ANÁLISE DE CLUSTERIZAÇÃO DE PAÍSES

Algoritmos de Inteligência Artificial para Clusterização

ÍNDICE

PARTE 2: ESCOLHA DE BASE DE DADOS

1.1 Quantidade de Países no Dataset

O dataset contém **167 países** com informações sobre indicadores socioeconômicos.

1.2 Análise da Faixa Dinâmica das Variáveis

As variáveis analisadas apresentam escalas muito diferentes. Por exemplo, a variável 'income' possui valores em milhares, enquanto 'health' possui valores em unidades menores. Isso indica a necessidade de normalização dos dados antes da clusterização.

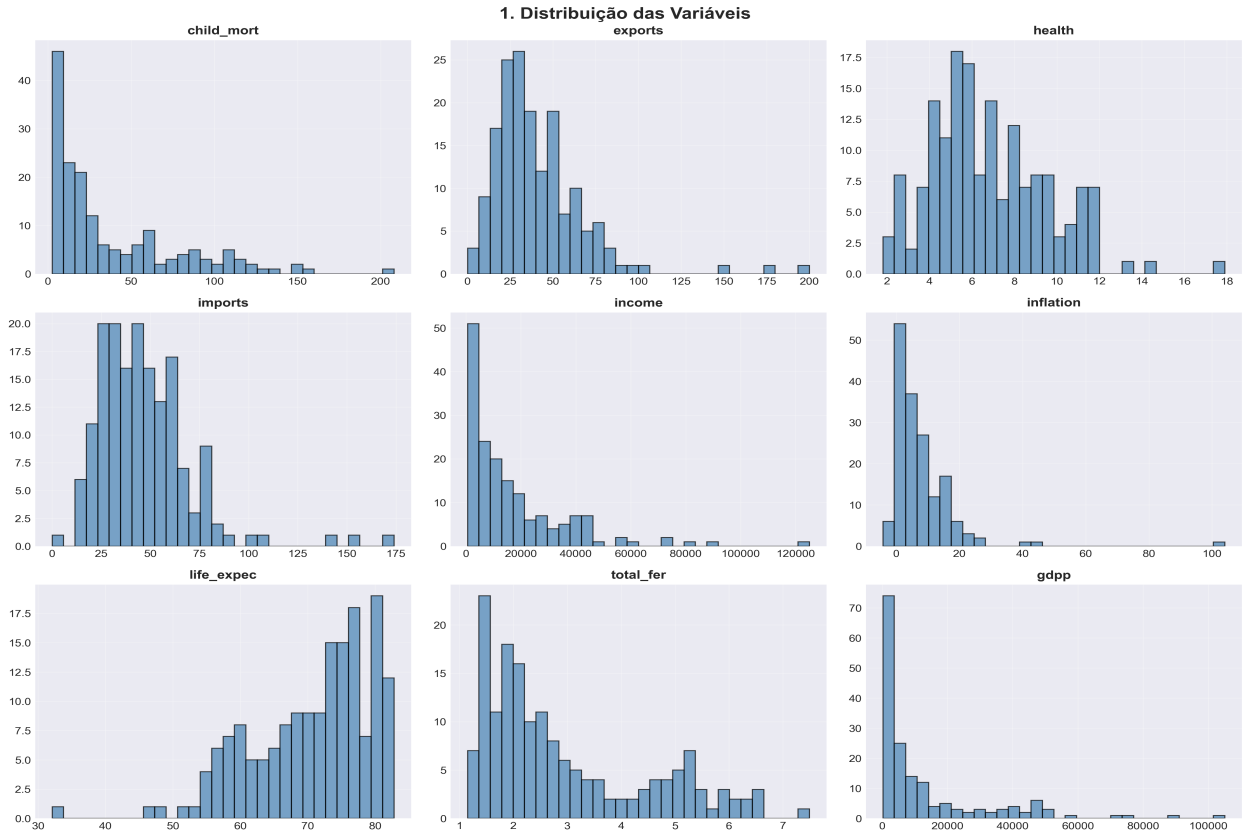

Figura 1: Distribuição das Variáveis do Dataset

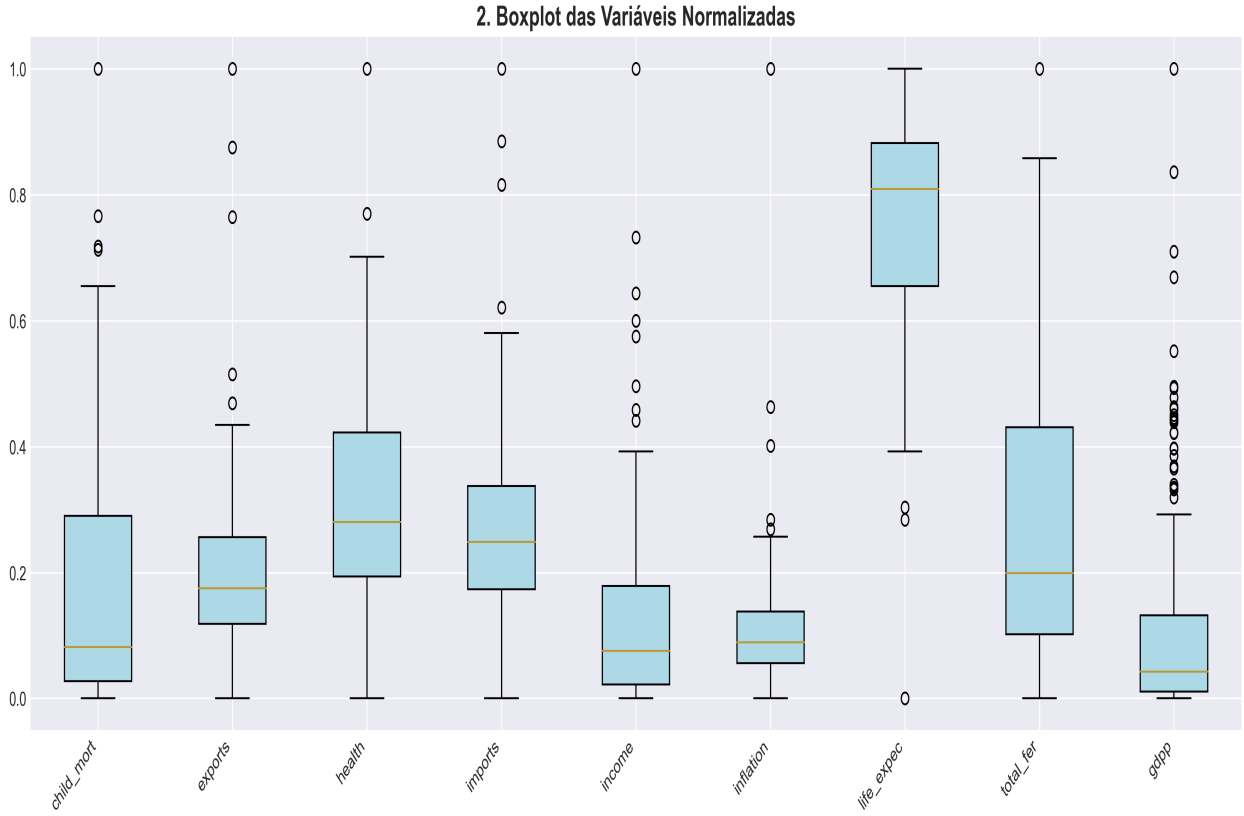

Figura 2: Boxplot das Variáveis Normalizadas - Análise de Outliers

1.3 Pré-processamento dos Dados

Etapas realizadas:

• Remoção de valores faltantes (NaN)
• Padronização usando StandardScaler
• Transformação: (X - média) / desvio_padrão
• Resultado: Dados com média=0 e desvio_padrão=1

Este pré-processamento é essencial para garantir que todas as variáveis tenham o mesmo peso nos algoritmos de clusterização.

PARTE 3: CLUSTERIZAÇÃO

2.1 Determinação do Número Ótimo de Clusters

Para determinar o número ideal de clusters, foram utilizados dois métodos: o Método do Cotovelo (Elbow Method) e o Silhueta Score. Ambos indicaram k=3 como o valor ótimo.

