Segmentação de Países da IMO por Desempenho

Aprendizado não supervisionado

João Vitor de Oliveira Santos

Objetivos

Agrupamento de países

Segmentar países com base em desempenho histórico utilizando técnicas de clusterização.

Redução de dimensionalidade

Aplicar métodos como PCA e UMAP para projetar dados em espaços de menor dimensão.

Comparação de métodos

Realizar clusterização com K-Means e DBSCAN, comparando resultados e avaliando qualidade dos agrupamentos.

Contexto

A **Olimpíada Internacional de Matemática (IMO)** é realizada anualmente desde 1959, reunindo estudantes de diversos países para competir em desafios matemáticos avançados.

Ao longo de mais de seis décadas, os países participantes apresentam desempenhos muito distintos, criando oportunidades para análise de padrões históricos.

Esta pesquisa utiliza o dataset público disponibilizado pelo repositório **TidyTuesday**, com interesse em identificar similaridades e agrupamentos naturais entre países com base em seus resultados históricos.

dados	
year	int
country	varchar
p1	float
p2	float
p3	float
p4	float
p5	float
p6	float
p7	float
awards_gold	float
awards_silver	float
awards_bronze	float
awards_honorable_mentions	float

Conjunto de dados

Estrutura original

Dados organizados por país e ano, onde cada linha representa o desempenho de um time em uma edição específica da IMO.

Variáveis principais

Atributos incluem pontuações nos problemas p1–p7, contagem de medalhas (ouro, prata, bronze) e menções honrosas.

Agregação final

Os dados foram agregados por país, gerando uma linha representando o desempenho médio histórico de cada nação.

Pré-processamento dos dados

Seleção de atributos

Foram selecionados apenas atributos numéricos relevantes para a análise de desempenho.

Normalização

Aplicação do **StandardScaler** para padronizar a escala dos atributos.

Conversão, limpeza e consolidação

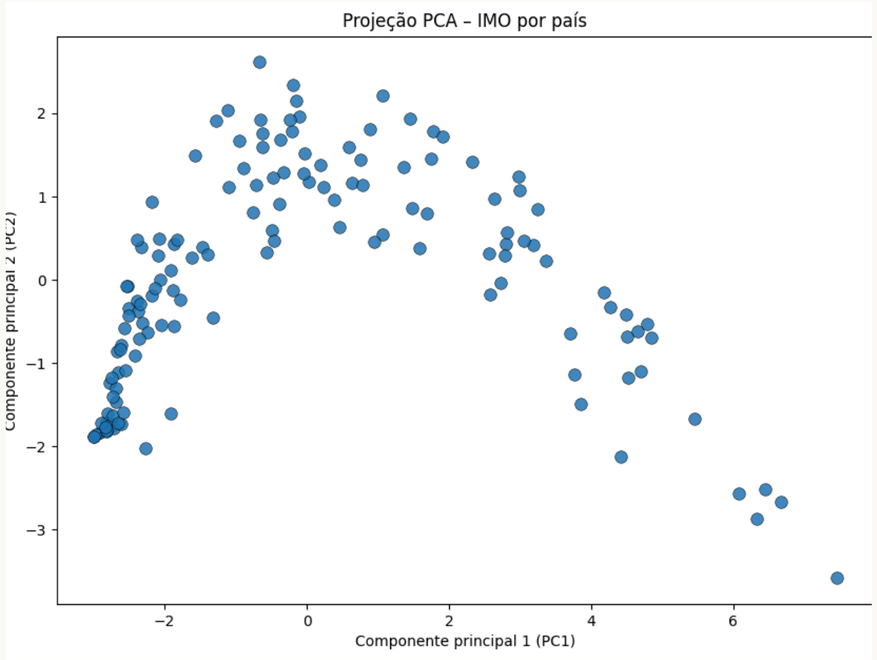
Realizada a conversão de tipos, tratamento de valores ausentes e consolidação dos registros por país, agrupando todas as participações históricas para formar um único vetor de desempenho por nação.

Matriz final

Geração da matriz **X** contendo atributos agregados e normalizados, pronta para modelagem.

Análise de Componentes Principais (PCA)

O PCA foi aplicado para transformar o conjunto original de 16 variáveis em duas componentes principais. Essa projeção reduz a dimensionalidade preservando a maior parcela da variância, permitindo analisar a estrutura dos dados em um espaço 2D sem descaracterizar seus padrões essenciais.



