma métrica única capaz de mensurar, de forma abrangente, a estrutura dos dados. Por esse motivo, recomenda-se a combinação de diferentes índices de validação para se obter uma análise mais precisa (PALACIO-NIñO; BERZAL, 2019).

As métricas de validação de agrupamento são ferramentas fundamentais para avaliar a qualidade das partições geradas por diferentes algoritmos e possibilitam a comparação entre seus desempenhos sobre um mesmo conjunto de dados. Como exemplo, Scaldelai, Santos e Matioli (2022) propuseram o índice de Densidade de Agrupamento (índice CD), uma métrica de validação interna que considera a relação entre coesão e separação dos grupos formados, apresentando desempenho comparável ao índice de *Davies-Bouldin* (DB) e ao coeficiente *Silhouette*.

No trabalho desenvolvido por Kumar et al. (2020), foi proposto um novo esquema de validação de agrupamento, denominado *Two-Phase Cluster Validation* (TPCV), que avalia a qualidade das partições com base na probabilidade de proximidade e separação entre grupos. Os experimentos indicaram que o TPCV é uma abordagem eficaz para analisar agrupamentos em conjuntos de dados estruturados e não estruturados, sem depender de informações prévias sobre as categorias formadas.

Liu et al. (2010) realizaram uma análise sistemática sobre medidas de validação interna aplicadas a algoritmos de agrupamento, com o objetivo de avaliar o desempenho de 11 índices amplamente utilizados na literatura. O estudo investigou a capacidade dessas métricas em lidar com diferentes características dos conjuntos de dados, incluindo variações em monotonicidade, presença de ruído, densidade desigual entre agrupamentos, existência de subagrupamentos e distribuições assimétricas..

Este trabalho tem como objetivo replicar o trabalho desenvolvido por Liu et al. (2010), ampliando o número de índices utilizados para a avaliação dos resultados através de um conjunto de dados gerado sinteticamente, mas utilizando as mesmas propriedades abordadas pelos autores.

1.1 Justificativa

A inexistência de um guia sistemático para a seleção de índices de validação interna dificulta a aplicação segura da análise de agrupamentos em contextos práticos. Essa lacuna é especialmente crítica na Engenharia de Produção, onde a definição do número de grupos pode impactar diretamente decisões estratégicas, como o projeto de arranjos físicos, a segmentação de clientes ou o controle estatístico de processos. Em muitos desses casos, a adoção de agrupamentos inadequados, influenciada pela escolha incorreta do índice, pode comprometer a eficiência operacional e a qualidade das decisões tomadas.

Com o interesse em entender e expandir o guia de seleção de índices desenvolvido por Liu et al. (2010), este estudo foi realizado ampliando o número de índices internos, oferecendo uma perspectiva mais ampla sobre o desempenho e a aplicabilidade dessas métricas em diferentes cenários.

A replicação e a ampliação de pesquisas previamente desenvolvidas são fundamentais para a validação e o progresso do conhecimento científico (BROWN et al., 2016). No contexto da análise de agrupamentos, esse tipo de abordagem permite verificar a consistência dos resultados e explorar novos contextos, contribuindo para uma compreensão mais aprofundada das limitações e potencialidades das métricas avaliadas.

Além de contribuir para o avanço acadêmico, este trabalho busca gerar um conjunto de evidências empíricas que sirva como apoio à construção de um guia de seleção de índices de validação interna. Essa contribuição é relevante para profissionais e pesquisadores da Engenharia de Produção que utilizam técnicas de agrupamento em problemas reais, oferecendo subsídios para decisões mais fundamentadas e eficazes no uso de algoritmos não supervisionados.

1.2 Escopo do trabalho

Este estudo tem como objetivo avaliar a qualidade de agrupamentos gerados pelo algoritmo *K-means*, com base em índices de validação interna. A investigação baseia-se na metodologia proposta por Liu et al. (2010), a qual emprega conjuntos de dados sintéticos construídos com propriedades estruturais controladas. Preservando as mesmas configurações de dados utilizadas no estudo original, esta pesquisa amplia o escopo da análise ao aplicar um número significativamente maior de índices de validação interna. Com isso, busca-se oferecer uma avaliação mais abrangente da capacidade dessas métricas

em identificar o número adequado de agrupamentos em diferentes contextos estruturais.

Delimitações do Trabalho:

1. **Contexto controlado**: O estudo foi conduzido com dados sintéticos, que não capturam totalmente ruídos, outliers ou distribuições complexas de bases reais, limitando a generalização dos resultados.
2. **Redundância potencial**: Com a ampliação do número de índices internos avaliados, algumas métricas podem apresentar comportamentos semelhantes, gerando análises parcialmente redundantes.
3. **Investigação parcial de falha**: Não foram explorados todos os motivos pelos quais certos índices falham em identificar o número correto de agrupamentos, focando apenas na taxa de acerto.
4. **Escopo metodológico**:
   - Estabilidade e sensibilidade: Não foram avaliados critérios como a consistência dos índices em repetições (estabilidade) ou sua resposta a variações no parâmetro k (sensibilidade).
   - Análise superficial dos algoritmos: O estudo utilizou as implementações padrão dos índices disponíveis no pacote R, sem modificações ou análise detalhada de seus algoritmos internos, conforme sua natureza exploratória.
5. **Dados sintéticos**: As bases foram geradas artificialmente devido à indisponibilidade da base original de Liu et al. (2010), o que pode influenciar a robustez das conclusões.

1.3 Objetivos do trabalho

Este trabalho tem como objetivo geral analisar o desempenho de diversos índices de agrupamento interno propostos na literatura usados para avaliar resultados dos algoritmos de agrupamentos.

Os objetivos específicos deste trabalho são os seguintes:

- Gerar um banco de dados sintéticos, com propriedades controladas, para servir de base para análise dos índices de agrupamento interno;
- Analisar o desempenho de índices de validação interna em diferentes contextos de agrupamento, utilizando conjuntos de dados sintéticos com características variadas.
- Analisar as vantagens e limitações da aplicação de um número maior de índices para validar o resultado de algoritmos de agrupamentos.

1.4 Metodologia da pesquisa

Este trabalho é caracterizado como uma pesquisa experimental, cujo propósito permeia as características descritivas, onde se busca definir com mais acurácia as melhores métricas de avaliação de agrupamentos, e as características explanatórias, quando se examina a relação de causa e efeito entre os diversos cenários gerados e os resultados das métricas de validação de agrupamentos.

Foi utilizada como base para a elaboração das pesquisas o estudo de Liu et al. (2010), em que os autores analisaram 11 índices de validação de agrupamentos. A Figura 1 apresenta as etapas do macrofluxo realizadas para o atingimento dos objetivos estabelecidos na seção 1.3.

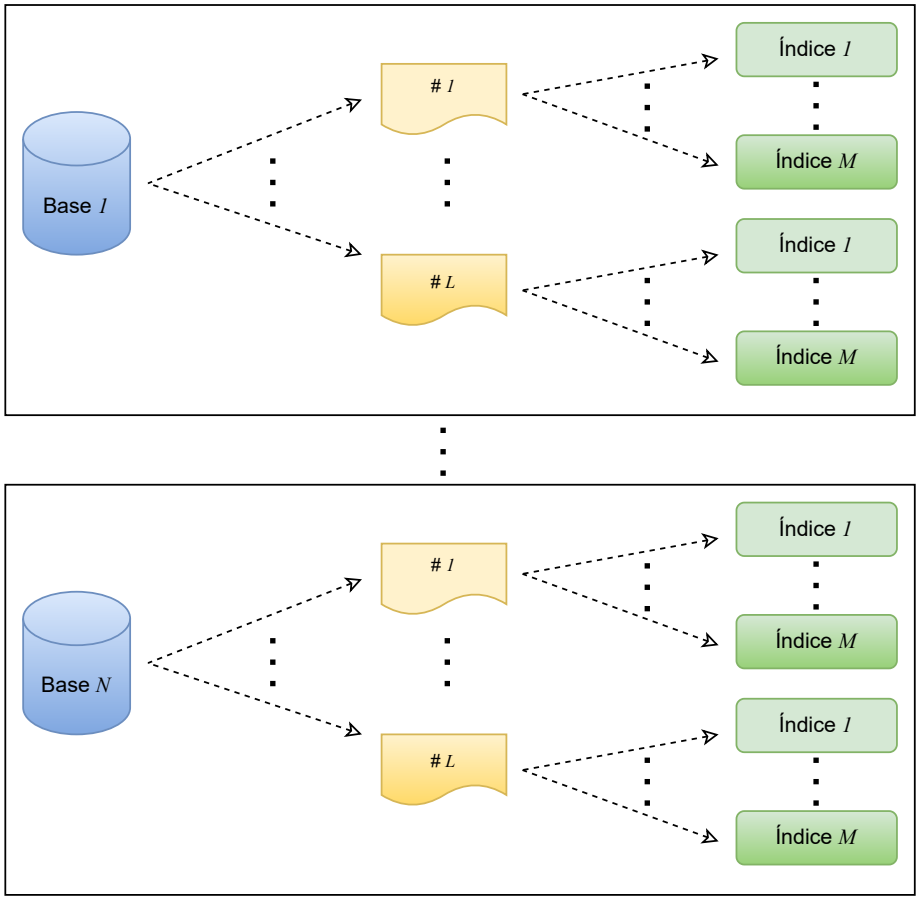


Figura 1 – Metodologia de pesquisa.
Fonte: Elaborado pelo autor (2025).

De forma geral, a metodologia adotada consiste na geração de um conjunto de dados sintéticos, na aplicação de um algoritmo de agrupamento e na posterior avaliação de índices de validação interna com base nos resultados obtidos.

Foram criados N conjuntos de dados sintéticos, com características semelhantes ao trabalho de referência (LIU et al., 2010) denominados neste trabalho como bases, nos quais variaram-se atributos como a dispersão dos dados e a posição dos centróides, garantindo

diferenciação nas características dos agrupamentos. Em todas as N bases, a quantidade de elementos foi mantida constante, assegurando uniformidade na comparação dos resultados. Uma característica importante nas bases de dados é que os agrupamentos gerados são visualmente separáveis.

Posteriormente, as N bases foram submetidas ao algoritmo *K-means* para a realização dos agrupamentos. Com o objetivo de ampliar o escopo da análise, o parâmetro k, correspondente ao número de agrupamentos, foi variado, resultando em um total de L execuções do *K-means* para cada uma das N bases.

Após a realização dos agrupamentos, aplicaram-se os M índices de validação interna selecionados a cada uma das L partições geradas. Esse procedimento permite avaliar a capacidade dos índices em refletir a qualidade dos agrupamentos e em identificar, com precisão, a quantidade ideal de grupos. Para fins de comparação, considerou-se como número ideal de agrupamentos aquele que pode ser visualmente identificado em cada base.

1.5 Estrutura do trabalho

O presente trabalho está estruturado em cinco capítulos.

O Capítulo 1 apresenta a introdução ao estudo, incluindo a justificativa para sua realização, o escopo do trabalho, a formulação dos objetivos e a metodologia empregada. Além disso, descreve a estrutura geral do documento.

O Capítulo 2 trata da revisão de literatura, abordando os principais conceitos sobre análise de agrupamento, os algoritmos utilizados, com ênfase no K-means, e os critérios de validação de agrupamentos, com destaque para os índices internos. Também são discutidos desafios relacionados à validação e avaliação da qualidade dos agrupamentos.

O Capítulo 3 é dedicado ao desenvolvimento da pesquisa. Nele, são apresentados os índices de validação interna de agrupamentos, a metodologia de geração dos conjuntos de dados sintéticos, a escolha do método de agrupamento e a implementação do algoritmo com as medidas de validação. Além disso, inclui uma análise exploratória dos dados e dos resultados obtidos.

O Capítulo 4 apresenta e analisa os resultados do estudo, detalhando os erros encontrados por conjunto de dados e os resultados obtidos pelos diferentes índices internos de validação aplicados.

Por fim, o Capítulo 5 traz as conclusões do trabalho, destacando as principais descobertas, as limitações identificadas e possíveis direções para pesquisas futuras.