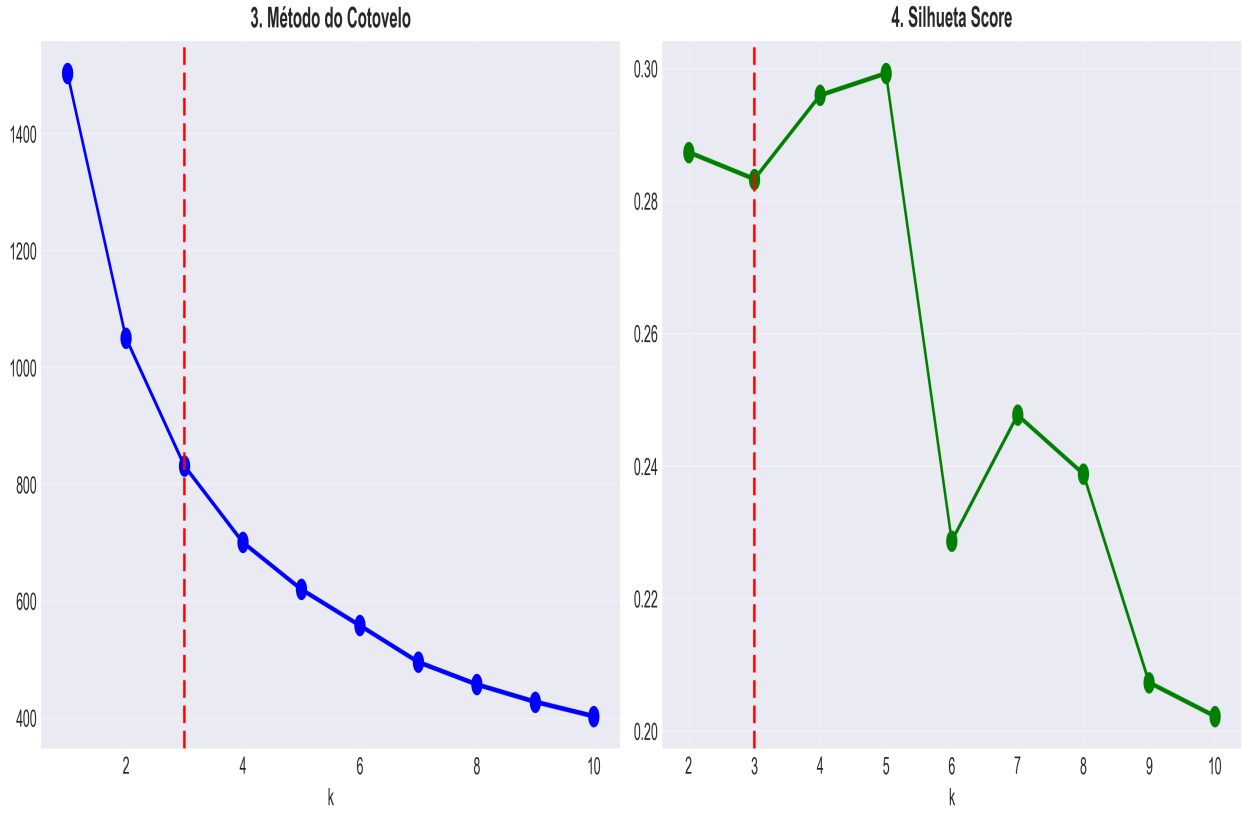

Figura 3: Método do Cotovelo e Silhueta Score para Determinação de k

2.2 Resultados K-Médias (k=3)

Distribuição dos clusters:
• Cluster 0: 36 países
• Cluster 1: 47 países
• Cluster 2: 84 países

País representativo de cada cluster:
• Cluster 0: Iceland (país mais próximo ao centróide)
• Cluster 1: Guinea
• Cluster 2: Jamaica

Figura 4: Visualização dos Clusters K-Médias (PCA)

