# Introdução e Motivação

## Introdução

A avaliação de dados educacionais é crucial para a criação de políticas públicas, o planejamento de investimentos e a promoção de uma sociedade mais justa. Em um cenário europeu, onde as desigualdades entre regiões e grupos sociais persistem, entender os níveis de educação pode fornecer dados valiosos para tomadas de decisões estratégicas.

## Motivação para o Projeto

A motivação para este projeto vem da necessidade de investigar as diferenças no nível educacional entre países, idades e gêneros na Europa. Com essa avaliação, busca-se reconhecer padrões, agrupamentos e potenciais disparidades que possam orientar futuras pesquisas ou iniciativas de aprimoramento por parte de instituições de ensino e governos.

## Descrição do Problema

O conjunto de dados empregado inclui dados sobre o grau de instrução da população europeia, divididos por idade, gênero e localização geográfica. Education attainment is given by International Standard Classification of Education (ISCED11).

ISCED11 education levels are the following:

**X** - No schooling

**0** - Early childhood education

**1** - Primary education

**2** - Lower secondary education

**3** - Upper secondary education

**4** - Post-secondary non-tertiary education

**5** - Short-cycle tertiary education

**6** - Bachelor’s or equivalent level

**7** - Master’s or equivalent level

**8** - Doctoral or equivalent level

**9** - Not elsewhere classified

Devido à vasta quantidade e variedade dos dados, é um desafio identificar padrões pertinentes sem o uso de métodos adequados de análise e agrupamento. Portanto, o projeto visa utilizar algoritmos de aprendizado de máquina (especialmente de clusterização) para identificar agrupamentos naturais nos dados que possuam características semelhantes.

## Objetivos do Projeto

Os objetivos centrais deste projeto são:

* Executar uma análise exploratório na base de dados de educação europeia;
* Utilizar métodos de agrupamento para agrupar observações com atributos parecidos;
* Analisar o rendimento de variados algoritmos de agrupamento;
* Analisar os agrupamentos identificados para obter percepções pertinentes sobre o perfil educacional dos europeus.

## Justificativa para o Conjunto de Dados Escolhido

As informações foram obtidas do site oficial da União Europeia, Eurostat, e registradas na plataforma Kaggle. Ele proporciona uma fundamentação sólida e segura sobre o grau de instrução da população, sendo perfeito para usos em aprendizagem não supervisionada. A sua segmentação por idade, gênero e localização faz com que o conjunto seja particularmente eficaz para detectar padrões educacionais ocultos que não seriam facilmente identificados por análises tradicionais.

# Algoritmos de Clusterização

Este projeto selecionou dois algoritmos de agrupamento: K-Means e Clustering Hierarquizado (Aglomerativo Clustering). Esses dois são frequentemente empregados em análises exploratórias de dados e proporcionam métodos distintos para a criação de grupos.

## Justificativa da Escolha

* **K-Means** foi selecionado devido à sua eficácia no processamento de grandes conjuntos de dados e simplicidade na sua interpretação. Ele é particularmente eficaz quando o número de agrupamentos pode ser estabelecido antecipadamente.
* **Hierarchical Clustering (Agglomerative)** foi escolhido devido à sua habilidade de criar uma estrutura hierárquica dos dados (dendrograma), útil para compreender de forma mais clara as relações entre os grupos, sem a exigência de estabelecer previamente a quantidade de grupos.

Em virtude do elevado uso de memória do algoritmo de Cluster Hierarquizado, foi preciso diminuir a amostra para 10.000 observações antes de implementar o modelo. Este processo foi crucial para prevenir o erro de memória originado pela complexidade do método computacional.

## Resultados e Comparação

|  |  |  |
| --- | --- | --- |
| Algoritmo | Silhouette Score | Davies-Bouldin Index |
| K-Means | 0.787 | 0.466 |
| Agglomerative | 0.748 | 0.482 |

* Pontuação Silhouette: Quanto mais perto de 1, mais eficaz é a distinção entre os agrupamentos. O K-Means demonstrou um desempenho um pouco superior, sugerindo uma definição mais precisa dos grupos.
* Índice Davies-Bouldin: Quanto menor, mais eficaz é a compactação e a distinção entre os agrupamentos. O K-Means também apresentou um desempenho um pouco superior neste parâmetro.

## Conlusão

Os dois algoritmos se mostraram eficientes na segmentação dos dados, contudo, o K-Means obteve resultados um pouco melhores em ambas as métricas empregadas. Ademais, ele se mostrou mais eficaz no uso de tempo e memória. O Clustering Hierarquizado, apesar de ser mais rico em recursos, proporcionou uma análise mais aprofundada das relações hierárquicas entre os dados.

Portanto, para grandes conjuntos de dados, o K-Means mostrou-se mais adequado, ao passo que o Hierarchical Clustering pode servir como um recurso adicional para análises mais detalhadas com amostras menores.