2 Revisão de literatura

Ao longo deste referencial teórico, são apresentados os principais fundamentos relacionados à análise de agrupamentos de dados, com foco em sua aplicação em contextos de aprendizado não supervisionado. Inicialmente, são abordados os diferentes tipos de algoritmos de agrupamento. Em seguida, é detalhado o funcionamento do algoritmo *K-means*, amplamente utilizado devido à sua simplicidade e eficiência computacional, além de ser o método adotado no desenvolvimento deste trabalho. Na sequência, são discutidos os critérios de validação de agrupamentos, com ênfase nos critérios internos, que possibilitam avaliar a qualidade das partições geradas com base exclusivamente na estrutura dos dados. Por fim, são explorados os principais desafios associados ao uso desses critérios.

2.1 Algoritmos de agrupamento

Os algoritmos de agrupamento particionam objetos de dados (padrões, entidades, instâncias, observâncias, unidades) em um certo número de *clusters* (grupos, subconjuntos ou categorias) e devido à dificuldade humana de interpretação aliada ao aumento da quantidade de dados, o desenvolvimento de técnicas de agrupamento tem se tornado cada vez mais importante (HAN MICHELINE KAMBER, 2006).

Cada algoritmo de agrupamento apresenta suas técnicas específicas e estratégias para agrupar os padrões identificados nos conjuntos de dados e usualmente são classificados como hierárquicos, particionais, baseados em densidade ou grade (JAIN; MURTY; FLYNN, 1999).

Xu e Tian (2015) apresentam um resumo mais amplo dos algoritmos comumente utilizados e suas respectivas categorias, que estão apresentados na tabela 1, e também apresentam em seu trabalho algoritmos mais modernos.

Tabela 1 – Relação dos algoritmos tradicionais

Categoria de Algoritmo de Agrupamento	**Algoritmos Típicos**
Baseado em partição	K-means, K-medoids, PAM, CLARA, CLARANS
Baseado em hierarquia	BIRCH, CURE, ROCK, Chameleon
Baseado em teoria fuzzy	FCM, FCS, MM
Baseado em distribuição	DBCLASD, GMM
Baseado em densidade	DBSCAN, OPTICS, Mean-Shift
Baseado em teoria de grafo	CLICK, MST
Baseado em grade	STING, CLIQUE
Baseado em teoria fractal	FC
Baseado em modelo	COBWEB, GMM, SOM, ART

Fonte: Adaptado de Xu e Tian (2015)

Na Engenharia de Produção, por exemplo, os algoritmos de agrupamento podem ser aplicados ao projeto de arranjos físicos modulares, com o objetivo de reduzir a movimentação de materiais e otimizar o fluxo de produção. Argoud, Filho e Tiberti (2008) propuseram a utilização de um algoritmo genético de agrupamento (AGA) para formar módulos de arranjo físico a partir de subsequências de operações comuns entre diferentes peças. O AGA permite que o número de módulos seja definido previamente ou determinado automaticamente, além de oferecer ao usuário flexibilidade na escolha da codificação dos cromossomos, medidas de similaridade e operadores genéticos. Os resultados experimentais demonstraram que o AGA gerou arranjos físicos mais eficientes do que abordagens anteriores, com destaque para a redução de até 18,28% na distância total percorrida pelas peças, evidenciando a eficácia do método no contexto da manufatura.

Deve-se ressaltar que a escolha do algoritmo é fundamental para o sucesso do agrupamento, pois dependendo do tipo de dados e da aplicação, cada método terá um desempenho diferente.

2.2 O algoritmo K-means

O *K-means* é um dos algoritmos de agrupamento mais utilizados em aprendizado de máquina e mineração de dados. Sua popularidade se deve à simplicidade e eficiência computacional, sendo amplamente empregado em aplicações que exigem a divisão de um conjunto de dados em grupos homogêneos (AHMED; SERAJ; ISLAM, 2020). O *K-means* é um método particional não supervisionado que busca minimizar a variância intra-cluster e maximizar a separação entre clusters.

O *K-means* é um algoritmo iterativo que segue quatro etapas principais:

1. **Inicialização**: Escolha de K centróides iniciais, seja de forma aleatória ou utilizando heurísticas como *K-means++*, que melhora a escolha inicial para evitar convergência para mínimos locais (COATES; NG, 2012).

2. **Atribuição de clusters**: Cada ponto x_j é atribuído ao cluster do centróide C_i mais próximo, com base na distância Euclidiana:

$$x_j \in C_i \text{ se } \|x_j - m_i\| \leq \|x_j - m_l\|, \quad \forall l \neq i, \tag{2.1}$$

   em que m_i representa a média dos N_i elementos atribuídos ao i-ésimo cluster.

3. **Atualização dos centróides**: Os centróides são recalculados como a média dos pontos atribuídos a cada cluster:

$$m_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j \tag{2.2}$$

4. **Critério de parada**: O processo é repetido até que os centróides não apresentem variações significativas ou até atingir um número máximo de iterações (CHONG, 2021).

Como principais vantagens do algoritmo, têm-se sua simplicidade de compreensão e implementação permitindo sua ampla utilização em diversos campos (CHONG, 2021). Para além disso, o *K-means* possui baixos custos computacionais para execução, o que garante uma eficiência em termos de execução, se comparado a outros métodos de agrupamento e pode ser construído para ter escalabilidade, permitindo sua aplicação a grandes conjuntos de dados (COATES; NG, 2012).

Apesar de apresentar tais vantagens, o *K-means* também possui suas limitações. Autores como Yuan e Yang (2019), Ahmed, Seraj e Islam (2020) e Jie et al. (2020) observam que o algoritmo é sensível a inicialização, ou seja, os grupos encontrados dependem fortemente da escolha inicial dos centróides, ocasionando convergências para diferentes agrupamentos dependendo dessa inicialização, o que pode levar a resultados subótimos. Constatam também que o algoritmo é insuficiente ao trabalhar com diferentes tipos de dados, limitando a aplicação para conjunto de dados apenas numéricos.

Para superar essas limitações, diversas melhorias foram propostas para o algoritmo, incluindo o *K-means++*, que aprimora a seleção dos centróides iniciais (COATES; NG, 2012); o *Kernel K-means*, que permite a identificação de agrupamentos não esféricos (AHMED; SERAJ; ISLAM, 2020); e o *Incremental K-means*, que acelera o processo de convergência (NGUYEN, 2020). Essas variações ampliaram a aplicabilidade do *K-means* e o tornaram mais adequado a conjuntos de dados com diferentes características estruturais.

O algoritmo *K-means* tem sido adotado em diversas áreas devido à sua simplicidade e eficiência computacional. Em visão computacional, por exemplo, é empregado para a extração de características e reconhecimento de padrões em imagens (COATES; NG, 2012). Na bioinformática, sua aplicação inclui a análise de sequências genéticas e a identificação de estruturas similares (AHMED; SERAJ; ISLAM, 2020). Já no contexto de Big Data, é frequentemente utilizado em tarefas de segmentação de clientes, permitindo uma categorização mais eficaz de perfis de consumo (JIE et al., 2020). Essa adaptabilidade metodológica torna o *K-means* uma técnica recorrente em estudos voltados à análise exploratória de dados.

Na Engenharia de Produção, o algoritmo *K-means* também tem sido empregado como ferramenta para controle e monitoramento de desempenho em ambientes industriais com alta variabilidade operacional. Kęsek (2020) propôs uma abordagem baseada no agrupamento de ciclos produtivos a partir de características extraídas diretamente dos dados do processo de ancoragem em minas subterrâneas. Os ciclos foram agrupados por similaridade utilizando o método *K-means*, permitindo a identificação de padrões de desempenho associados a diferentes grupos de condições operacionais. Essa classificação

possibilitou, por exemplo, prever o desempenho de novos ciclos com base no grupo ao qual foram atribuídos, promovendo uma forma proativa de controle da eficiência do processo produtivo. A aplicação prática foi realizada com o auxílio das linguagens R e VBA, demonstrando a viabilidade do uso do *K-means* como apoio à tomada de decisão em contextos industriais reais.

2.3 Avaliação da qualidade dos agrupamentos

Avaliar a qualidade dos agrupamentos gerados por algoritmos de agrupamento é uma etapa essencial no processo de análise, pois permite interpretar os padrões identificados nos dados e orientar a escolha do número apropriado de grupos. No entanto, como não existe uma definição universalmente aceita do que caracteriza um "bom agrupamento", também não há consenso sobre uma única métrica de avaliação que seja aplicável em todos os contextos (LEWIS; ACKERMAN; SA, 2012).

Segundo Brun et al. (2007), os critérios de validação de agrupamentos podem ser classificados em três categorias principais: internos, relativos e externos. Cada uma dessas abordagens oferece uma perspectiva distinta sobre a qualidade da partição dos dados. Enquanto os métodos internos avaliam propriedades geométricas dos grupos, os externos utilizam rótulos de referência (quando disponíveis) e os relativos comparam diferentes partições geradas sob variações de parâmetros.

Em ambientes industriais com produção em pequena escala, a limitação do tamanho amostral compromete a aplicação de ferramentas estatísticas convencionais para o controle de processo. Com o intuito de contornar esse desafio, Greipel, Nottenkämper e Schmitt (2020) investigaram diferentes algoritmos de agrupamento, visando formar grupos homogêneos de amostras a partir de dados históricos de usinagem. Para avaliar a qualidade das partições geradas, os autores utilizaram critérios de validação interna, com destaque para o índice *SDbw*, que considera simultaneamente a compacidade dos agrupamentos e a separação entre eles. Testes estatísticos complementares, como *Levene* e *Kruskal–Wallis*, foram empregados para confirmar a homogeneidade interna dos grupos. A metodologia demonstrou-se eficaz na criação de agrupamentos apropriados para uso em cartas de controle, viabilizando a aplicação do controle estatístico em contextos com restrição de amostragem.

2.4 Tipos de métricas de validação

As métricas de validação têm como objetivo quantificar a qualidade das partições geradas pelos algoritmos de agrupamento. Elas são classificadas conforme o tipo de informação que utilizam:

- **Validação interna**: considera apenas a estrutura dos dados e a configuração dos

grupos resultantes. Métricas como coesão, separação e compacidade são comuns nesse grupo (PALACIO-NIñO; BERZAL, 2019).

- **Validação relativa**: compara diferentes partições geradas por variações no algoritmo ou nos parâmetros, como diferentes valores de k ou inicializações distintas (BRUN et al., 2007).
- **Validação externa**: utiliza uma partição de referência conhecida para comparar os agrupamentos gerados. Sua aplicação é restrita a cenários com dados rotulados e sua eficácia depende diretamente da qualidade da referência utilizada (DOM, 2012).

Apesar das diferentes abordagens disponíveis, a escolha entre elas depende fortemente da natureza do problema e da disponibilidade de informações externas. Em contextos aplicados, como na Engenharia de Produção, é comum a ausência de rótulos de referência, o que torna a validação interna a estratégia mais viável. Ainda assim, a interpretação dos resultados deve considerar as limitações inerentes de cada tipo de métrica.

2.5 Métricas internas de validação

As métricas de validação interna têm como objetivo mensurar a qualidade de um agrupamento com base apenas nas propriedades internas dos dados e na configuração dos grupos gerados, sem depender de rótulos externos. Essas métricas procuram capturar o quão coerente é a partição obtida a partir de características como a proximidade entre os elementos de um mesmo grupo (coesão) e a separação entre grupos distintos (separabilidade). A avaliação interna é especialmente útil em cenários onde não há informações de referência disponíveis, como em muitos contextos de análise exploratória ou em aplicações industriais baseadas em dados operacionais.

Essas métricas podem ser agrupadas em diferentes categorias, conforme os critérios utilizados para quantificar a qualidade do agrupamento:

- **Índices baseados em distância**: consideram relações entre os elementos e os centróides de seus respectivos grupos, ou entre os próprios grupos. Exemplos incluem o índice de Davies-Bouldin e o índice de Dunn, que avaliam, respectivamente, a compacidade dos grupos e a separação entre eles (ARBELAITZ et al., 2013).
- **Índices baseados em densidade**: consideram a densidade local dos dados, sendo particularmente eficazes em conjuntos com ruído ou com agrupamentos de formas irregulares. O índice *SDbw* é um exemplo dessa categoria, combinando medidas de dispersão dentro dos grupos com estimativas de densidade entre os agrupamentos (GREIPEL; NOTTENKäMPER; SCHMITT, 2020).