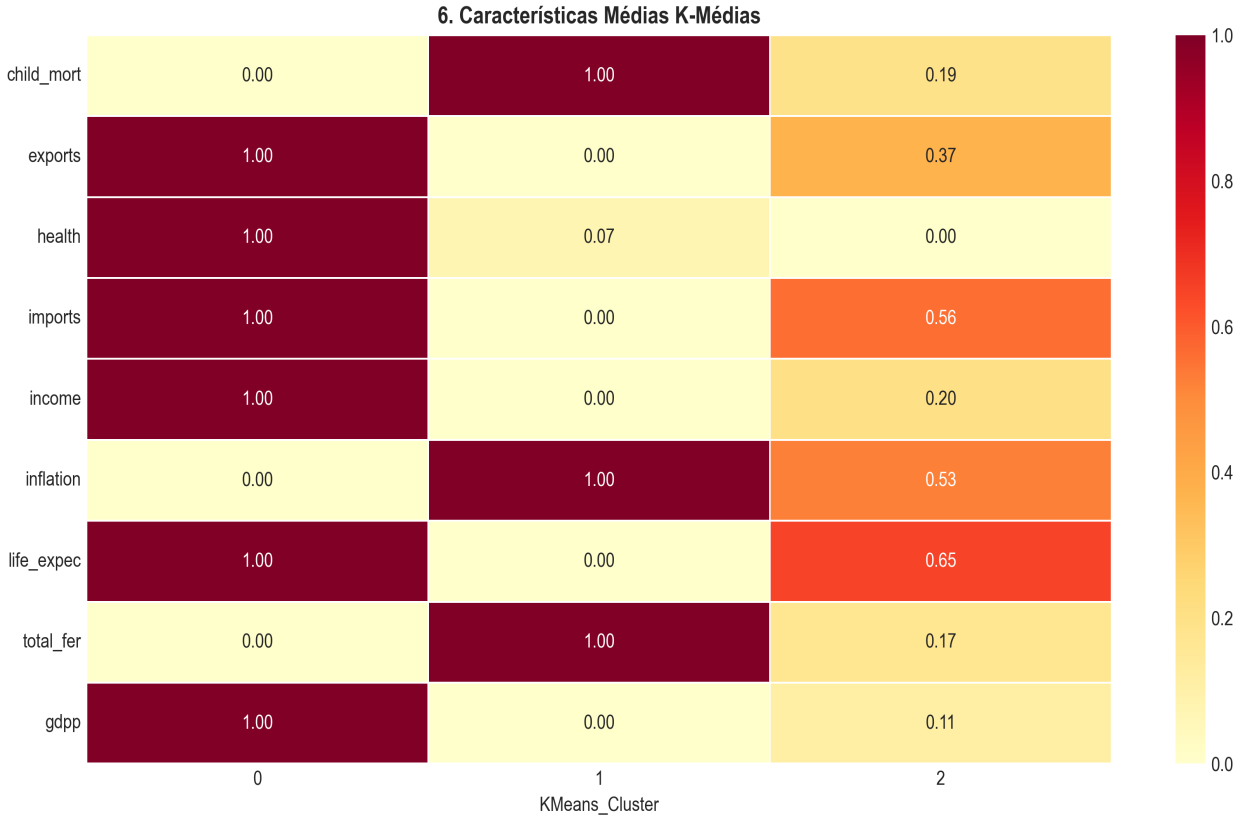

Figura 5: Características Médias dos Clusters K-Médias

2.3 Interpretação dos Clusters K-Médias

Cluster 0 (Países Desenvolvidos):
Caracterizado por alta expectativa de vida, baixa mortalidade infantil, alto PIB per capita e alta renda. Inclui países como Austrália, Áustria, Bélgica e outros países europeus desenvolvidos.

Cluster 1 (Países em Desenvolvimento):
Apresenta indicadores intermediários. Inclui países africanos e asiáticos com desenvolvimento moderado.

Cluster 2 (Países em Desenvolvimento/Subdesenvolvidos):
Maior grupo, com indicadores socioeconômicos mais baixos. Inclui países com menor PIB per capita e expectativa de vida mais reduzida.

2.4 Clusterização Hierárquica

A clusterização hierárquica foi realizada usando o método de ligação de Ward, que minimiza a variância dentro dos clusters.

