UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METRÓPOLE DIGITAL
BACHARELADO EM TECNOLOGIA DA INFORMAÇÃO

JOÃO VITOR DE OLIVEIRA SANTOS

CLUSTERIZAÇÃO E REDUÇÃO DE DIMENSIONALIDADE APLICADAS AO DESEMPENHO DE PAÍSES NA IMO

NATAL

2025

SUMÁRIO

1 INTRODUÇÃO

A Olimpíada Internacional de Matemática é uma das competições acadêmicas mais tradicionais do mundo e reúne anualmente participantes de dezenas de países. Ao longo das edições, esses resultados acumulados formam um conjunto de dados que revela diferenças importantes no desempenho entre as países. No entanto, observar essas diferenças de maneira estruturada não é trivial apenas com estatísticas individuais.

Este trabalho utiliza técnicas de aprendizado não supervisionado para investigar a organização dos países com base em seus indicadores históricos na IMO. Foram selecionadas métricas como médias de pontuação, distribuição de medalhas e medidas derivadas, que permitem representar cada país de forma comparável. Após o pré-processamento e a normalização dos dados, aplicamos métodos de redução de dimensionalidade e algoritmos de clusterização para explorar possíveis padrões e agrupamentos.

O objetivo é identificar como os países se distribuem em termos de desempenho e verificar se existem grupos naturalmente formados, como países de alto rendimento, intermediários e de baixo desempenho. As projeções bidimensionais e as métricas de avaliação permitem analisar a qualidade dos agrupamentos e comparar o comportamento dos métodos utilizados. Com isso, o estudo oferece uma visão mais clara da estrutura presente nos dados e ajuda a compreender de forma geral o cenário internacional da competição.

2 CONJUNTO DE DADOS

2.1 Fonte dos dados

Os dados utilizados neste trabalho foram obtidos a partir do projeto TidyTuesday, uma iniciativa que disponibiliza semanalmente conjuntos de dados públicos voltados para análise e aprendizagem em ciência de dados. Para este estudo, foi utilizado o dataset relativo à Olimpíada Internacional de Matemática. Essa base contém registros estruturados por país e por ano de participação na competição.

2.2 Variáveis originais

No conjunto original, cada linha representa a participação de um país em uma edição específica da IMO, ou seja, o desempenho do time daquele país naquele ano. As variáveis incluem as pontuações do time nas sete questões da prova (p1 a p7), bem como o número de medalhas de ouro, prata, bronze e menções honrosas obtidas pelos seus participantes. Essas informações funcionam como indicadores diretos do desempenho anual de cada país na competição.

2.3 Construção de features derivadas

Como o objetivo deste trabalho é comparar países entre si, e não edições específicas, os dados originais foram agregados por país. Para cada nação, calculamos médias das pontuações p1 a p7 e médias das quantidades de medalhas obtidas ao longo da série histórica disponível. Também foram criadas métricas derivadas, como o award_score, que pondera medalhas de acordo com sua relevância, além do total médio de medalhas e proporções por tipo. O resultado dessa etapa é uma estrutura final em que cada país é representado por um vetor numérico que resume seu desempenho médio na IMO, servindo de entrada para os métodos de análise não supervisionada utilizados nas seções seguintes.

3 PRÉ-PROCESSAMENTO

3.1. Limpeza dos dados

Os dados passaram por uma etapa de limpeza que incluiu a conversão de colunas para tipos numéricos e o tratamento de valores ausentes. Valores inválidos foram convertidos para NaN e posteriormente substituídos por zero, garantindo consistência para as etapas seguintes da análise.

3.2. Seleção de atributos

Foram selecionados apenas atributos numéricos diretamente relacionados ao desempenho dos países na IMO, como as pontuações das questões p1 a p7 e as quantidades de medalhas de cada tipo presentes no dataset original.

3.3. Construção de features derivadas e consolidação por país

Após a seleção dos atributos, foram criadas novas variáveis derivadas com o objetivo de enriquecer a representação do desempenho dos países. Em seguida, os dados foram consolidados por país, agregando todas as participações históricas em uma única observação que resume o desempenho médio de cada nação. Essa etapa foi necessária, pois o foco do trabalho é a análise comparativa entre países, e não entre edições específicas da competição.