- **Critérios estatísticos auxiliares**: embora não sejam métricas de validação formalmente estabelecidas, medidas como variância intra e intergrupos podem ser utilizadas como suporte à avaliação da homogeneidade dos agrupamentos, especialmente quando combinadas a testes estatísticos complementares (GREIPEL; NOTTENKäMPER; SCHMITT, 2020).

Embora existam diferentes tipos de métricas internas, todas compartilham o objetivo de oferecer uma medida objetiva da estrutura dos agrupamentos formados. A escolha do índice mais adequado depende de características específicas dos dados, como a presença de ruído, a forma dos grupos e o grau de separação entre eles. Por isso, é comum que essas métricas sejam utilizadas de forma combinada, buscando uma avaliação mais robusta da qualidade do agrupamento.

2.6 Limitações das métricas internas

Apesar de suas vantagens, os índices de validação interna apresentam limitações importantes. A primeira delas é a ausência de um índice que se destaque universalmente em todos os cenários. Estudos mostram que a performance desses índices varia significativamente conforme o formato dos dados, a presença de ruído, o grau de sobreposição entre grupos e a dimensionalidade da base (LIU et al., 2013).

Além disso, muitos desses índices assumem a existência de agrupamentos esféricos e bem separados, o que nem sempre se verifica em bases reais. Agrupamentos com formatos alongados, densidades heterogêneas ou com subestruturas internas tendem a ser penalizados, mesmo quando semanticamente coerentes (PAKGOHAR; LENGYEL; BOTTA-DUKÁT, 2024).

Outro desafio está na sensibilidade a variações nos parâmetros do algoritmo. Mudanças no número de grupos (k), na inicialização ou na ordem de processamento dos dados podem afetar significativamente a avaliação obtida por esses índices (ARBELAITZ et al., 2013).

2.7 Desafios na avaliação de agrupamentos

As métricas de validação interna são frequentemente adotadas para avaliar propriedades como coesão e separação dos agrupamentos. No entanto, apresentam limitações relevantes, como a sensibilidade a características específicas dos dados, incluindo número de grupos, densidade ou presença de ruído.

Um dos principais desafios está na inexistência de um índice universalmente eficaz. Conforme apontado por Arbelaitz et al. (2013), diferentes métricas produzem resultados inconsistentes a depender da estrutura dos dados, como sobreposição entre agrupamentos, variabilidade de densidade ou alta dimensionalidade. Liu et al. (2013) reforçam que mesmo

os índices mais utilizados podem apresentar viés quando aplicados a bases com formas ou densidades heterogêneas, comprometendo a confiabilidade da avaliação.

Outro obstáculo diz respeito à instabilidade dos índices frente à variação de parâmetros. Alterações no número de grupos ou na inicialização do algoritmo podem impactar significativamente os resultados, dificultando a replicabilidade e o uso confiável de uma métrica isolada como critério decisório (XIE et al., 2020).

Por fim, é importante considerar que a maioria das métricas internas se baseia exclusivamente em propriedades geométricas, como distâncias e densidades, desconsiderando o significado semântico dos agrupamentos. Isso pode levar à penalização de soluções visualmente consistentes ou semanticamente relevantes que não atendem a pressupostos geométricos, como esfericidade ou simetria dos grupos.

3 Desenvolvimento

Neste capítulo, descrevem-se de forma sistemática as etapas metodológicas adotadas na condução do estudo. Inicialmente, apresenta-se o processo de geração dos conjuntos de dados sintéticos, construídos com propriedades controladas para simular diferentes desafios relacionados ao agrupamento de dados. Em seguida, detalha-se a aplicação do algoritmo *K-means* sobre os dados gerados, considerando variações no número de agrupamentos, com o objetivo de analisar o comportamento dos índices de validação interna em diferentes cenários. Por fim, são apresentados os 42 índices utilizados e os critérios adotados para a interpretação dos resultados, com base em referências consolidadas na literatura.

3.1 Geração dos conjuntos de dados sintéticos

Para avaliar o desempenho dos índices de validação interna, foram gerados cinco conjuntos de dados sintéticos utilizando *Python*. Os parâmetros foram definidos de forma a refletir a diversidade estrutural das bases analisadas por Liu et al. (2010), cuja proposta metodológica serve de referência para este estudo. Como os conjuntos originais não são publicamente acessíveis, optou-se por criar bases com propriedades semelhantes, de modo a viabilizar a replicação e a ampliação das análises conduzidas pelos autores.

O código utilizado para a geração dos conjuntos de dados encontra-se disponível no Apêndice A e permite a parametrização de critérios que influenciam diretamente a estrutura e as características dos agrupamentos:

- **Centróides**: definidos de forma a posicionar os agrupamentos no espaço multidimensional;
- **Número de elementos**: ajustado para assegurar uma quantidade representativa de pontos em cada agrupamento;
- **Dispersão**: configurada para controlar a variabilidade interna dos pontos e, consequentemente, a separação entre os agrupamentos.

Cada conjunto foi projetado para representar um dos cinco aspectos do agrupamento discutidos por Liu et al. (2010), com o objetivo de avaliar o desempenho dos índices de validação interna em contextos que refletem desafios típicos da análise de agrupamentos:

- **Conjunto Bem Separado (CBS)**: Representa o aspecto da separabilidade entre agrupamentos, sendo composto por cinco grupos bem definidos e com grande distância entre os centróides. A Figura 2 ilustra essa configuração;

- **Conjunto Bem Separado com Ruído (CBSR)**: Aborda a robustez frente a *outliers*, simulando pontos ruidosos distribuídos aleatoriamente no espaço. A Figura 3 apresenta esse cenário;
- **Conjunto com Diferentes Densidades (CDD)**: Representa a heterogeneidade na densidade dos agrupamentos, com três grupos de tamanhos e dispersões distintos. Ilustrado na Figura 4;
- **Conjunto com Subagrupamentos (CSC)**: Reflete a existência de hierarquias internas, simulando subdivisões dentro de três agrupamentos principais. A Figura 5 demonstra essa estrutura;
- **Conjunto Assimétrico (CA)**: Enfoca a não esfericidade dos agrupamentos, com três grupos de formas irregulares e alongadas. A Figura 6 exibe essa configuração.

3.2 Aplicação do método K-means para agrupamento dos dados sintéticos

O agrupamento dos conjuntos de dados sintéticos foi realizado utilizando o algoritmo *K-means* no *RStudio*, em conformidade com a metodologia empregada por Liu et al. (2010). Considerando as diferenças de variabilidade entre as variáveis x e y presentes nos conjuntos de dados, optou-se pela padronização prévia dos dados por meio da função `scale()`, que os transforma em *z-scores* (média zero e desvio padrão um). Esse procedimento é fundamental porque o *K-means* utiliza a distância Euclidiana, de modo que variáveis com maior dispersão poderiam exercer influência desproporcional no processo de agrupamento. A normalização garantiu que ambas as variáveis contribuíssem de forma equilibrada para o cálculo das distâncias, resultando em partições mais consistentes e representativas.A conversão para *Z-scores* é definida por:

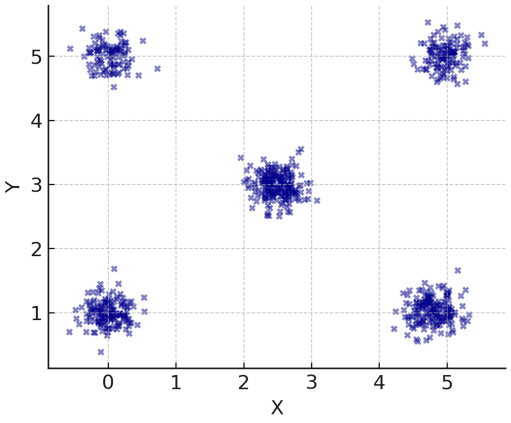
$$Z = \frac{x - \mu}{\sigma}, \tag{3.1}$$

em que x é o valor observado, μ a média da amostra e σ é o desvio padrão da amostra.

Considerando a sensibilidade do algoritmo *K-means* à escolha dos centróides iniciais, adotou-se a prática recomendada de executar o algoritmo múltiplas vezes para cada valor de k. Especificamente, utilizou-se o parâmetro `nstart = 25` na implementação em R, de modo que o agrupamento selecionado fosse aquele que apresentou a menor soma dos quadrados intragrupo dentre 25 inicializações aleatórias. Essa abordagem reduz o risco de convergência para soluções locais e assegura maior confiabilidade na avaliação dos índices de validação interna.

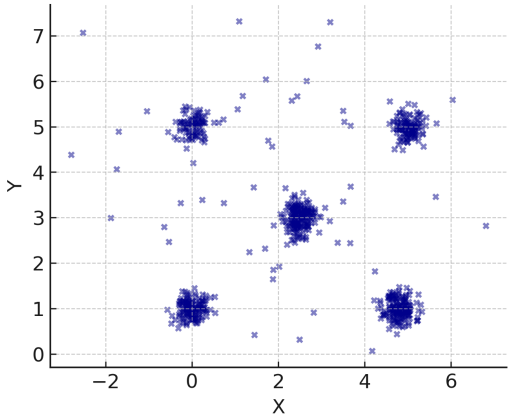
Para investigar o comportamento dos índices de validação interna, o algoritmo *K-means* foi executado considerando diferentes valores para o número de agrupamentos,

Figura 2 – Conjunto bem separado (CBS)



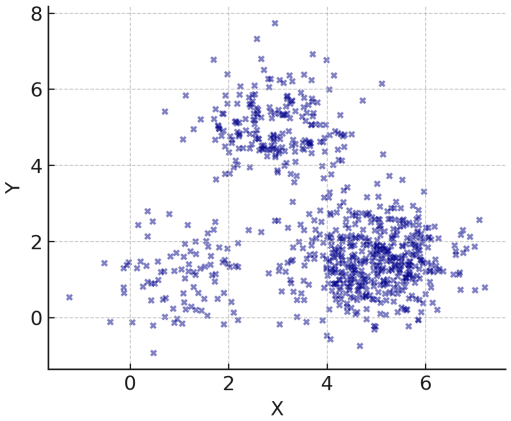
Fonte: Elaborado pelo autor (2025)

Figura 3 – Conjunto bem separado com ruído (CBSR)



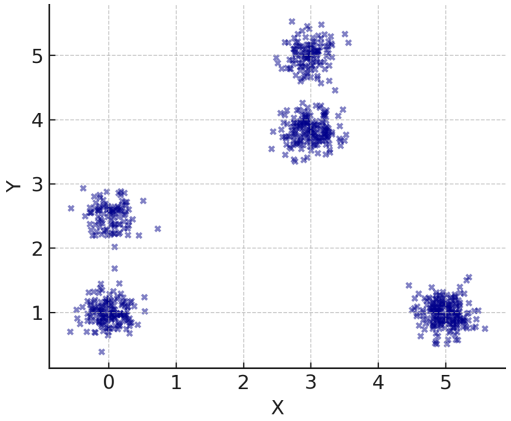
Fonte: Elaborado pelo autor (2025)

Figura 4 – Conjunto com diferentes densidades (CDD)



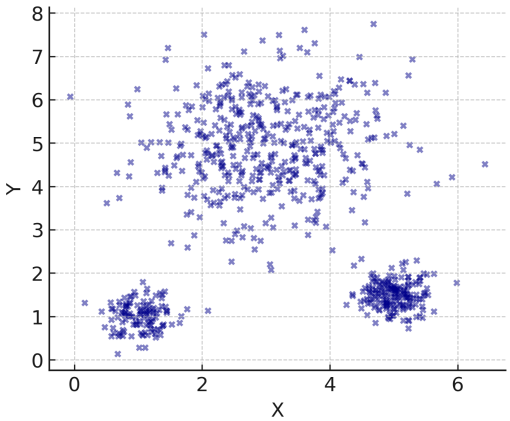
Fonte: Elaborado pelo autor (2025)

Figura 5 – Conjunto com subclusters (CSC)



Fonte: Elaborado pelo autor (2025)

Figura 6 – Conjunto assimétrico (CA)

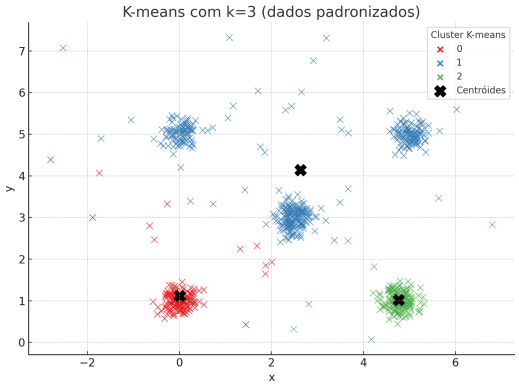


Fonte: Elaborado pelo autor (2025)

com $k = 2, 3, \ldots, 9$. Isso resultou em oito agrupamentos distintos para cada uma das cinco bases de dados sintéticos, totalizando 40 partições. O objetivo foi verificar como os índices respondem a diferentes estruturas de agrupamento. O código correspondente à implementação do algoritmo pode ser consultado no Apêndice B.