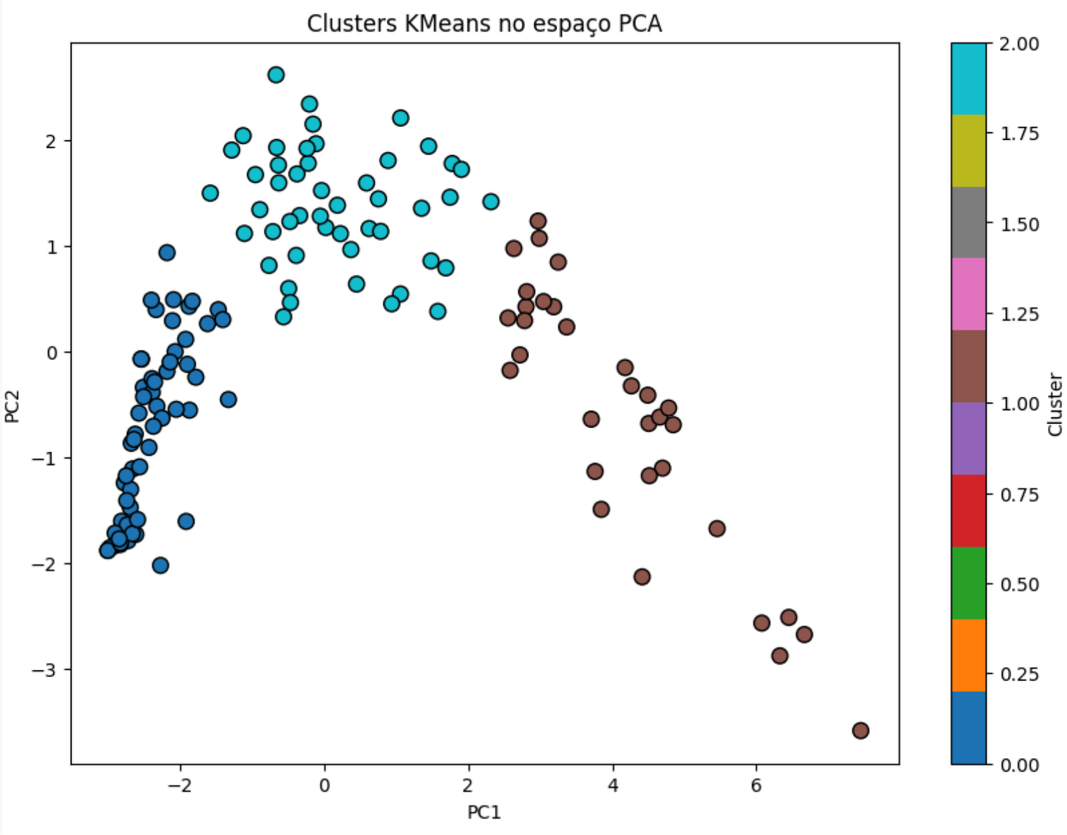
	Componente	Variância explicada	Percentual
0	PC1	0.6548	65.48%
1	PC2	0.1591	15.91%
2	Total	0.8139	81.39%

- PC1 captura o principal gradiente de desempenho (~65%).
- PC2 complementa com ~16% de variabilidade adicional.
- Juntas, PC1 + PC2 preservam ~81% da informação original.
- Essa retenção elevada permite visualizar similaridades e separações entre países.