3.4. Normalização (StandardScaler)

Após a agregação por país, todas as variáveis numéricas foram normalizadas utilizando o StandardScaler. Esse procedimento padroniza os atributos para média zero e desvio padrão igual a um, garantindo que diferenças de escala entre as variáveis não afetem o comportamento dos métodos aplicados.

3.5. Preparação da matriz X

Com os dados normalizados, foi construída a matriz X contendo apenas os atributos numéricos que representam o desempenho consolidado de cada país. Essa matriz foi utilizada como entrada para os métodos PCA e UMAP, assim como para os algoritmos de clusterização aplicados nas etapas seguintes da análise.

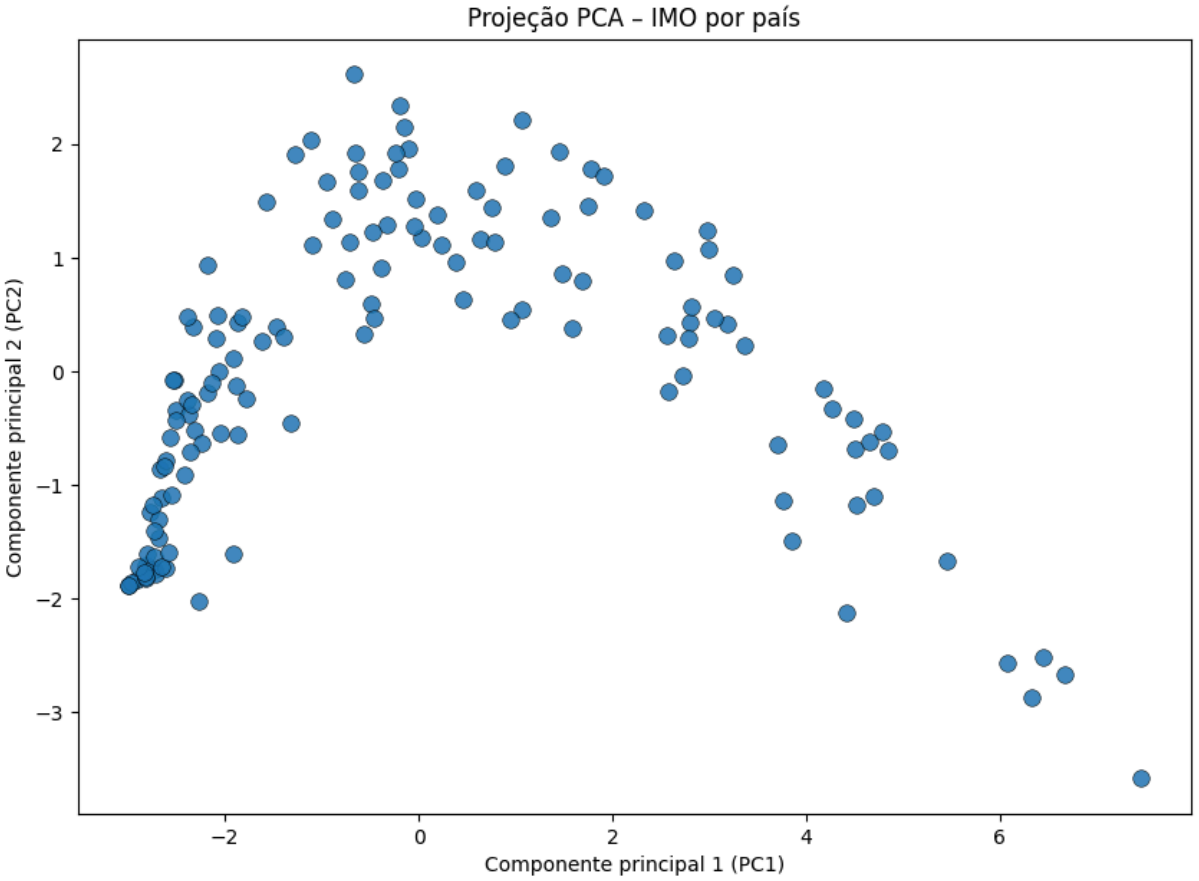
4 REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade foi utilizada com dois objetivos principais: facilitar a visualização dos países em baixa dimensão e apoiar os métodos de clusterização. Para isso, foram aplicados dois métodos amplamente utilizados: PCA, que resume a variação global dos dados em componentes principais, e UMAP, que destaca relações locais e pode revelar padrões não lineares.

