# 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 escolaridade da população europeia, divididos por faixas etárias, 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 é responder a seguinte pergunta:

**Como agrupar países europeus com base em padrões de nível educacional ao longo dos anos?**

Ela deverá ser respondida seguindo alguns passos:

* 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 | 0.310 | 0.802 |
| Agglomerative | 0.743 | 0.466 |

Pontuação Silhouette: Quanto mais perto de 1, mais eficaz é a distinção entre os agrupamentos. O Agglomerative agrupou os dados de forma muito mais coesa e separada entre clusters.

Índice Davies-Bouldin: Quanto menor, mais eficaz é a compactação e a distinção entre os agrupamentos. O Agglomerative formou clusters mais compactos e distintos.

## Conclusão

Usando duas dimensões, Date(Ano) e Value, foram identificados quatro clusters de dados de Grau de Escolaridade, Faixas Etárias, Gênero, Países e Classificação ISCED11.

O algoritmo K-Means alcançou uma pontuação Silhouette de 0.310 e um índice Davies-Bouldin de 0.802, sinalizando uma separação moderada entre os agrupamentos e uma certa sobreposição entre eles. Por outro lado, o método Hierarchical Clustering (Agglomerative Clustering) mostrou um desempenho consideravelmente superior, com uma pontuação Silhouette de 0,743 e um índice Davies-Bouldin de 0,466, demonstrando a criação de agrupamentos mais harmoniosos e bem delimitados. Estes achados sugerem que a clusterização hierárquica se mostrou mais eficiente para identificar os padrões educacionais contidos nos dados, tornando-se, assim, a estratégia mais apropriada para esta avaliação.

# EDA de Série Temporal

Para se responder a pergunta de: Como agrupar países europeus com base em padrões de nível educacional ao longo dos anos? Devemos entender o comportamento dos dados ao longo do tempo, encontrar dados faltantes, valores extremos ou padrões sazonais importante.

### Média/mediana de nível educacional por ano

### Evolução temporal por país

Mapenado os 5 paises com mais registros para identificar a evolução do nível educacional de cada pais.

### Evolução temporal por faixa etária

Mapeado as 5 faixas etárias com mais registros para identificar nivel educacional por Faixa Etária.

### Heatmap (Ano x País)

### Rolling Mean

### Decomposição

A Decomposição demonstrou que não foi detectada uma sazonalidade clara na série. Isso pode ter ocorrido porque os dados têm frequência anual, e a decomposição espera múltiplas observações por período (ex: mensal ou trimestral dentro de cada ano).

O componente Residual também sugere que não sobrou variação não explicada após separar a tendência e a ausente sazonalidade, ou seja, tudo o que podia ser modelado está contido na tendência.

# ARIMA

A modelagem ARIMA foi aplicada a uma série temporal representando a média de nível educacional ao longo dos anos em um país europeu, com o objetivo de entender padrões que possam sustentar os agrupamentos realizados na análise de clusterização. Foi utilizado os dados da Alemanha como exemplo para a modelagem:

## Análise de Estacionariedade

### Teste de Dickey-Fuller Aumentado (ADF)

* **ADF Statistic:** -2.455
* **p-value:** 0.127

Como o p-valor é maior que 0.05, não rejeitamos a hipótese nula de que a série possui uma raiz unitária, ou seja, a série não é estacionária em seu estado original. Isso implica que os dados apresentam tendência ao longo do tempo, o que reforça a necessidade de aplicar uma diferenciação antes da modelagem com ARIMA, tornando-a apropriada para séries com comportamento não constante.

## Desempenho do Modelo ARIMA (1,1,1)

* **RMSE (Root Mean Squared Error):** 161.33
* **MAE (Mean Absolute Error):** 80.68

O RMSE, por penalizar mais fortemente grandes erros, sugere que há flutuações expressivas não capturadas pelo modelo em alguns anos.

O MAE mostra que, em média, o erro absoluto é de cerca de 80 unidades na escala da variável, o que pode ser considerado alto ou baixo dependendo da amplitude da série.

Esses valores indicam que, apesar do modelo capturar parcialmente a tendência da série, existem desvios consideráveis entre os valores reais e os previstos, especialmente em anos com maior variação.

## Conclusão

Os resultados indicam que o modelo ARIMA (1,1,1) é razoável, mas pode não capturar todas as nuances da série temporal, possivelmente devido a variações estruturais ou sazonalidades não modeladas.

Estes resultados indicam que o país estudado (Alemanha) apresenta um padrão educacional que sofre alterações significativas ao longo do tempo, o que pode ser um fator relevante para a classificação. Países com tendências similares podem formar agrupamentos diferentes daqueles que exibem estabilidade ou sazonalidade acentuada. A característica não estacionária da série evidencia que a dinâmica da educação muda em resposta a elementos políticos, econômicos e sociais, o que pode ser observado na comparação entre nações européias.

Em última análise, mesmo que o modelo ARIMA não tenha um desempenho impecável, ele oferece percepções valiosas sobre o percurso histórico da educação em um país específico, reforçando a análise exploratória e auxiliando na justificação dos agrupamentos gerados pelos algoritmos de clusterização.