# Relatório de Análise de Clustering Socioeconômico

## 1. Definição do Problema

O presente trabalho tem como objetivo aplicar técnicas de aprendizado de máquina não supervisionado para identificar padrões e segmentar indivíduos com base em variáveis demográficas e socioeconômicas. Utilizando o algoritmo KMeans, buscamos agrupar 2000 indivíduos em clusters que revelem tendências relacionadas à distribuição de renda, ocupação, tamanho do assentamento e outros fatores socioeconômicos. Este estudo é relevante para compreender relações entre idade, estado civil, bem-estar econômico e localização geográfica, com aplicações em políticas públicas e estratégias de mercado.

## 2. Justificativa

A análise de dados socioeconômicos é fundamental para entender as dinâmicas que influenciam a distribuição de renda, o acesso ao emprego e o bem-estar econômico. Este conjunto de dados contém informações demográficas e socioeconômicas de 2000 indivíduos, incluindo atributos como idade, nível de escolaridade, ocupação, renda e tamanho do assentamento. É ideal para estudos relacionados à distribuição de renda, tendências de emprego e fatores socioeconômicos que influenciam a situação financeira. A segmentação de indivíduos em grupos homogêneos pode fornecer insights valiosos para análises socioeconômicas (estudando tendências no emprego e tamanho do assentamento) e estudos demográficos (explorando as relações entre idade, estado civil e bem-estar econômico). Este conjunto de dados pode ser usado para análise exploratória de dados, modelagem estatística e aplicativos de aprendizado de máquina relacionados a insights socioeconômicos.

## 3. Descrição do Conjunto de Dados

O dataset contém 2000 registros com 8 variáveis: `ID`, `Sex`, `Marital status`, `Age`, `Education`, `Income`, `Occupation` e `Settlement size`. As variáveis incluem dados demográficos (idade, sexo, estado civil), socioeconômicos (renda, ocupação, nível de escolaridade) e geográficos (tamanho do assentamento). A variável `ID` foi usada apenas para identificação e não incluída na análise. Não há uma variável alvo definida, pois o objetivo é clustering não supervisionado. Algumas variáveis categóricas (`Marital status`, `Education`, `Occupation`) foram codificadas, e valores ausentes, se presentes, foram tratados por imputação (mediana para variáveis numéricas e moda para categóricas).

## 4. Análise Exploratória dos Dados (EDA)

A análise exploratória revelou uma distribuição ampla de idades (18 a 76 anos) e rendas (mínimo de 89 mil, máximo de 159 mil, aproximadamente). Um heatmap de correlação mostrou associações moderadas entre `Income` e `Education`, e entre `Settlement size` e `Occupation`, indicando que indivíduos em áreas urbanas tendem a ter ocupações mais qualificadas. Boxplots mostraram que indivíduos com maior renda são, em média, mais velhos e possuem maior nível de escolaridade. A variável `Sex` apresentou distribuição equilibrada, enquanto `Marital status` mostrou uma leve predominância de indivíduos não solteiros (casados, divorciados, viúvos).

# 5. Preparação dos Dados

As variáveis categóricas (`Sex`, `Marital status`, `Education`, `Occupation`) foram transformadas com codificação ordinal usando `OrdinalEncoder`, já que o KMeans requer dados numéricos. As variáveis numéricas (`Age`, `Income`, `Settlement size`) foram padronizadas com Z-score usando `StandardScaler` para evitar que variáveis com escalas diferentes (como renda e idade) dominassem o processo de clustering. Não foi realizada divisão em treino e teste, pois o objetivo é clustering não supervisionado, e todos os dados foram usados para a análise.