4.1. PCA

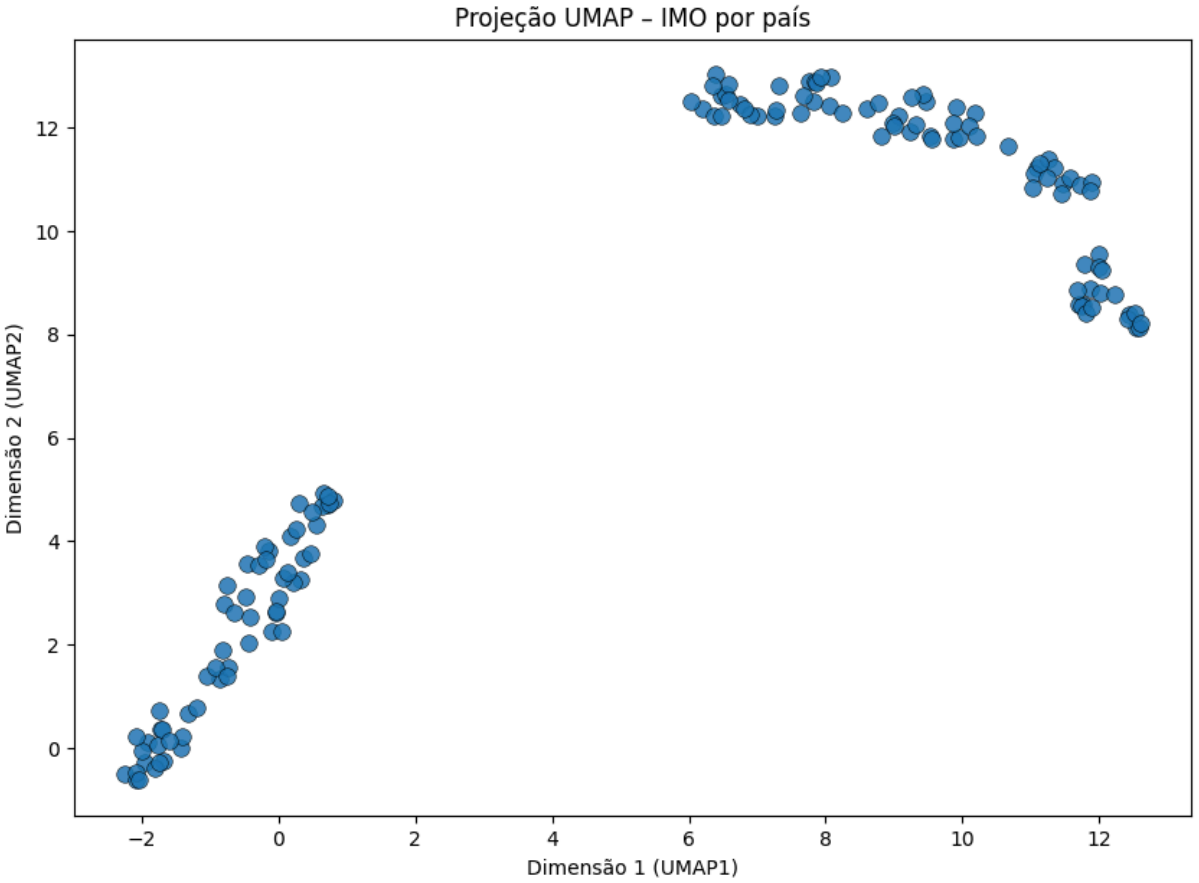
O PCA foi aplicado à matriz X normalizada com o objetivo de projetar os países em um espaço bidimensional que preservasse a maior parte da variabilidade presente nos dados originais.

A projeção 2D permitiu visualizar a distribuição geral dos países com base em seus atributos de desempenho. Os pontos foram representados em relação aos dois primeiros componentes principais, que capturam as maiores direções de variação do dataset. Essa visualização serviu como referência inicial para observar possíveis agrupamentos e relações de proximidade entre países. Os dois primeiros componentes explicaram aproximadamente 81% da variância total, o que mostrou que a redução para duas dimensões mantém uma boa representação da estrutura global dos dados.