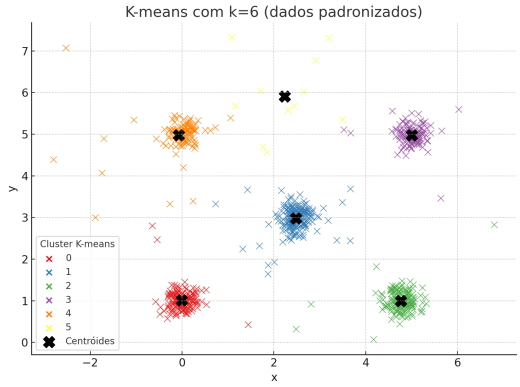
Como exemplo, as Figuras 7 e 8 apresentam os resultados da aplicação do *K-means* sobre o Conjunto Bem Separado com Ruído (CBSR), para os casos em que $k = 3$ e $k = 6$. Essas figuras ilustra como a escolha do número de agrupamentos pode influenciar diretamente na interpretação dos dados e na resposta dos índices de validação observados no próximo tópico.

Figura 7 – Clusterização conjunto CBSR com $k = 3$



Fonte: Elaborado pelo autor (2025)

Figura 8 – Clusterização conjunto CBSR com $k = 6$



Fonte: Elaborado pelo autor (2025)

3.3 Aplicação dos índices de validação interna

O trabalho apresentado por Liu et al. (2010) utilizou 11 índices internos de validação de agrupamento para avaliar a qualidade das partições geradas. No presente trabalho, ampliamos essa abordagem, aplicando um total de 42 índices, com o objetivo de aprofundar a análise exploratória e investigar a metodologia utilizada.

Os valores dos índices foram calculados para cada conjunto de dados sintético e para cada partição gerada com diferentes valores de k. Esse processo foi realizado por meio da função *intCriteria*, disponível na biblioteca *clusterCrit*, que recebe como entrada a matriz de dados original e a atribuição dos agrupamentos produzida pelo algoritmo *K-means*. O pacote *clusterCrit* é parte do trabalho apresentado por Desgraupes (2013).

Para interpretação dos resultados, foram utilizadas as regras definidas no trabalho de Desgraupes (2013) em que o melhor valor obtido para um determinado índice indica uma melhor estrutura de agrupamento. No entanto, é importante ressaltar que a direção da 'melhoria' (maior ou menor valor) varia de acordo com o índice utilizado. Uma lista contendo os 42 índices de validação interna utilizados no presente trabalho e suas regras que definem o melhor resultado estão indicadas na figura 9.

ÍNDICE EM R	REGRA	ÍNDICE EM R	REGRA
ball_hall	MÁXIMO	gdi51	MÁXIMO
banfeld_raftery	MÍNIMO	gdi52	MÁXIMO
c_index	MÍNIMO	gdi53	MÁXIMO
calinski_harabasz	MÁXIMO	ksq_detw	MÁXIMO
davies_bouldin	MÍNIMO	log_det_ratio	MÍNIMO
det_ratio	MÍNIMO	log_ss_ratio	MÍNIMO
dunn	MÁXIMO	mcclain_rao	MÍNIMO
gamma	MÁXIMO	pbm	MÁXIMO
g_plus	MÍNIMO	point_biserial	MÁXIMO
gdi11	MÁXIMO	ray_turi	MÍNIMO
gdi12	MÁXIMO	ratkowsky_lance	MÁXIMO
gdi13	MÁXIMO	scott_symons	MÍNIMO
gdi21	MÁXIMO	sd_scat	MÍNIMO
gdi22	MÁXIMO	sd_dis	MÍNIMO
gdi23	MÁXIMO	s_dbw	MÍNIMO
gdi31	MÁXIMO	silhouette	MÁXIMO
gdi32	MÁXIMO	tau	MÁXIMO
gdi33	MÁXIMO	trace_w	MÁXIMO
gdi41	MÁXIMO	trace_wib	MÁXIMO
gdi42	MÁXIMO	wemmert_gancarski	MÁXIMO
gdi43	MÁXIMO	xie_beni	MÍNIMO

Figura 9 – Índices de validação interna de agrupamento

Fonte: Adaptado de Desgraupes (2013)

Os valores obtidos para cada índice foram organizados em planilhas, permitindo a visualização das respostas fornecidas em cada conjunto de dados e em diferentes valores de k. Essa estrutura possibilitou avaliar, posteriormente, quais índices identificaram corretamente o número de agrupamentos.

Para avaliar a eficácia de cada índice na identificação do número ideal de agrupamentos, foi calculada a taxa de sucesso de cada métrica. Define-se S_i como o número de vezes em que o índice i indicou corretamente o número de agrupamentos, e W como o total de conjuntos de dados analisados. Assim, a taxa de sucesso do índice i é dada por:

$$T_i = \frac{S_i}{W} \times 100 \tag{3.2}$$

Essa métrica, expressa em percentual, indica o grau de acerto do índice na identificação do número ideal de agrupamentos. Valores mais altos de T_i refletem maior confiabilidade do índice nos diferentes cenários analisados.

4 Resultados

Este capítulo apresenta os resultados obtidos a partir da aplicação dos índices de validação interna sobre os agrupamentos gerados. As análises estão organizadas de forma a permitir a visualização do desempenho dos índices tanto individualmente por base de dados quanto de forma consolidada. Inicialmente, é avaliado o percentual de acerto dos índices em cada um dos cinco conjuntos sintéticos, possibilitando observar como diferentes estruturas de dados impactam a eficácia das métricas. Em seguida, os índices são agrupados conforme sua taxa de sucesso geral, destacando-se aqueles com maior robustez e capacidade de generalização frente aos diferentes cenários simulados.

4.1 Desempenho dos índices por base de dados

Ao todo, foram obtidos 1.680 resultados decorrentes da aplicação dos 42 índices de validação interna às 40 partições geradas pelo algoritmo *K-means*. Esses resultados foram organizados e armazenados em planilhas do Excel. A Figura 10 ilustra uma amostra dessa organização, na qual cada índice apresenta oito resultados para cada conjunto de dados, indicando se o número correto de agrupamentos foi identificado.

	A	B	C	D	E	F	G	H	I
1	Bem separado (5)		MÁX	MÍN	MÍN	MÁX	MÍN	MÍN	MÁX
2	VARIAÇÃO	K	ball_hall	banfeld_raftery	c_index	calinski_harabasz	davies_bouldin	det_ratio	dunn
3	-3	2	1,23	139,81	0,16	477,36	1,20	4,66	0,01
4	-2	3	0,39	-1162,35	0,08	776,81	0,54	10,89	0,36
5	-1	4	0,18	-1791,74	0,03	1522,81	0,48	327,80	0,46
6	0	5	0,02	-2800,38	0,00	15949,45	0,16	7830,24	1,31
7	1	6	0,02	-2900,54	0,01	14262,04	0,46	9717,16	0,01
8	2	7	0,02	-2959,77	0,01	12758,71	0,72	11430,93	0,01
9	3	8	0,02	-3032,30	0,01	12177,65	0,91	14543,74	0,01
10	4	9	0,02	-3078,62	0,01	11391,35	1,04	16021,88	0,01
11		REFERÊNCIA	1,23	-3078,62	0,00	15949,45	0,16	4,66	1,31
12		ACERTOU?	NÃO	NÃO	SIM	SIM	SIM	NÃO	SIM
13									
14									
15	Ruído (5)		MÁX	MÍN	MÍN	MÁX	MÍN	MÍN	MÁX
16	VARIAÇÃO	K	ball_hall	banfeld_raftery	c_index	calinski_harabasz	davies_bouldin	det_ratio	dunn
17	-3	2	1,22	147,05	0,18	469,90	1,20	4,50	0,01
18	-2	3	0,45	-680,70	0,10	696,44	0,58	9,21	0,03
19	-1	4	0,23	-1273,65	0,04	1252,48	0,53	91,35	0,02
20	0	5	0,08	-2004,40	0,00	4716,52	0,23	711,96	0,09
21	1	6	0,11	-2173,94	0,00	4638,86	0,44	1078,46	0,06
22	2	7	0,13	-2278,69	0,00	4477,89	0,52	1422,23	0,07
23	3	8	0,12	-2351,41	0,02	4176,57	0,65	1699,20	0,01
24	4	9	0,13	-2420,56	0,02	4057,91	0,65	2075,43	0,01
25		REFERÊNCIA	1,22	-2420,56	0,00	4716,52	0,23	4,50	0,09
26		ACERTOU?	NÃO	NÃO	NÃO	SIM	SIM	NÃO	SIM

Figura 10 – Resultados dos índices de validação

Fonte: Elaborado pelo autor (2025)

Observa-se que, ao aplicar o algoritmo *K-means* ao Conjunto Bem Separado (CBS) com $k = 5$, valor correspondente ao número correto de agrupamentos nessa base, apenas os índices *c_index*, *calinski_harabasz*, *davies_bouldin* e *dunn* indicaram corretamente a partição como válida.

Avaliando o desempenho dos índices em cada conjunto de dados utilizado, identificou-se que há uma variação significativa no percentual de acerto conforme a estrutura dos agrupamentos. A figura 11 ilustra os resultados, indicando quais bases foram mais facilmente identificadas e quais apresentaram maiores desafios para os métodos de validação interna.

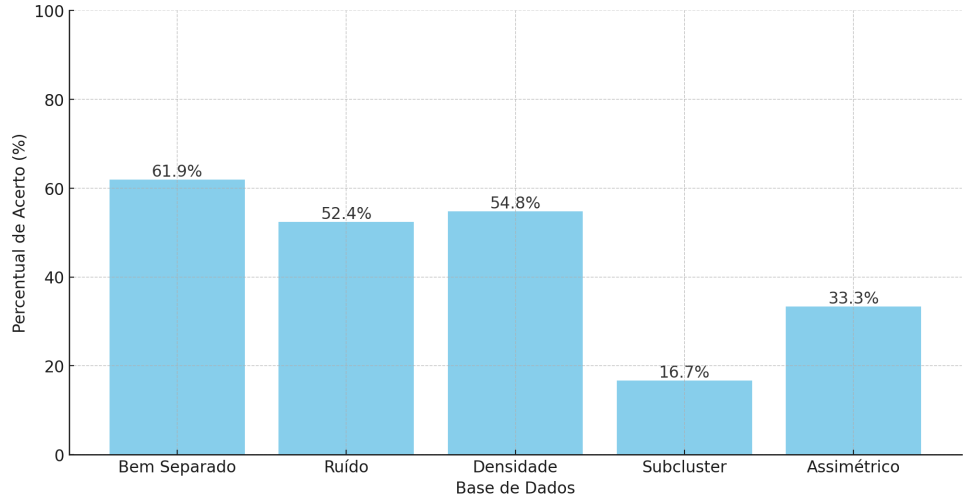


Figura 11 – Percentual de acerto dos índices por base de dados

Fonte: Elaborado pelo autor (2025)

Esse comportamento pode ser explicado pelas características estruturais das bases:

- **Bem Separado**: Esta foi a base em que os índices apresentaram melhor desempenho. A separação nítida entre os agrupamentos favorece o uso de métricas baseadas na compacidade interna dos grupos e na distância entre eles, permitindo que os índices tradicionais forneçam estimativas mais precisas do número ideal de agrupamentos.