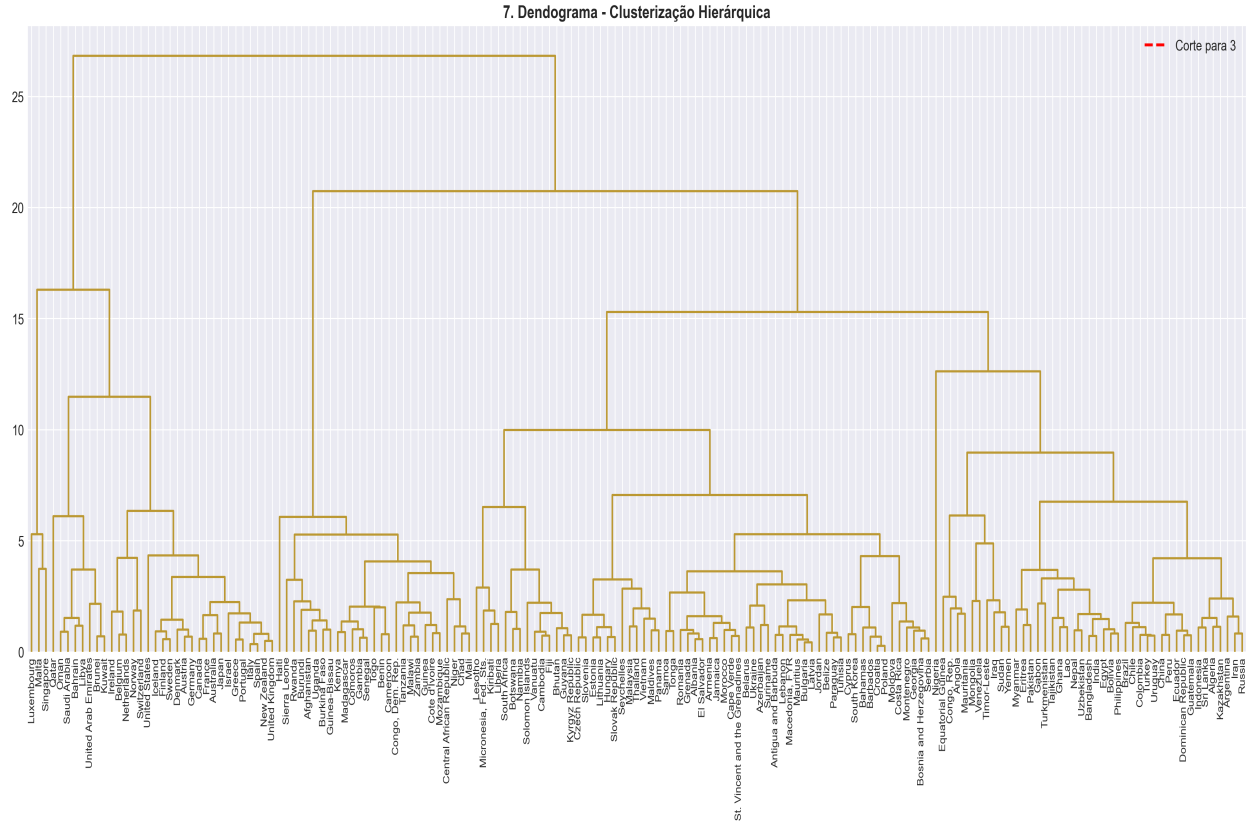

Figura 6: Dendograma - Clusterização Hierárquica

Interpretação do Dendograma:
O dendograma mostra a hierarquia de agrupamentos. O corte em altura 50 produz 3 clusters distintos:
• Cluster 0: 34 países (países desenvolvidos)
• Cluster 1: 27 países (países em desenvolvimento)
• Cluster 2: 106 países (países em desenvolvimento/subdesenvolvidos)

A estrutura hierárquica revela que os países desenvolvidos se agrupam primeiro, indicando uma separação clara entre países ricos e pobres.

Figura 7: Visualização dos Clusters Hierárquicos (PCA)