4.2 UMAP

Após o PCA, foi aplicado o UMAP com o objetivo de capturar estruturas não lineares presentes nos dados e destacar relações de proximidade entre os países. Foram testadas diferentes combinações dos hiperparâmetros n_neighbors e min_dist, variando o grau de preservação das relações locais e a compactação dos grupos na projeção em baixa dimensão. A avaliação visual das projeções mostrou como essas escolhas influenciam a distribuição dos países, produzindo resultados mais dispersos ou mais agrupados conforme os valores adotados. A configuração selecionada gerou uma separação clara entre grupos de países com padrões semelhantes de desempenho, permitindo que a projeção fosse utilizada como base para a aplicação do DBSCAN na etapa de clusterização.

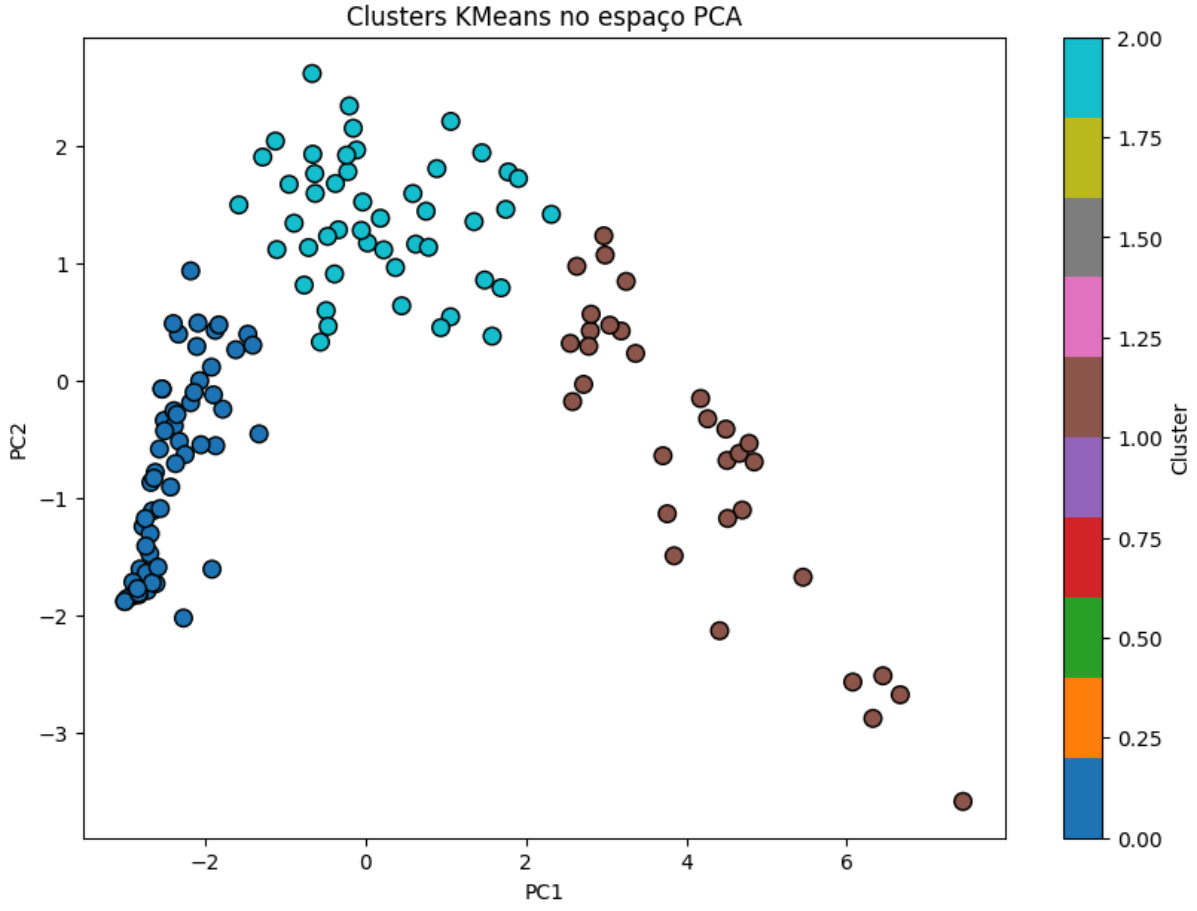

Projeção UMAP – IMO por país
Dimensão 2 (UMAP2)
Dimensão 1 (UMAP1)
12
10
8
6
4
2
0
−2
0
2
4
6
8
10
12

5 CLUSTERIZAÇÃO

5.1 K-Means no espaço PCA

O método K-Means foi aplicado sobre a projeção bidimensional obtida pelo PCA. Como o PCA conseguiu preservar uma parte significativa da variância do conjunto de dados, a projeção serviu como um espaço adequado para a aplicação do algoritmo. O K-Means foi configurado para três clusters, número escolhido com base na observação preliminar da distribuição dos países no gráfico do PCA. O resultado gerou três grupos bem definidos, com separações coerentes com as concentrações de pontos observadas na projeção.