Clusterização com K-Means

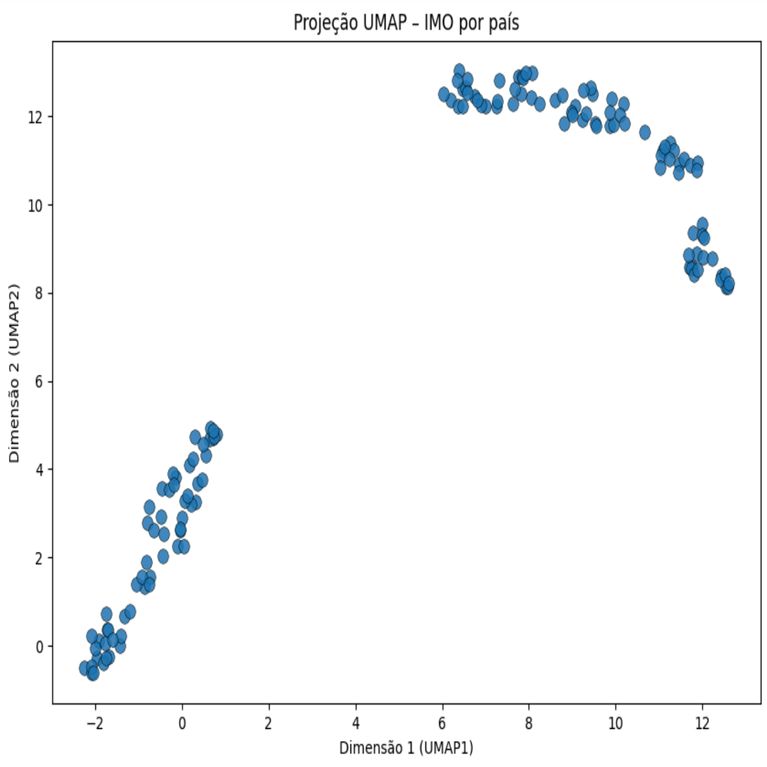
Aplicação sobre projeção PCA

O algoritmo **K-Means** foi aplicado sobre a projeção bidimensional gerada pelo PCA, buscando identificar agrupamentos naturais nos dados.



	Métrica	Valor
0	Silhouette	0.5793
1	Davies-Bouldin	0.5591

Projeção UMAP



Ajuste de hiperparâmetros

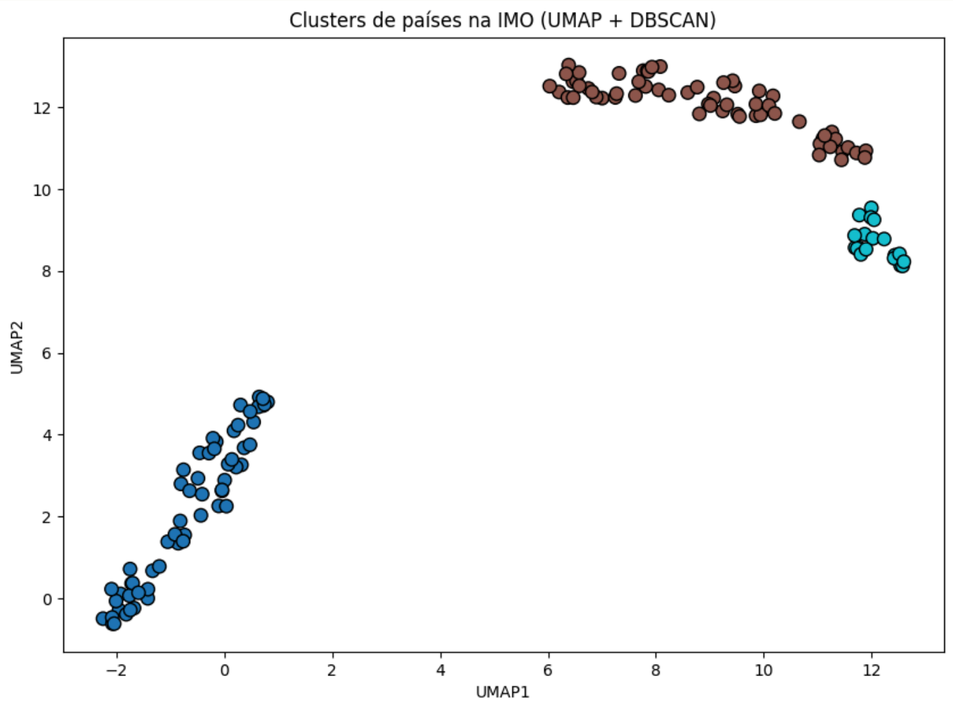
Foram realizados testes sistemáticos variando os parâmetros n_neighbors (número de vizinhos) e min_dist (distância mínima entre pontos).

Separação de grupos

A projeção final demonstrou uma separação clara e bem definida entre grupos de países, superior ao PCA.

Clusterização com DBSCAN

O algoritmo **DBSCAN** (Density-Based Spatial Clustering) foi aplicado sobre a projeção UMAP, identificando clusters baseados em densidade de pontos.



	Métrica	Valor
0	Silhouette	0.6642
1	Davies-Bouldin	0.3912

Conclusões

Estrutura dos dados

Tanto PCA quanto UMAP revelaram estrutura clara nos dados de desempenho, com o UMAP proporcionando separação mais nítida entre grupos.

Melhor desempenho

A combinação **DBSCAN + UMAP** apresentou o melhor desempenho global nas métricas de qualidade de clusterização.

Padrões históricos

Os agrupamentos identificados refletem consistentemente os padrões históricos da IMO, diferenciando países por níveis de excelência matemática.

Análise exploratória

Métodos não supervisionados mostraram-se ferramentas adequadas e eficazes para análise exploratória de desempenho competitivo.