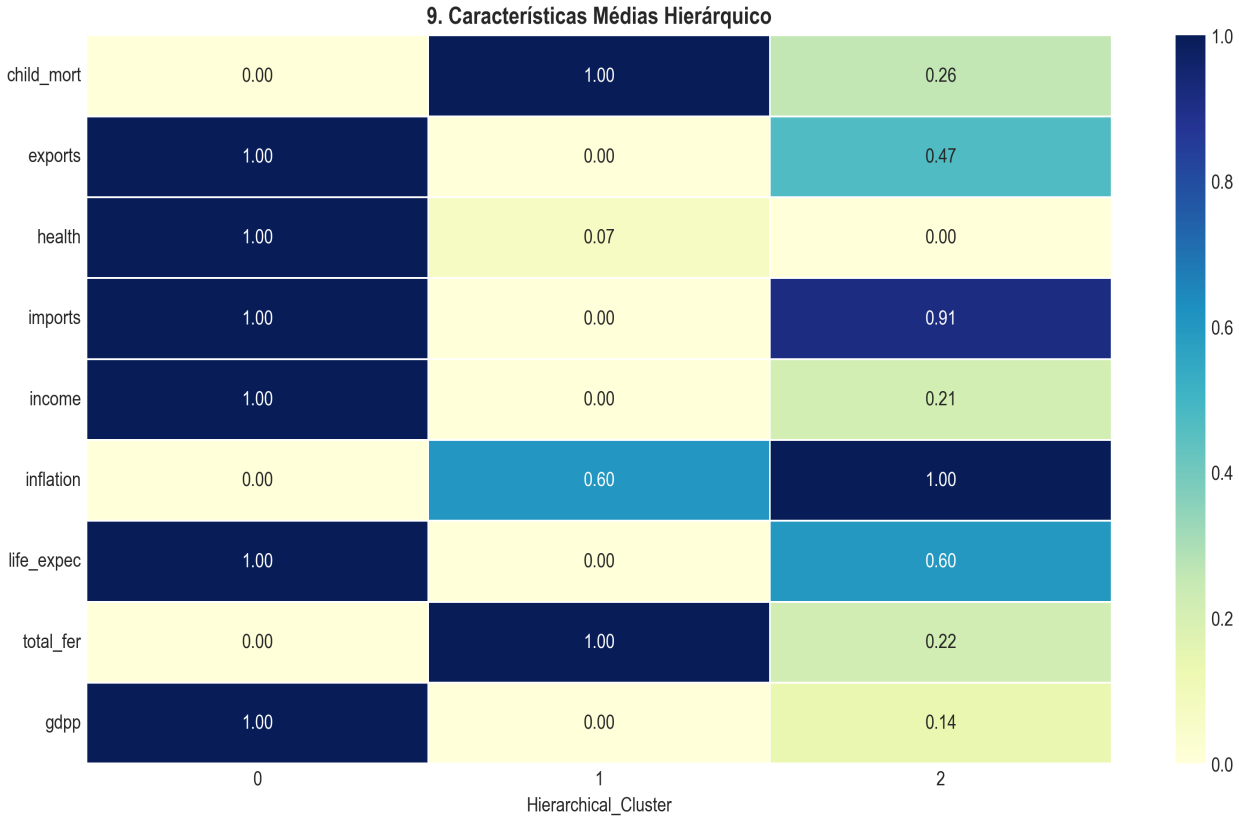

Figura 8: Características Médias dos Clusters Hierárquicos

2.5 Comparação K-Médias vs Clusterização Hierárquica

Adjusted Rand Index (ARI): 0.5370
Concordância: 139 de 167 países (83.2%)

Semelhanças:
• Ambos identificam um grupo de países desenvolvidos
• Ambos separam países por nível de desenvolvimento
• Alta concordância nas atribuições (83.2%)
• Padrões similares de agrupamento

Diferenças:
• K-Médias: distribuição mais equilibrada (36, 47, 84)
• Hierárquica: clusters mais desbalanceados (34, 27, 106)
• K-Médias otimiza a inércia global
• Hierárquica preserva a estrutura de similaridade
• K-Médias é mais rápido computacionalmente

PARTE 4: ESCOLHA DE ALGORITMOS

3.1 Etapas do Algoritmo K-Médias até Convergência

Algoritmo K-Médias:
1. Inicializar k centróides aleatoriamente
2. Atribuir cada ponto ao centróide mais próximo
3. Recalcular centróides como média dos pontos do cluster
4. Repetir passos 2-3 até convergência
5. Convergência: deslocamento dos centróides < limiar

Convergência alcançada neste dataset:
• Iterações necessárias: 19
• Inércia final: 831.42
• Centróides encontrados: 3

3.2 K-Medoids (Usando Medóides em vez de Centróides)

Diferença fundamental:
Enquanto K-Médias usa a MÉDIA dos pontos como centróide, K-Medoids usa um ponto REAL do dataset como representante do cluster (medóide).

Medóides encontrados:
• Cluster 0: Kiribati
• Cluster 1: Ghana
• Cluster 2: Poland

Vantagens do K-Medoids:
• Mais interpretável (usa dados reais)
• Menos sensível a outliers
• Medóide sempre existe no dataset