- **Ruído**: O desempenho intermediário observado nesta base pode estar relacionado à presença de *outliers* e pontos dispersos que não pertencem a nenhum agrupamento claramente delimitado. Muitos índices de validação interna baseiam seus cálculos na distância média ou na variabilidade dentro dos agrupamentos, e a presença de ruído tende a distorcer essas medidas, comprometendo a precisão das estimativas.

- **Densidade**: O desempenho intermediário observado nesta base sugere que alguns índices de validação interna têm baixa sensibilidade à variação de densidade entre os agrupamentos. Métricas que assumem uniformidade na compacidade interna dos

grupos tendem a apresentar limitações em cenários com distribuição desigual dos dados.

- **Subcluster**: Esta foi uma das bases com pior desempenho. Uma possível explicação é que muitos índices de validação interna avaliam a qualidade dos agrupamentos com base na separação global entre os grupos, sem considerar a presença de subestruturas internas. Isso pode levar a avaliações equivocadas, uma vez que a partição gerada não reflete com precisão a complexidade da estrutura dos dados.
- **Assimétrico**: Os índices apresentaram limitações nesta base, possivelmente em razão da suposição implícita, presente em muitos métodos, de que os agrupamentos possuem forma esférica ou distribuição homogênea. No entanto, a presença de grupos com formas irregulares compromete a efetividade de métricas tradicionais baseadas em separabilidade e compacidade.

4.2 Desempenho geral dos indices

Para apoiar uma interpretação estruturada, os índices foram agrupados em três categorias conforme suas taxas de sucessos:

- **Ruim (0% a 20%)**: 15 índices. São eles: gdi51, gdi52, ratkowsky_lance, sd_scat, ball_hall, banfeld_raftery, det_ratio, ksq_detw, log_det_ratio, log_ss_ratio, mcclain_rao, point_biserial, scott_symons, trace_w, trace_wib;
- **Intermediário (40% a 60%)**: 12 índices. São eles: gdi31, gdi53, c_index, dunn, gamma, g_plus, gdi11, gdi12, gdi13, gdi21, sd_dis, xie_beni;
- **Bom (80% a 100%)**: 15 índices. São eles: gdi33, gdi43, calinski_harabasz, davies_bouldin, gdi22, gdi23, gdi32, gdi41, gdi42, pbm, ray_turi, s_dbw, silhouette, tau, wemmert_gancarski

A tabela 2 indica em quais bases os índices de validação interna tiveram sucesso.

Observa-se que os índices *gdi33* e *gdi43* foram os únicos que identificaram corretamente o número de agrupamentos em todas as bases testadas. Esse resultado pode estar relacionado a características específicas dessas variantes, conforme discutido por Bezdek e Pal (1998), que generalizaram o índice de *Dunn* com o objetivo de torná-lo mais robusto a diferentes distribuições de agrupamentos.

Esses índices pertencem à família dos *Generalized Dunn Indexes (GDI)*, desenvolvida para superar limitações do índice de *Dunn* original. Segundo Bezdek e Pal (1998), algumas variantes do GDI, incluindo aquelas utilizadas neste estudo, apresentam melhor equilíbrio entre separação e compacidade interna, proporcionando avaliações mais estáveis e consistentes da qualidade das partições.

Tabela 2 – Desempenho Consolidado dos Índices

Índice	Bem Separado	Ruído	Densidade	Subcluster	Assimétrico
gdi33	x	x	x	x	x
gdi43	x	x	x	x	x
calinski_harabasz	x	x	x	x	
silhouette	x	x	x	x	
davies_bouldin	x	x	x		x
gdi22	x	x	x	x	
gdi23	x	x	x	x	
pbm	x	x		x	x
ray_turi	x	x	x		x
s_dbw	x	x	x		x
tau	x	x	x		x
wemmert_gancarski	x	x	x		x
c_index	x		x		
dunn	x	x			
gamma	x		x		
g_plus	x		x		
gdi11	x	x			
gdi12	x	x			
gdi13	x	x			
gdi21	x		x		
sd_dis			x		x
xie_beni	x	x			
gdi31	x	x	x		
gdi53	x		x	x	
ball_hall					
banfeld_raftery					
det_ratio					
ksq_detw					
log_det_ratio					
log_ss_ratio					
mcclain_rao					
point_biserial					
scott_symons					
trace_w					
trace_wib					
gdi51			x		
gdi52			x		
ratkowsky_lance					x
sd_scat		x			

Fonte: Elaborado pelo autor (2025).

Os principais fatores que podem ter contribuído para o desempenho superior do *gdi33* e *gdi43* são:

- **Uso de novas métricas para medir separação entre agrupamentos**: Diferentemente do índice de Dunn original, que utiliza apenas a menor distância entre dois agrupamentos, os GDIs incorporam distâncias baseadas na média entre todos os pontos, o que reduz a influência de *outliers* e torna a medida mais representativa da separação entre os grupos.

- **Menor sensibilidade a variações de forma e densidade**: O índice de Dunn original assume agrupamentos compactos e bem separados, sendo pouco eficaz em conjuntos com diferentes densidades ou presença de subestruturas internas. Os GDIs superam essa limitação ao utilizar medidas de dispersão baseadas na média dos pontos, o que contribui para seu bom desempenho nas bases Densidade e Subcluster.

- **Melhor adaptação a agrupamentos assimétricos**: A presença de formas irregulares nos dados pode comprometer o desempenho de métricas clássicas que assumem agrupamentos esféricos. Os índices *gdi33* e *gdi43*, por utilizarem médias ponderadas no cálculo da dispersão, demonstram maior robustez a essa limitação. Esse fator contribui para os bons resultados observados na base Assimétrica.

- **Menor impacto do ruído nos cálculos**: Bezdek e Pal (1998) destacam que índices baseados exclusivamente em distâncias mínimas entre agrupamentos são mais suscetíveis a distorções na presença de *outliers*. Como os índices *gdi33* e *gdi43* utilizam médias das distâncias intergrupos, apresentam menor sensibilidade a pontos dispersos, o que favorece seu desempenho na base Ruído.

Essas características tornam os índices *gdi33* e *gdi43* mais robustos e generalizáveis em comparação com outras variantes do GDI, bem como com índices clássicos como *Davies-Bouldin* e *Silhouette*, que tiveram bom desempenho, mas não foram consistentes em todas as bases.

O índice *S_Dbw* foi apontado por Liu et al. (2010) como a métrica de validação interna com melhor desempenho geral, em razão de sua capacidade de capturar múltiplos aspectos da estrutura dos agrupamentos. No entanto, no presente estudo, esse índice não foi capaz de identificar corretamente o número de agrupamentos no conjunto Subcluster. Esse resultado indica que sua eficácia pode ser sensível às características específicas da base de dados, o que desaconselha seu uso isolado como critério único de avaliação.

5 Conclusão

Este estudo teve como objetivo ampliar a avaliação dos índices de validação interna de agrupamento, explorando o comportamento de 42 métricas distintas aplicadas a conjuntos de dados sintéticos com diferentes características estruturais. A partir do agrupamento realizado por meio do algoritmo *K-means*, foi possível analisar o desempenho de cada índice quanto à sua capacidade de indicar o número correto de agrupamentos. Os resultados indicaram que os índices *gdi33* e *gdi43* foram os únicos a alcançar 100% de acerto em todos os conjuntos testados, evidenciando consistência em diferentes cenários e sugerindo sua relevância como critérios de referência em estudos futuros sobre validação interna.

Por outro lado, observou-se que grande parte dos índices analisados apresentou variações significativas de desempenho entre os diferentes conjuntos de dados. Essas variações expõem uma limitação recorrente nos critérios tradicionais de validação, que frequentemente apresentam baixa sensibilidade a estruturas complexas, como subagrupamentos, diferentes densidades ou distribuições assimétricas.

Apesar da consistência observada em alguns índices, como o *gdi33* e o *gdi43*, este estudo apresenta limitações que devem ser consideradas. A principal refere-se ao uso exclusivo de dados sintéticos. Embora tenham sido projetados para simular cenários variados de agrupamento, esses conjuntos não refletem integralmente a complexidade e os ruídos presentes em bases reais, o que pode impactar o desempenho dos índices em contextos aplicados. Outra limitação está na escolha do algoritmo de agrupamento: o *K-means*, apesar de ser uma técnica consolidada, possui restrições quanto à forma dos agrupamentos e à sensibilidade a *outliers*. Além disso, este trabalho avaliou os índices com base apenas na taxa de acerto do número de agrupamentos, sem considerar aspectos como a estabilidade das soluções ou a sensibilidade a variações no parâmetro k.

Diante dessas limitações, estudos futuros podem aplicar a metodologia proposta a conjuntos de dados reais, a fim de verificar se os resultados obtidos se mantêm consistentes em contextos mais complexos e menos controlados. A incorporação de outros algoritmos de agrupamento, especialmente aqueles que não impõem suposições de esfericidade ou que sejam menos sensíveis à presença de ruído, pode enriquecer a análise comparativa entre os índices. Embora tenha sido utilizado o parâmetro `nstart = 25` para mitigar os efeitos da aleatoriedade na inicialização do algoritmo *K-means*, reconhece-se que esse número pode ser insuficiente em cenários de maior dificuldade. Investigações posteriores podem adotar um número maior de repetições com o objetivo de reduzir o risco de convergência para soluções de menor qualidade. Essa precaução pode contribuir para que os índices de validação reflitam de forma mais precisa a qualidade da partição analisada. Por fim, recomenda-se a avaliação da sensibilidade dos índices a variações nos dados e nos parâmetros do modelo,

bem como a experimentação de abordagens que combinem múltiplas métricas de avaliação, ampliando as possibilidades de diagnóstico da qualidade dos agrupamentos gerados.

REFERÊNCIAS

AHMED, Mohiuddin; SERAJ, Raihan; ISLAM, Syed Mohammed Shamsul. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, 2020.

ARBELAITZ, Olatz; GURRUTXAGA, Ibai; MUGUERZA, Javier; PÉREZ, Jesús M.; PERONA, Iñigo. An extensive comparative study of cluster validity indices. **Pattern Recognit.**, v. 46, p. 243–256, 2013. Disponível em: <https://api.semanticscholar.org/CorpusID:12473012>.

ARGOUD, Ana Rita Tiradentes Terra; FILHO, Eduardo Vila Gonçalves; TIBERTI, Alexandre José. Algoritmo genético de agrupamento para formação de módulos de arranjo físico. **Gestão & Produção**, SciELO Brasil, v. 15, p. 393–405, 2008.

BEZDEK, James C.; PAL, Nikhil R. Some new indexes of cluster validity. **IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics**, IEEE, v. 28, n. 3, p. 301–315, 1998.

BROWN, Sue; DENNIS, A.; SAMUEL, Binny M.; TAN, Barney; VALACICH, J.; WHITLEY, Edgar A. Replication research: Opportunities, experiences and challenges. 2016.

BRUN, Marcel; SIMA, Chao; HUA, Jianping; LOWEY, James; CARROLL, Brent; SUH, Edward; DOUGHERTY, Edward R. Model-based evaluation of clustering validation measures. **Pattern Recognit.**, v. 40, p. 807–824, 2007. Disponível em: <https://api.semanticscholar.org/CorpusID:9108372>.

CHONG, Bao. K-means clustering algorithm: a brief review. **Academic Journal of Computing & Information Science**, 2021.

COATES, Adam; NG, A. Learning feature representations with k-means. **Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700**, p. 561–580, 2012.

DESGRAUPES, Bernard. Clustering indices. **University of Paris Ouest-Lab Modal'X**, Paris, France:, v. 1, n. 1, p. 34, 2013.

DOM, Byron E. An information-theoretic external cluster-validity measure. **arXiv preprint arXiv:1301.0565**, 2012.