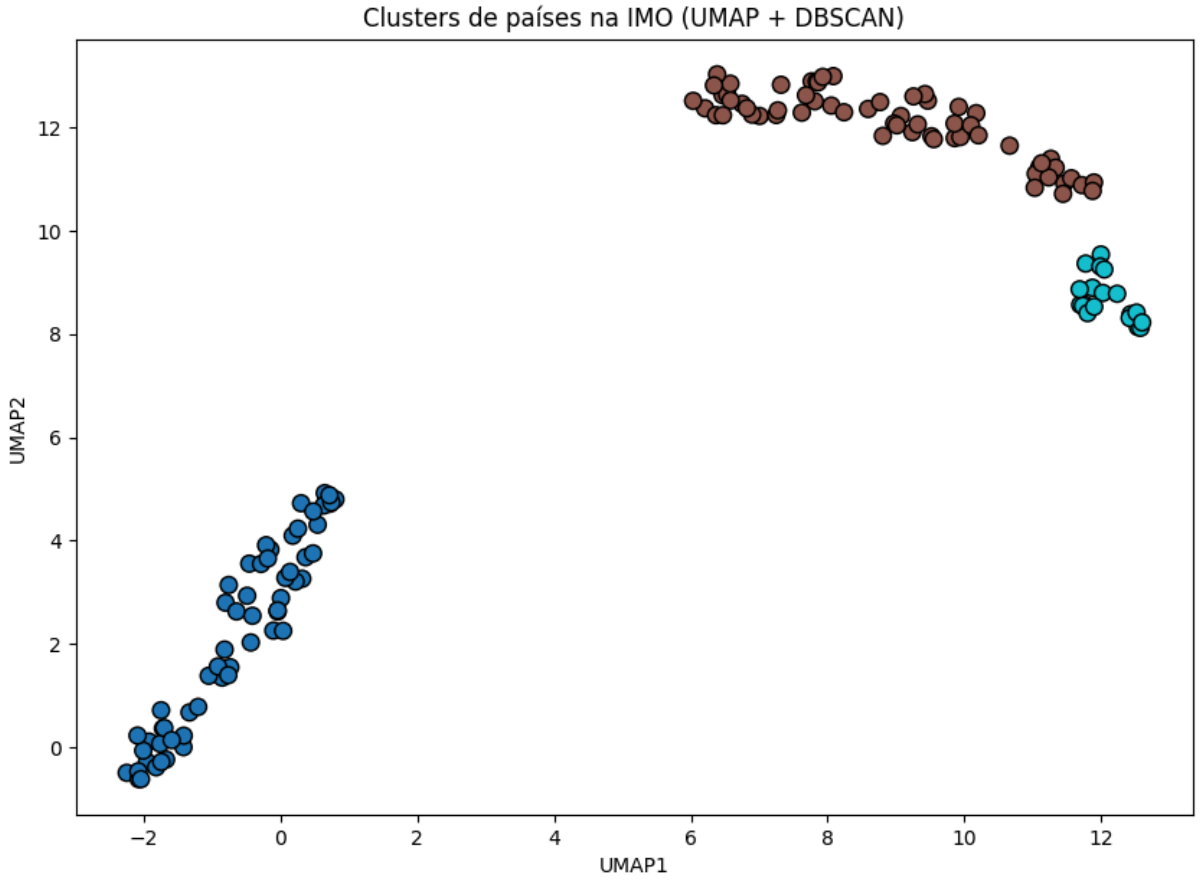
A interpretação visual mostrou que o PCA já evidenciava uma organização geral entre países com desempenho mais baixo, intermediário e elevado, e o K-Means reforçou essa divisão ao separar os países conforme similaridades globais. Após a clusterização, os países foram agrupados e analisados, permitindo identificar claramente quais nações pertenciam a cada grupo e como esses grupos se relacionavam em termos de desempenho histórico na IMO.