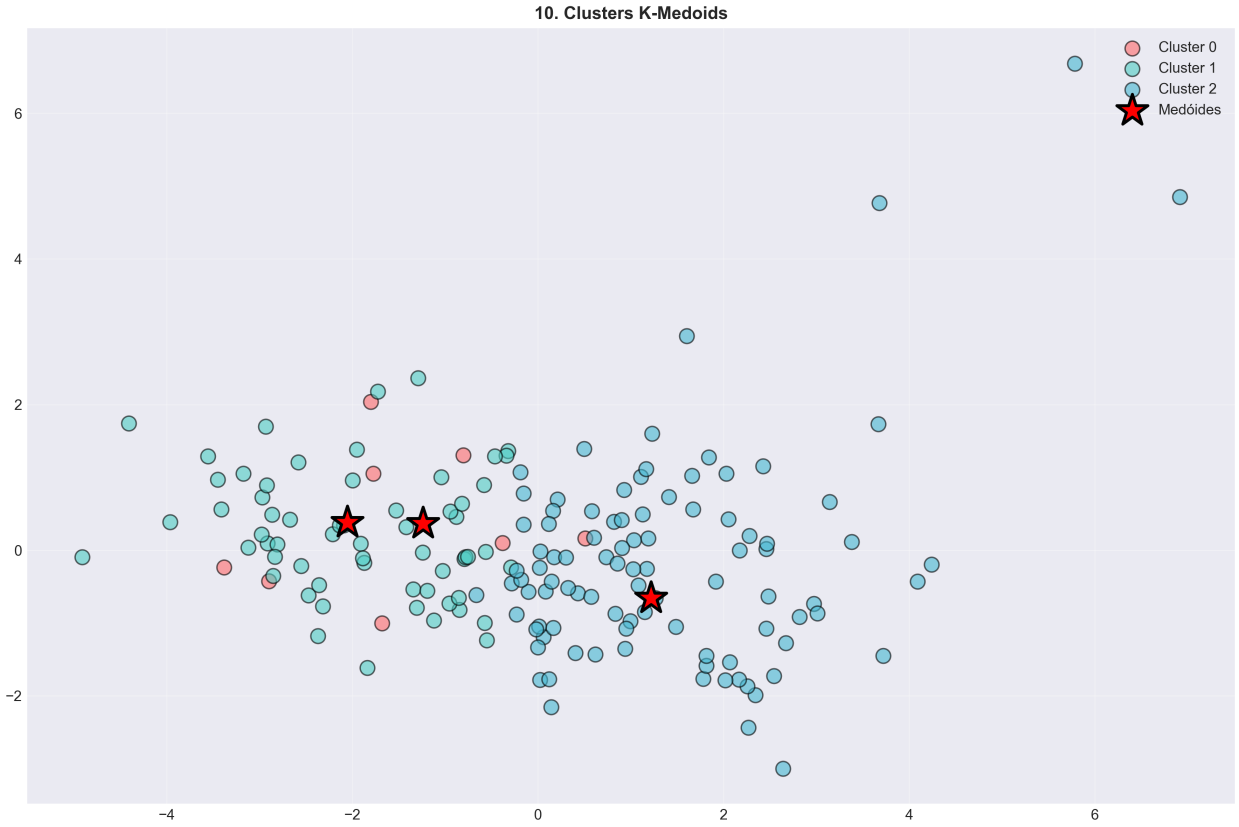

Figura 9: Visualização dos Clusters K-Medoids

3.3 Sensibilidade a Outliers

Por que K-Médias é sensível a outliers?

1. **Usa a MÉDIA como centróide:** A média é altamente influenciada por valores extremos. Um único outlier pode deslocar significativamente o centróide.

2. **Distância euclidiana amplifica diferenças:** A distância euclidiana dá mais peso a diferenças grandes, fazendo outliers terem impacto desproporcional.

3. **Deslocamento do centróide:** Um outlier pode deslocar o centróide para longe do verdadeiro centro do cluster, afetando toda a clusterização.

Exemplo: Se um cluster de países pobres contiver um país muito rico (outlier), a média será puxada para cima, distorcendo a representação do cluster.

3.4 Por que DBScan é Mais Robusto a Outliers?

Características do DBScan que o tornam robusto:

1. **Baseado em DENSIDADE:** DBScan agrupa pontos que estão próximos uns dos outros (alta densidade), em vez de usar distância euclidiana pura.

2. **Outliers ficam isolados:** Pontos que não têm vizinhos suficientes (eps e min_samples) são marcados como ruído (-1), não forçados em clusters.

3. **Não assume clusters esféricos:** DBScan encontra clusters de forma arbitrária, não limitado a formas geométricas específicas.

4. **Não tenta forçar todos os pontos:** Diferente de K-Médias, que atribui cada ponto a um cluster, DBScan permite pontos não-atribuídos (outliers).