GREIPEL, J.; NOTTENKäMPER, Gina; SCHMITT, R. Comparison of grouping algorithms to increase the sample size for statistical process control. **SN Applied Sciences**, v. 2, p. 1–20, 2020.

HAN MICHELINE KAMBER, Jian Pei Jiawei. **Data Mining: Concepts and Techniques**. 2. ed. Morgan Kaufmann, 2006. (The Morgan Kaufmann Series in Data Management Systems). ISBN 1558609016; 9781558609013. Disponível em: <libgen.li/file.php?md5=d62bcd0f33a976303b1415fba99c9bc6>.

HENNIG, C. An empirical comparison and characterisation of nine popular clustering methods. **Advances in Data Analysis and Classification**, v. 16, p. 201 – 229, 2021.

HU, Zhengbing; LURIE, Irina; TYSHCHENKO, Oleksii K; SAVINA, Nataliia; LYTVYNENKO, Volodymyr. Comparative analysis of inductive density clustering algorithms meanshift and dbscan. In: SPRINGER. **Advances in Artificial Systems for Power Engineering**. [S.l.], 2021. p. 232–242.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 31, n. 3, p. 264–323, sep 1999. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/331499.331504>.

JIE, Chen; JIYUE, Zhang; JUNHUI, Wu; YUSHENG, Wu; HUIPING, Si; KAIYAN, Lin. Review on the research of k-means clustering algorithm in big data. **2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)**, p. 107–111, 2020.

KUMAR, S. S.; AHMED, Syed Thouheed; VIGNESHWARAN, P.; SANDEEP, H.; SINGH, H. M. Retracted article: Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets. **Journal of Ambient Intelligence and Humanized Computing**, v. 12, p. 7581 – 7594, 2020.

KęSEK, M. The k-means grouping method as a mean to control the performance of the production process. **Immunotechnology**, v. 1, 2020.

LEWIS, Joshua; ACKERMAN, Margareta; SA, Virginia de. Human cluster evaluation and formal quality measures: A comparative study. In: **Proceedings of the Annual Meeting of the Cognitive Science Society**. [S.l.: s.n.], 2012. v. 34, n. 34.

LIU, Yanchi; LI, Zhongmou; XIONG, Hui; GAO, Xuedong; WU, Junjie. Understanding of internal clustering validation measures. In: IEEE. **2010 IEEE international conference on data mining**. [S.l.], 2010. p. 911–916.

LIU, Yanchi; LI, Zhongmou; XIONG, Hui; GAO, Xuedong; WU, Junjie; WU, Sen. Understanding and enhancement of internal clustering validation measures. **IEEE Transactions on Cybernetics**, v. 43, p. 982–994, 2013.

NAQA, Issam El; MURPHY, Martin J. What is machine learning? In: _____. **Machine Learning in Radiation Oncology: Theory and Applications**. Cham: Springer International Publishing, 2015. p. 3–11. ISBN 978-3-319-18305-3. Disponível em: <https://doi.org/10.1007/978-3-319-18305-3_1>.

NGUYEN, Tien-Dung. **Improving The Performance Of The K-means Algorithm**. 2020. Disponível em: <https://arxiv.org/abs/2005.04689>.

PAKGOHAR, Naghmeh; LENGYEL, Attila; BOTTA-DUKÁT, Zoltán. Quantitative evaluation of internal cluster validation indices using binary data sets. **Journal of Vegetation Science**, 2024. Disponível em: <https://api.semanticscholar.org/CorpusID:273340623>.

PALACIO-NIñO, Julio-Omar; BERZAL, Fernando. **Evaluation Metrics for Unsupervised Learning Algorithms**. 2019. Disponível em: <https://arxiv.org/abs/1905.05667>.

SAXENA, Amit; PRASAD, Mukesh; GUPTA, Akshansh; BHARILL, Neha; PATEL, Om Prakash; TIWARI, Aruna; ER, Meng Joo; DING, Weiping; LIN, Chin-Teng. A review of clustering techniques and developments. **Neurocomputing**, v. 267, p. 664–681, 2017.

ISSN 0925-2312. Disponível em:
<https://www.sciencedirect.com/science/article/pii/S0925231217311815>.

SCALDELAI, Dirceu; SANTOS, Solange R dos; MATIOLI, Luiz C. Índice de densidade da clusterização: Uma nova métrica para validação interna de agrupamentos. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 9, n. 1, 2022.

SHUKLA, Moksh; SHARMA, Krishna Kumar. A comparative study to detect tumor in brain mri images using clustering algorithms. **2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)**, p. 773–777, 2020.

XIE, Jiang; XIONG, Zhongyang; DAI, Qi-Zhu; WANG, Xiao-Xia; ZHANG, Yu-Fang. A new internal index based on density core for clustering validation. **Inf. Sci.**, v. 506, p. 346–365, 2020.

XU, Dongkuan; TIAN, Yingjie. A comprehensive survey of clustering algorithms. **Annals of data science**, Springer, v. 2, p. 165–193, 2015.

YUAN, Chunhui; YANG, Haitao. Research on k-value selection method of k-means clustering algorithm. **J**, 2019.

APÊNDICE A – Código para geração dos conjuntos de dados

```python
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import metrics
from sklearn.cluster import KMeans

# Define semente para reprodutibilidade
np.random.seed(2021)

# Parametros de configuracao
numAtributos = 3
centroides = [[1, 1], [2, 2], [3, 3], [0, 3.5]] # Centro de cada grupo
quantElementos = [100, 80, 70, 65]    # Numero de elementos em cada grupo
dispersao = [0.2, 0.2, 0.2, 0.2]      # Dispersao de cada grupo
rotulo = [0, 1, 2, 1]                 # Rotulo de cada grupo

numC = len(centroides)

# Gera dados sinteticos para cada grupo
for i in range(numC):
    MatrizDadosAux = np.zeros((quantElementos[i], numAtributos))

    mu_x, mu_y = centroides[i]     # Centro do grupo (x, y)
    sigma = dispersao[i]           # Dispersao (desvio padrao)

    # Gera valores aleatorios com distribuicao normal
    aux_x = np.random.normal(mu_x, sigma, quantElementos[i])
    aux_y = np.random.normal(mu_y, sigma, quantElementos[i])

    # Atribui valores as colunas
    MatrizDadosAux[:, 0] = aux_x
    MatrizDadosAux[:, 1] = aux_y
    MatrizDadosAux[:, 2] = rotulo[i]

    # Concatena com a matriz de dados principal
    if i == 0:
        MatrizDados = MatrizDadosAux
    else:
        MatrizDados = np.concatenate((MatrizDados, MatrizDadosAux))

# Cria um DataFrame para manipular e salvar os dados
df = pd.DataFrame(MatrizDados, columns=['x','y','grupo'])
```

Algoritimo 1 – Código para geração dos conjuntos de dados na linguagem Python. Fonte: Elaborado pelo autor

APÊNDICE B – Código para agrupamento e cálculo dos indices de validação interna

```r
# setwd("C:/GERAL/DOCUMENTOS/UFF/TCC/Bases novas 16-10-2022/Resultado Indices")
rm(list = ls())
setwd("C:/GERAL/DOCUMENTOS/UFF/TCC/Bases novas 16-06-2025")

library(readxl)
library(dplyr)
library(openxlsx)

# Le o arquivo CSV com dados iniciais
Base1 <- read.csv(
  "C:/GERAL/DOCUMENTOS/UFF/ICv2/afcp/DataFrame com novos dados (basecl e classificacoes iniciais).csv",
  sep = ";"
)

# Normaliza (padroniza) os dados
Base1.scale <- scale(Base1)

# Realizando diversas clusterizacoes com k-means para diferentes numeros de agrupamentos
resultadoBase1 <- list()

criteriosBase1cl1 <- data.frame(ncriterias = NA)
criteriosBase1cl2 <- data.frame(ncriterias = NA)
criteriosBase1cl3 <- data.frame(ncriterias = NA)
criteriosBase1cl4 <- data.frame(ncriterias = NA)
criteriosBase1cl5 <- data.frame(ncriterias = NA)
criteriosBase1cl6 <- data.frame(ncriterias = NA)

# Converte a base para matriz
Base1 <- as.matrix(Base1)

# Loop para criar k clusters (de 1 ate 6)
for (i in 1:6) {
  km <- kmeans(Base1, i)             # Aplica o k-means
  resultadoBase1[[i]] <- km          # Armazena o resultado

  # Armazena os valores de cluster (exemplo: "all1", "all2" etc. precisam existir na sua base)
  criteriosBase1cl1 <- rbind(
    criteriosBase1cl1,
    data.frame(ncriterias = km$cluster[,"all1"])
  )
  criteriosBase1cl2 <- rbind(
```

```
    criteriosBase1cl2,
    data.frame(ncriterias = km$cluster[,"all2"])
  )
  criteriosBase1cl3 <- rbind(
    criteriosBase1cl3,
    data.frame(ncriterias = km$cluster[,"all3"])
  )
  criteriosBase1cl4 <- rbind(
    criteriosBase1cl4,
    data.frame(ncriterias = km$cluster[,"all4"])
  )
  criteriosBase1cl5 <- rbind(
    criteriosBase1cl5,
    data.frame(ncriterias = km$cluster[,"all5"])
  )
  criteriosBase1cl6 <- rbind(
    criteriosBase1cl6,
    data.frame(ncriterias = km$cluster[,"all6"])
  )
}

# Combina os data.frames de criterios em um unico
dfCriterias1 <- data.frame(
  criteriosBase1cl1,
  criteriosBase1cl2,
  criteriosBase1cl3,
  criteriosBase1cl4,
  criteriosBase1cl5,
  criteriosBase1cl6
)

# Exporta os resultados para um arquivo Excel
write.xlsx(dfCriterias1, file = "Base1 - Resultados.xlsx")
```

Algoritimo 2 – Código para agrupamento e cálculo dos índices internos de validação na linguagem R. Fonte: Elaborado pelo autor

APÊNDICE C – Desempenho dos 42 índices de validação interna por conjunto de dados

Bem separado (5)		MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO
VARIAÇÃO	K	ball_hall	banfeld_ra	c_index	calinski_ha	davies_bo	det_ratio	dunn	gamma	g_plus	gdi11	gdi12	gdi13	gdi21	gdi22	gdi23	gdi31	gdi32	gdi33	gdi41	gdi42	gdi43
-3	2	1,23	139,81	0,16	477,36	1,20	4,66	0,01	0,50	0,12	0,01	0,03	0,01	1,36	6,39	1,74	0,77	3,62	0,99	0,61	2,85	0,78
-2	3	0,39	-1162,35	0,08	776,81	0,54	10,89	0,36	0,68	0,08	0,36	1,69	0,53	0,97	4,58	1,43	0,75	3,58	1,12	0,70	3,33	1,04
-1	4	0,18	-1791,74	0,03	1522,81	0,48	327,80	0,46	0,87	0,03	0,46	2,45	0,69	1,24	6,63	1,88	0,83	4,46	1,27	0,77	4,12	1,17
0	5	0,02	-2800,38	0,00	15949,45	0,16	7830,24	1,31	1,00	0,00	1,31	10,23	3,63	2,87	22,46	7,96	2,08	16,30	5,77	2,07	16,22	5,75
1	6	0,02	-2900,54	0,01	14262,04	0,46	9717,16	0,01	0,97	0,00	0,01	0,08	0,03	0,87	6,82	2,42	0,30	2,35	0,83	0,24	1,92	0,68
2	7	0,02	-2959,77	0,01	12758,71	0,72	11430,93	0,01	0,95	0,01	0,01	0,08	0,03	0,94	6,82	2,42	0,31	2,24	0,79	0,24	1,71	0,61
3	8	0,02	-3032,30	0,01	12177,65	0,91	14543,74	0,01	0,94	0,01	0,01	0,08	0,03	1,01	7,09	2,50	0,33	2,33	0,82	0,25	1,78	0,63
4	9	0,02	-3078,62	0,01	11391,35	1,04	16021,88	0,01	0,94	0,01	0,01	0,08	0,03	0,98	6,86	2,46	0,33	2,33	0,84	0,25	1,79	0,64
	REFERÊNCIA	1,23	-3078,62	0,00	15949,45	0,16	4,66	1,31	1,00	0,00	1,31	10,23	3,63	2,87	22,46	7,96	2,08	16,30	5,77	2,07	16,22	5,75
	ACERTOU?	NÃO	NÃO	SIM	SIM	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM

Bem separado (5)		MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO
VARIAÇÃO	K	gdi51	gdi52	gdi53	ksq_detw	log_det_ra	log_ss_rat	mcclain_ra	pbm	point_bise	ray_turi	ratkowsky	scott_sym	sd_scat	sd_dis	s_dbw	silhouette	tau	trace_w	trace_wib	wemmert_	xie_beni
-3	2	0,36	1,70	0,46	467712,09	1139,44	-0,44	0,55	1,10	-0,51	0,38	0,44	-1723,51	0,76	1,12	2,00	0,44	0,36	897,48	3,66	0,46	3493,09
-2	3	0,05	0,22	0,07	450809,20	1766,77	0,75	0,44	2,38	-0,63	0,15	0,48	-3406,56	0,20	0,77	0,20	0,74	0,48	475,54	5,00	0,70	0,59
-1	4	0,06	0,30	0,09	26616,35	4286,38	1,83	0,26	7,43	-0,74	0,09	0,46	-5454,56	0,09	1,13	0,09	0,81	0,56	205,08	92,33	0,79	0,25
0	5	0,17	1,32	0,47	1741,03	6634,65	4,46	0,09	39,64	-0,80	0,01	0,44	-6647,17	0,01	1,19	0,01	0,88	0,58	16,83	176,59	0,92	0,02
1	6	0,13	1,04	0,37	2020,25	6794,42	4,58	0,09	31,38	-0,72	0,57	0,41	-6875,07	0,01	11,51	0,79	0,71	0,52	15,06	195,66	0,81	359,35
2	7	0,14	1,02	0,36	2337,52	6914,62	4,65	0,09	25,15	-0,67	0,67	0,38	-7042,94	0,01	12,43	2,15	0,57	0,49	14,02	213,79	0,72	334,52
3	8	0,15	1,06	0,37	2399,63	7092,84	4,76	0,08	21,52	-0,61	0,60	0,35	-7232,96	0,01	12,11	2,53	0,47	0,45	12,58	243,88	0,62	300,30
4	9	0,15	1,06	0,38	2756,85	7164,47	4,83	0,08	18,20	-0,58	0,56	0,33	-7336,95	0,01	11,59	4,54	0,40	0,43	11,76	252,18	0,56	280,67
	REFERÊNCIA	0,36	1,70	0,47	467712,09	1139,44	-0,44	0,08	39,64	-0,51	0,01	0,48	-7336,95	0,01	0,77	0,01	0,88	0,58	897,48	252,18	0,92	0,02
	ACERTOU?	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	SIM	SIM	SIM	NÃO	NÃO	SIM	SIM

Ruído (5)		MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO
VARIAÇÃO	K	ball_hall	banfeld_ra	c_index	calinski_ha	davies_bo	det_ratio	dunn	gamma	g_plus	gdi11	gdi12	gdi13	gdi21	gdi22	gdi23	gdi31	gdi32	gdi33	gdi41	gdi42	gdi43
-3	2	1,22	147,05	0,18	469,90	1,20	4,50	0,01	0,52	0,12	0,01	0,05	0,01	1,02	7,93	2,32	0,45	3,45	1,01	0,35	2,73	0,80
-2	3	0,45	-680,70	0,10	696,44	0,58	9,21	0,03	0,68	0,08	0,03	0,22	0,07	1,03	6,69	2,11	0,54	3,49	1,10	0,50	3,23	1,02
-1	4	0,23	-1273,65	0,04	1252,48	0,53	91,35	0,02	0,86	0,03	0,02	0,17	0,05	1,13	8,89	2,61	0,52	4,07	1,19	0,47	3,72	1,09
0	5	0,08	-2004,40	0,00	4716,52	0,23	711,96	0,09	0,99	0,00	0,09	1,08	0,43	1,03	12,88	5,17	0,61	7,69	3,09	0,61	7,59	3,05
1	6	0,11	-2173,94	0,00	4638,86	0,44	1078,46	0,06	1,00	0,00	0,06	0,34	0,13	1,05	6,28	2,39	0,53	3,18	1,21	0,49	2,92	1,11
2	7	0,13	-2278,69	0,00	4477,89	0,52	1422,23	0,07	1,00	0,00	0,07	0,41	0,15	1,04	6,17	2,26	0,49	2,89	1,06	0,42	2,49	0,91
3	8	0,12	-2351,41	0,02	4176,57	0,65	1699,20	0,01	0,97	0,00	0,01	0,07	0,02	0,67	3,94	1,45	0,13	0,74	0,27	0,11	0,63	0,23
4	9	0,13	-2420,56	0,02	4057,91	0,65	2075,43	0,01	0,98	0,00	0,01	0,06	0,02	0,67	3,47	1,40	0,13	0,66	0,26	0,11	0,56	0,22
	REFERÊNCIA	1,22	-2420,56	0,00	4716,52	0,23	4,50	0,09	1,00	0,00	0,09	1,08	0,43	1,13	12,88	5,17	0,61	7,69	3,09	0,61	7,59	3,05
	ACERTOU?	NÃO	NÃO	NÃO	SIM	SIM	NÃO	SIM	NÃO	NÃO	SIM	SIM	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM

Ruído (5)		MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO
VARIAÇÃO	K	gdi51	gdi52	gdi53	ksq_detw	log_det_ra	log_ss_rat	mcclain_ra	pbm	point_bise	ray_turi	ratkowsky	scott_sym	sd_scat	sd_dis	s_dbw	silhouette	tau	trace_w	trace_wib	wemmert_	xie_beni
-3	2	0,21	1,64	0,48	485078,24	1113,09	-0,45	0,57	1,09	-0,48	0,39	0,44	-1406,84	0,75	1,13	2,33	0,42	0,37	903,03	3,50	0,46	1287,88
-2	3	0,05	0,31	0,10	533105,81	1643,32	0,64	0,47	2,05	-0,59	0,16	0,47	-2480,65	0,24	0,77	0,28	0,69	0,47	511,43	4,38	0,67	36,01
-1	4	0,05	0,37	0,11	95589,73	3340,89	1,63	0,29	5,86	-0,70	0,11	0,46	-4001,70	0,12	1,14	0,13	0,76	0,55	242,09	28,53	0,76	55,04
0	5	0,06	0,76	0,31	19164,17	4860,34	3,25	0,12	20,73	-0,76	0,03	0,44	-5154,47	0,04	1,21	0,04	0,82	0,57	55,42	51,48	0,88	1,52
1	6	0,06	0,35	0,13	18218,17	5167,64	3,45	0,11	15,82	-0,76	0,04	0,40	-5422,06	0,06	1,52	0,10	0,73	0,57	45,34	64,04	0,87	3,09
2	7	0,06	0,33	0,12	18803,34	5372,38	3,60	0,11	12,48	-0,76	0,05	0,37	-5650,16	0,07	1,71	-	0,68	0,56	39,25	73,61	0,87	1,77
3	8	0,05	0,30	0,11	20556,18	5504,06	3,69	0,11	10,76	-0,69	0,69	0,35	-5786,48	0,06	7,31	-	0,58	0,51	36,10	81,17	0,79	63,41
4	9	0,05	0,27	0,11	21300,22	5652,06	3,79	0,10	12,96	-0,69	0,62	0,33	-5909,28	0,07	8,42	-	0,54	0,51	32,55	89,28	0,78	57,17
	REFERÊNCIA	0,21	1,64	0,48	533105,81	1113,09	-0,45	0,10	20,73	-0,48	0,03	0,47	-5909,28	0,04	0,77	0,04	0,82	0,57	903,03	89,28	0,88	1,52
	ACERTOU?	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	SIM	NÃO	NÃO	SIM	NÃO	SIM	SIM	SIM	NÃO	NÃO	SIM	SIM

Densidade (3)		MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO
VARIAÇÃO	K	ball_hall	banfeld_ra	c_index	calinski_ha	davies_bo	det_ratio	dunn	gamma	g_plus	gdi11	gdi12	gdi13	gdi21	gdi22	gdi23	gdi31	gdi32	gdi33	gdi41	gdi42	gdi43
-1	2	1,09	-171,19	0,10	798,12	0,85	5,46	0,02	0,82	0,04	0,02	0,11	0,04	1,05	6,93	2,31	0,48	3,18	1,06	0,41	2,69	0,90
0	3	0,46	-593,66	0,02	1271,82	0,51	18,64	0,01	0,96	0,01	0,01	0,09	0,03	1,70	11,05	3,93	0,86	5,55	1,97	0,82	5,33	1,89
1	4	0,35	-812,53	0,06	1197,71	0,81	32,49	0,01	0,84	0,03	0,01	0,09	0,03	1,04	7,32	2,56	0,37	2,61	0,91	0,30	2,10	0,74
2	5	0,29	-966,13	0,07	1103,66	0,75	45,51	0,00	0,83	0,03	0,00	0,03	0,01	1,01	7,01	2,47	0,37	2,55	0,90	0,33	2,26	0,80
3	6	0,26	-1065,01	0,06	1051,45	0,93	61,72	0,01	0,84	0,02	0,01	0,06	0,02	1,02	6,57	2,31	0,37	2,41	0,85	0,30	1,96	0,69
4	7	0,23	-1177,29	0,06	1006,75	0,93	80,12	0,01	0,84	0,02	0,01	0,07	0,03	1,01	6,12	2,15	0,36	2,20	0,77	0,32	1,95	0,68
5	8	0,20	-1259,59	0,06	976,43	0,87	98,52	0,02	0,86	0,02	0,02	0,09	0,03	0,98	5,63	1,98	0,37	2,15	0,76	0,33	1,89	0,66
6	9	0,19	-1301,14	0,05	949,46	0,96	119,90	0,01	0,87	0,01	0,01	0,09	0,03	1,08	7,44	2,60	0,40	2,75	0,96	0,35	2,42	0,85
	REFERÊNCIA	1,09	-1301,14	0,02	1271,82	0,51	5,46	0,02	0,96	0,01	0,02	0,11	0,04	1,70	11,05	3,93	0,86	5,55	1,97	0,82	5,33	1,89
	ACERTOU?	NÃO	NÃO	SIM	SIM	SIM	NÃO	NÃO	SIM	SIM	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM

Densidade (3)		MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO
VARIAÇÃO	K	gdi51	gdi52	gdi53	ksq_detw	log_det_ra	log_ss_rat	mcclain_ra	pbm	point_bise	ray_turi	ratkowsky	scott_sym	sd_scat	sd_dis	s_dbw	silhouette	tau	trace_w	trace_wib	wemmert_	xie_beni
-1	2	0,16	1,07	0,36	368545,23	1256,24	0,08	0,43	2,59	-0,70	0,22	0,51	-1613,08	0,57	0,96	0,82	0,50	0,58	710,08	4,46	0,62	124,27
0	3	0,21	1,35	0,48	242942,88	2164,72	1,24	0,34	3,08	-0,81	0,08	0,51	-2230,58	0,23	0,65	0,50	0,61	0,67	332,03	7,14	0,71	267,33
1	4	0,17	1,16	0,41	247806,15	2575,82	1,59	0,34	3,19	-0,63	0,46	0,46	-2723,62	0,18	2,10	1,23	0,48	0,53	251,28	9,74	0,58	272,35
2	5	0,14	0,98	0,34	276425,99	2825,19	1,79	0,32	2,53	-0,57	0,34	0,41	-2981,71	0,15	2,12	2,87	0,44	0,49	210,95	12,24	0,57	2664,45
3	6	0,15	0,95	0,34	293470,77	3050,75	1,97	0,30	2,25	-0,53	0,38	0,38	-3219,86	0,13	2,31	3,35	0,38	0,45	181,07	14,72	0,51	394,59
4	7	0,14	0,86	0,30	307704,44	3243,85	2,11	0,29	1,92	-0,48	0,36	0,36	-3411,13	0,11	2,49	5,79	0,36	0,42	159,94	17,15	0,50	248,96
5	8	0,15	0,86	0,30	326848,58	3396,81	2,23	0,28	1,62	-0,47	0,34	0,34	-3560,06	0,10	2,47	8,48	0,36	0,41	142,97	18,83	0,50	145,13
6	9	0,16	1,09	0,38	339922,94	3542,10	2,34	0,26	1,76	-0,47	0,30	0,32	-3662,11	0,10	2,55	8,76	0,34	0,41	129,75	20,52	0,48	201,96
	REFERÊNCIA	0,21	1,35	0,48	368545,23	1256,24	0,08	0,26	3,19	-0,47	0,08	0,51	-3662,11	0,10	0,65	0,50	0,61	0,67	710,08	20,52	0,71	124,27
	ACERTOU?	SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO	NÃO	SIM	NÃO