# 6. Modelagem e Avaliação

Foi utilizado o algoritmo KMeans para realizar o clustering, com o número de clusters ((k)) definido como 5, determinado pelo método do cotovelo com auxílio da biblioteca `kneed`. Além do KMeans, exploramos clustering hierárquico com dendrogramas para confirmar a estrutura dos dados. As métricas de avaliação utilizadas foram o Silhouette Score (para medir a separação entre clusters) e o Davies-Bouldin Score (para avaliar a compactação e separação dos clusters). O KMeans com (k=5) apresentou um Silhouette Score de aproximadamente 0,35 e um Davies-Bouldin Score de 1,2, indicando segmentação razoável, mas com sobreposição entre alguns clusters.

Os clusters foram analisados com base nas médias das variáveis numéricas e nas proporções das variáveis categóricas. Abaixo, apresentamos os perfis detalhados de cada cluster, seguidos pelos padrões gerais e insights identificados.

### Perfis dos Clusters

- **Cluster 0: "Profissionais Solteiros de Renda Média-Alta"**
    - **Tamanho**: 579 indivíduos.
    - **Idade**: Média de 36,9 anos, mediana de 36, desvio padrão de 8,59.
    - **Renda**: Média de 139.478, mediana de 132.243, desvio padrão de 32.788.
    - **Sexo**: 100% do mesmo gênero (valor 0, possivelmente masculino).
    - **Estado Civil**: 100% solteiros.
    - **Educação**: 66,9% com ensino médio, 33,1% com nível desconhecido/outro.

    - **Ocupação**: 75% empregados qualificados/oficiais, 24,2% gerentes/autônomos/funcionários altamente qualificados.
    - **Tamanho do Assentamento**: Média de 1,33 (cidades médias ou grandes).
    - Este grupo é formado por profissionais estabelecidos, solteiros, com boa renda e vivendo em áreas urbanas.
- **Cluster 1: "Jovens Mistos de Renda Média em Áreas Rurais"**

    - **Tamanho**: 254 indivíduos.
    - **Idade**: Média de 29,8 anos, mediana de 29, desvio padrão de 6,20.
    - **Renda**: Média de 113.688, mediana de 112.227, desvio padrão de 13.071.
    - **Sexo**: Média de 0,53 (mistura de gêneros, ligeiramente equilibrada).
    - **Estado Civil**: 29,8% solteiros, 70,2% não solteiros (casados, divorciados, viúvos).
    - **Educação**: 92,6% com ensino médio, 7,2% com nível desconhecido/outro, 0,2% com universidade.
    - **Ocupação**: 98,7% empregados qualificados/oficiais, 1,3% desempregados/não qualificados.
    - **Tamanho do Assentamento**: Média de 0,28 (áreas rurais).
    - Este grupo representa jovens trabalhadores qualificados, muitos casados ou em relacionamentos, vivendo em áreas rurais com renda moderada.
- **Cluster 2: "Desempregados Solteiros de Renda Baixa em Áreas Rurais"**

    - **Tamanho**: 534 indivíduos.
    - **Idade**: Média de 35,8 anos, mediana de 35, desvio padrão de 9,49.
    - **Renda**: Média de 89.885, mediana de 84.779, desvio padrão de 23.377.
    - **Sexo**: 100% do mesmo gênero (valor 0, possivelmente masculino).
    - **Estado Civil**: 85,7% solteiros, 14,3% não solteiros.
    - **Educação**: 74,8% com ensino médio, 25% com nível desconhecido/outro, 0,2% com universidade.
    - **Ocupação**: 100% desempregados/não qualificados.
    - **Tamanho do Assentamento**: Média de 0,12 (majoritariamente áreas rurais).
    - Este grupo é formado por indivíduos desempregados ou não qualificados, com baixa renda, vivendo em áreas rurais e majoritariamente solteiros.
- **Cluster 3: "Desempregadas Jovens de Renda Baixa em Áreas Rurais"**