5.2 DBSCAN no espaço UMAP

Além do K-Means, foi aplicada a clusterização com o DBSCAN utilizando a projeção gerada pelo UMAP. Como o UMAP preserva melhor as relações locais, essa projeção tende a evidenciar regiões de maior densidade, tornando o DBSCAN uma escolha adequada para identificar agrupamentos naturais. Diferentes valores de eps e min_samples foram testados até encontrar uma configuração capaz de separar os países de forma consistente, sem gerar excesso de ruído.

O DBSCAN produziu três clusters principais, com divisão clara entre países de alto, médio e baixo desempenho. A interpretação visual da projeção do UMAP mostrou grupos bem concentrados e com limites definidos, reforçando a qualidade da clusterização. A análise dos países pertencentes a cada cluster confirmou padrões esperados, como a formação de um grupo composto pelas nações tradicionalmente mais fortes na IMO e outro com países de desempenho intermediário, além de um grupo maior com países de desempenho mais baixo.



6 AVALIAÇÃO DOS MODELOS

A avaliação dos agrupamentos foi realizada por meio de métricas internas, uma vez que não existem rótulos verdadeiros para comparar os resultados. Para o K-Means aplicado no espaço PCA, foram calculados o Silhouette Score e o índice de Davies–Bouldin. O Silhouette apresentou valor aproximado de 0,58, indicando uma separação moderada entre os clusters formados. O índice de Davies–Bouldin ficou em torno de 0,56, sugerindo que os grupos apresentam compactação razoável, mas ainda com alguma sobreposição. Esses resultados indicam que o K-Means conseguiu identificar uma estrutura coerente no espaço projetado pelo PCA, embora com limitações.

No caso do DBSCAN, a avaliação foi realizada considerando apenas os pontos atribuídos a clusters, desconsiderando o ruído. O Silhouette obtido foi de aproximadamente 0,66, superior ao valor observado no K-Means, indicando uma separação mais clara entre os grupos. O índice de Davies–Bouldin apresentou valor próximo de 0,39, sugerindo clusters mais compactos e bem definidos.

De modo geral, os resultados indicam que os agrupamentos gerados pelo DBSCAN no espaço do UMAP apresentaram qualidade superior aos obtidos pelo K-Means no PCA. A combinação UMAP + DBSCAN foi capaz de revelar grupos mais bem delimitados e coerentes com os padrões históricos de desempenho dos países na IMO.

7 CONCLUSÃO

A aplicação de técnicas de aprendizado não supervisionado permitiu organizar os países da IMO com base em indicadores consolidados de desempenho. Após o pré-processamento e a normalização dos dados, o PCA e o UMAP foram utilizados para reduzir a dimensionalidade e facilitar a identificação de padrões. O PCA forneceu uma visão global da variação entre os países, enquanto o UMAP destacou relações locais e evidenciou separações mais claras entre grupos.

Os métodos de clusterização apresentaram comportamentos distintos. O K-Means no espaço do PCA gerou três clusters coerentes, refletindo níveis diferentes de desempenho. No entanto, o DBSCAN aplicado ao UMAP produziu agrupamentos mais bem definidos, com métricas de avaliação superiores e maior separação entre os grupos. Os resultados obtidos confirmam que a combinação UMAP + DBSCAN foi mais eficaz para revelar a estrutura dos dados e identificar conjuntos de países com características semelhantes.

O estudo mostrou que métodos não supervisionados são adequados para explorar padrões em dados educacionais agregados e podem sintetizar informações complexas de forma clara e interpretável.

8 REFERÊNCIAS

- **PCA (Scikit-learn)**

  Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. Disponível em: https://jmlr.org/papers/v12/pedregosa11a.html

- **K-Means (Scikit-learn)**

  Scikit-learn Developers. *K-Means clustering*.
  Disponível em: https://scikit-learn.org/stable/modules/clustering.html#k-means

- **DBSCAN (Scikit-learn)**

  Scikit-learn Developers. *DBSCAN — Density-Based Spatial Clustering*.
  Disponível em: https://scikit-learn.org/stable/modules/clustering.html#dbscan

- **UMAP**

  McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426.
  Disponível em: https://arxiv.org/abs/1802.03426

- **Olimpíada Internacional de Matemática (IMO)**

  International Mathematical Olympiad. *Official Website*.
  Disponível em: https://www.imo-official.org/