Subcluster (5)		MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO
VARIAÇÃO	K	ball_hall	banfeld_ra	c_index	calinski_ha	davies_bo	det_ratio	dunn	gamma	g_plus	gdi11	gdi12	gdi13	gdi21	gdi22	gdi23	gdi31	gdi32	gdi33	gdi41	gdi42	gdi43
-3	2	0,82	-189,51	0,10	717,02	0,76	-30,65	0,32	0,68	0,08	0,32	1,64	0,48	1,07	5,42	1,58	0,72	3,66	1,07	0,63	3,16	0,92
-2	3	0,14	-1623,80	0,00	4517,70	0,33	649,93	0,68	1,00	0,00	0,68	3,86	1,16	2,08	11,79	3,55	1,45	8,21	2,47	1,42	8,06	2,43
-1	4	0,06	-2194,80	0,01	6046,07	0,27	1358,42	0,15	0,99	0,00	0,15	0,88	0,28	1,17	6,74	2,12	0,70	4,04	1,27	0,69	3,99	1,25
0	5	0,03	-2717,43	0,00	14243,75	0,30	6280,70	0,22	1,00	0,00	0,22	1,79	0,63	1,67	13,30	4,70	0,92	7,37	2,61	0,91	7,22	2,55
1	6	0,02	-2818,46	0,01	12751,13	0,58	7775,15	0,01	0,97	0,00	0,01	0,08	0,03	0,87	6,90	2,44	0,30	2,39	0,84	0,24	1,95	0,69
2	7	0,02	-2890,25	0,01	11727,71	0,79	9549,45	0,01	0,95	0,01	0,01	0,08	0,03	0,87	7,08	2,50	0,30	2,44	0,86	0,24	1,96	0,69
3	8	0,02	-2953,86	0,01	10930,82	0,95	11600,95	0,01	0,94	0,01	0,01	0,08	0,03	1,04	7,08	2,50	0,34	2,32	0,82	0,27	1,82	0,64
4	9	0,02	-3034,14	0,01	10404,86	0,82	13228,52	0,01	0,94	0,01	0,01	0,08	0,03	0,67	5,50	1,94	0,29	2,36	0,83	0,24	2,00	0,71
	REFERÊNCIA	0,82	-3034,14	0,00	14243,75	0,27	-30,65	0,68	1,00	0,00	0,68	3,86	1,16	2,08	13,30	4,70	1,45	8,21	2,61	1,42	8,06	2,55
	ACERTOU?	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM	NÃO	NÃO	SIM	NÃO	NÃO	SIM

Subcluster (5)		MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO
VARIAÇÃO	K	gdi51	gdi52	gdi53	ksq_detw	log_det_ra	log_ss_rat	mcclain_ra	pbm	point_bise	ray_turi	ratkowsky	scott_sym	sd_scat	sd_dis	s_dbw	silhouette	tau	trace_w	trace_wib	wemmert_	xie_beni
-3	2	0,27	1,37	0,40	-71221,76	-	-0,03	0,48	2,52	-0,63	0,23	0,50	-	0,50	0,95	0,50	0,61	0,48	749,66	29,65	0,62	0,86
-2	3	0,18	1,00	0,30	7557,45	4792,88	2,51	0,19	11,56	-0,92	0,03	0,56	-5110,30	0,09	0,71	0,09	0,81	0,67	111,47	96,08	0,84	0,13
-1	4	0,10	0,58	0,18	6428,12	5338,42	3,20	0,15	15,77	-0,88	0,09	0,49	-5812,79	0,04	1,78	0,05	0,81	0,63	57,63	104,05	0,86	1,78
0	5	0,17	1,35	0,48	2172,36	6471,47	4,35	0,09	32,67	-0,79	0,05	0,44	-6481,93	0,01	2,41	0,06	0,79	0,58	18,82	158,12	0,85	0,75
1	6	0,13	1,05	0,37	2526,93	6629,43	4,46	0,09	25,88	-0,71	0,57	0,41	-6709,81	0,01	8,56	0,75	0,62	0,52	16,82	174,81	0,74	355,34
2	7	0,13	1,08	0,38	2800,38	6781,54	4,56	0,09	21,15	-0,65	0,53	0,38	-6895,22	0,01	9,33	1,86	0,52	0,48	15,24	194,42	0,66	321,87
3	8	0,15	1,05	0,37	3010,83	6925,54	4,65	0,09	18,50	-0,60	0,57	0,35	-7048,03	0,01	9,87	2,93	0,43	0,44	14,01	217,07	0,58	295,85
4	9	0,11	0,88	0,31	3341,75	7022,70	4,74	0,09	16,26	-0,59	0,43	0,33	-7152,75	0,01	10,40	3,76	0,46	0,44	12,87	228,24	0,63	271,79
	REFERÊNCIA	0,27	1,37	0,48	7557,45	4792,88	-0,03	0,09	32,67	-0,59	0,03	0,56	-7152,75	0,01	0,71	0,05	0,81	0,67	749,66	228,24	0,86	0,13
	ACERTOU?	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO

Assimétrico (3)		MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO
VARIAÇÃO	K	ball_hall	banfeld_ra	c_index	calinski_ha	davies_bo	det_ratio	dunn	gamma	g_plus	gdi11	gdi12	gdi13	gdi21	gdi22	gdi23	gdi31	gdi32	gdi33	gdi41	gdi42	gdi43
-1	2	1,20	10,42	0,15	556,70	1,07	6,66	0,01	0,68	0,08	0,01	0,08	0,02	1,05	6,36	1,85	0,52	3,16	0,92	0,42	2,51	0,73
0	3	0,31	-789,45	0,03	1366,32	0,45	22,85	0,02	0,93	0,02	0,02	0,12	0,04	1,24	9,11	3,19	0,63	4,60	1,61	0,60	4,39	1,54
1	4	0,25	-1108,68	0,05	1545,32	0,72	55,17	0,02	0,91	0,02	0,02	0,11	0,04	1,39	9,35	3,29	0,48	3,25	1,15	0,40	2,70	0,95
2	5	0,21	-1316,24	0,04	1600,51	0,72	94,55	0,01	0,92	0,01	0,01	0,04	0,02	1,14	8,30	2,94	0,41	3,01	1,07	0,35	2,57	0,91
3	6	0,18	-1451,42	0,03	1654,84	0,71	150,49	0,02	0,94	0,01	0,02	0,10	0,04	1,31	8,29	2,91	0,49	3,12	1,10	0,43	2,74	0,96
4	7	0,16	-1554,73	0,03	1612,83	0,75	203,50	0,02	0,95	0,01	0,02	0,12	0,04	1,03	7,99	2,86	0,39	2,98	1,07	0,34	2,64	0,94
5	8	0,14	-1629,82	0,03	1562,05	0,77	258,62	0,01	0,95	0,01	0,01	0,07	0,03	0,92	6,28	2,22	0,39	2,66	0,94	0,34	2,30	0,81
6	9	0,14	-1693,89	0,02	1546,50	0,77	321,63	0,02	0,96	0,01	0,02	0,14	0,05	1,18	6,38	2,22	0,49	2,64	0,92	0,43	2,31	0,81
	REFERÊNCIA	1,20	-1693,89	0,02	1654,84	0,45	6,66	0,02	0,96	0,01	0,02	0,14	0,05	1,39	9,35	3,29	0,63	4,60	1,61	0,60	4,39	1,54
	ACERTOU?	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	SIM

Assimétrico (3)		MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÍNIMO	MÁXIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÍNIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÁXIMO	MÍNIMO
VARIAÇÃO	K	gdi51	gdi52	gdi53	ksq_detw	log_det_ra	log_ss_rat	mcclain_ra	pbm	point_bise	ray_turi	ratkowsky	scott_sym	sd_scat	sd_dis	s_dbw	silhouette	tau	trace_w	trace_wib	wemmert_	xie_beni
-1	2	0,22	1,30	0,38	327369,52	1402,74	-0,28	0,51	1,61	-0,57	0,33	0,46	-1650,48	0,76	1,07	1,23	0,46	0,48	842,48	5,66	0,54	307,00
0	3	0,08	0,58	0,20	214550,33	2315,51	1,31	0,36	4,86	-0,76	0,09	0,51	-2642,45	0,16	0,76	0,33	0,70	0,65	313,95	7,78	0,73	120,58
1	4	0,09	0,63	0,22	158004,11	2967,66	1,84	0,30	4,56	-0,67	0,24	0,46	-3312,75	0,13	1,62	0,86	0,58	0,57	202,50	13,14	0,65	145,42
2	5	0,09	0,64	0,23	144041,92	3366,38	2,16	0,25	4,29	-0,62	0,23	0,42	-3698,26	0,10	2,08	3,77	0,53	0,53	152,21	17,57	0,64	764,73
3	6	0,11	0,70	0,25	130318,70	3710,30	2,42	0,22	3,99	-0,60	0,20	0,39	-3955,83	0,09	2,28	4,07	0,50	0,50	120,43	22,54	0,64	150,73
4	7	0,09	0,70	0,25	131173,05	3933,61	2,58	0,21	3,75	-0,59	0,20	0,36	-4177,54	0,08	2,64	5,52	0,47	0,49	104,07	26,58	0,64	88,30
5	8	0,11	0,74	0,26	134816,08	4110,97	2,70	0,20	3,09	-0,57	0,26	0,34	-4323,33	0,07	3,29	9,64	0,44	0,48	92,74	30,47	0,63	257,44
6	9	0,14	0,77	0,27	137199,92	4272,31	2,83	0,19	2,66	-0,57	0,25	0,32	-4456,53	0,07	3,34	5,65	0,43	0,47	82,46	33,93	0,63	72,74
	REFERÊNCIA	0,22	1,30	0,38	327369,52	1402,74	-0,28	0,19	4,86	-0,57	0,09	0,51	-4456,53	0,07	0,76	0,33	0,70	0,65	842,48	33,93	0,73	72,74
	ACERTOU?	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO	SIM	SIM	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO	NÃO	SIM	NÃO

Figura 12 – Tabela de resultados consolidados.

Fonte: Elaborado pelo autor (2025)

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA

Termo de Declaração de Autenticidade de Autoria

Declaro, sob as penas da lei e para os devidos fins, junto à Universidade Federal de Juiz de Fora, que meu Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Produção é original, de minha única e exclusiva autoria. E não se trata de cópia integral ou parcial de textos e trabalhos de autoria de outrem, seja em formato de papel, eletrônico, digital, áudio-visual ou qualquer outro meio.
Declaro ainda ter total conhecimento e compreensão do que é considerado plágio, não apenas a cópia integral do trabalho, mas também de parte dele, inclusive de artigos e/ou parágrafos, sem citação do autor ou de sua fonte.
Declaro, por fim, ter total conhecimento e compreensão das punições decorrentes da prática de plágio, através das sanções civis previstas na lei do direito autoral[1] e criminais previstas no Código Penal[2], além das cominações administrativas e acadêmicas que poderão resultar em reprovação no Trabalho de Conclusão de Curso.

Juiz de Fora, 28 de agosto de 2025.

Diego Aparecido da Silva	202049030
NOME LEGÍVEL DO ALUNO (A)	Matrícula
Diego Aparecido da Silva	107.257.426-82
ASSINATURA	CPF

[1] LEI N° 9.610, DE 19 DE FEVEREIRO DE 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.
[2] Art. 184. Violar direitos de autor e os que lhe são conexos: Pena - detenção, de 3 (três) meses a 1 (um) ano, ou multa.