Resultado DBScan neste dataset:
• Clusters encontrados: 1
• Outliers detectados: 30
• Percentual de outliers: 17.96%

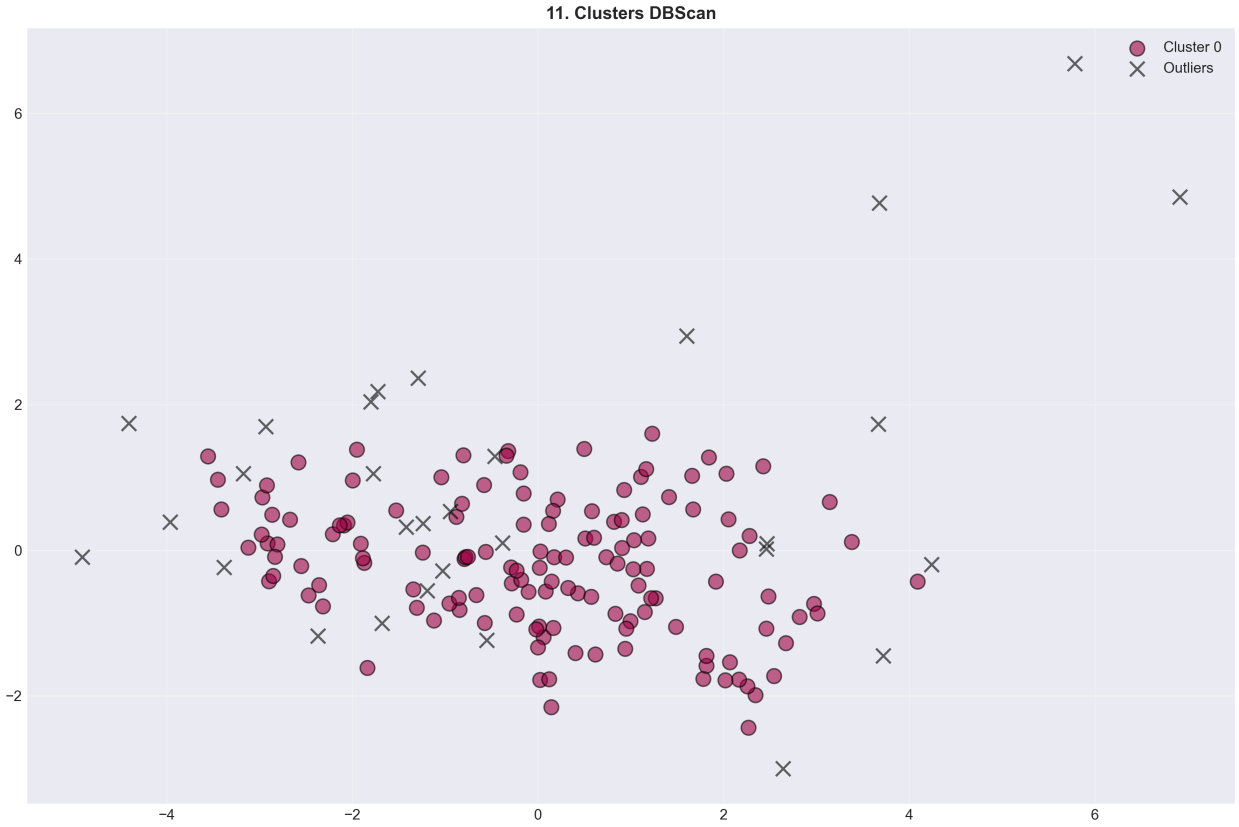

Figura 10: Visualização dos Clusters DBScan - Detecção de Outliers

CONCLUSÕES

1. Sobre a Base de Dados:
O dataset contém 167 países com 9 variáveis socioeconômicas. A análise exploratória revelou grande variabilidade nos dados, necessitando normalização antes da clusterização.

2. Sobre a Clusterização:
Tanto K-Médias quanto clusterização hierárquica identificaram padrões similares, separando países por nível de desenvolvimento. K-Médias produziu clusters mais equilibrados, enquanto a hierárquica revelou a estrutura de similaridade entre países.

3. Sobre os Algoritmos:
K-Médias é eficiente mas sensível a outliers. K-Medoids oferece uma alternativa mais robusta usando dados reais. DBScan é particularmente útil para detectar outliers e encontrar clusters de forma arbitrária.

4. Recomendações:
• Use K-Médias para datasets bem estruturados e sem muitos outliers
• Use K-Medoids quando interpretabilidade é importante
• Use DBScan quando há suspeita de outliers ou clusters de forma irregular
• Sempre realize pré-processamento adequado dos dados