    - **Tamanho**: 272 indivíduos.
    - **Idade**: Média de 32,9 anos, mediana de 30, desvio padrão de 5,12.
    - **Renda**: Média de 87.327, mediana de 86.319, desvio padrão de 20.671.
    - **Sexo**: Média de 0,96 (quase todos do gênero oposto, possivelmente feminino).
    - **Estado Civil**: 69,1% solteiros, 30,9% não solteiros.
    - **Educação**: 71,7% com ensino médio, 28,3% com nível desconhecido/outro.
    - **Ocupação**: 100% desempregados/não qualificados.
    - **Tamanho do Assentamento**: Média de 0 (todos em áreas rurais).
    - Este grupo é similar ao Cluster 2, mas com predominância de mulheres jovens, também desempregadas e com baixa renda.
- **Cluster 4: "Jovens Casados de Renda Média-Alta em Áreas Urbanas"**

    - **Tamanho**: 361 indivíduos.
    - **Idade**: Média de 28,1 anos, mediana de 28, desvio padrão de 5,12.
    - **Renda**: Média de 127.664, mediana de 118.428, desvio padrão de 30.789.
    - **Sexo**: Média de 0,57 (mistura de gêneros).
    - **Estado Civil**: 100% não solteiros.
    - **Educação**: 99,5% com ensino médio, 0,5% com universidade.
    - **Ocupação**: 82% empregados qualificados/oficiais, 15,8% gerentes/autônomos/funcionários altamente qualificados, 2,2% desempregados/não qualificados.
    - **Tamanho do Assentamento**: Média de 1,21 (cidades médias ou grandes).
    - Este grupo é formado por jovens casados, com boa renda, vivendo em áreas urbanas e trabalhando em posições qualificadas.
- **Cluster 5: "Profissionais Mais Velhos de Alta Renda e Educação Superior"**

    - **Tamanho**: Possivelmente vazio (verificação necessária).
    - **Idade**: Média de 55,9 anos, mediana de 57, desvio padrão de 10,45.
    - **Renda**: Média de 159.967, mediana de 149.410, desvio padrão de 44.328.
    - **Sexo**: Média de 0,48 (gêneros equilibrados).
    - **Estado Civil**: 31,3% solteiros, 68,7% não solteiros.
    - **Educação**: 14,5% com ensino médio, 85,5% com nível universitário.
    - **Ocupação**: 57,8% empregados qualificados/oficiais, 29,3% gerentes/autônomos/funcionários altamente qualificados, 12,9% desempregados/não qualificados.
    - **Tamanho do Assentamento**: Média de 1,15 (cidades médias ou grandes).
    - Este grupo representa profissionais mais velhos, com alta educação e renda, vivendo em áreas urbanas e ocupando posições variadas, incluindo cargos de liderança.

## Padrões Gerais

1. **Renda e Educação**:

    - Clusters com maior renda (Clusters 0, 4 e 5) têm maior proporção de empregados qualificados ou em cargos de liderança. O Cluster 5, em particular, destaca-se com renda média de 159.967 e 85,5% dos indivíduos com educação universitária.
    - Clusters com baixa renda (Clusters 2 e 3) têm apenas ensino médio e são majoritariamente desempregados/não qualificados, com rendas médias de 89.885 e 87.327, respectivamente.
2. **Idade e Estado Civil**:

    - Clusters mais jovens (Clusters 1 e 4, com médias de 29,8 e 28,1 anos) tendem a ter rendas intermediárias e maior proporção de casados/não solteiros.
    - O Cluster 5, com idade média de 55,9 anos, apresenta a maior renda e uma mistura de solteiros e não solteiros, refletindo maior maturidade profissional.
3. **Localização**:

    - Clusters com maior renda (Clusters 0, 4 e 5) estão associados a áreas urbanas, com valores médios de `Settlement size` superiores a 1 (1,33, 1,21 e 1,15, respectivamente).
    - Clusters com baixa renda (Clusters 2 e 3) estão em áreas rurais, com valores médios de `Settlement size` próximos de 0 (0,12 e 0).
4. **Sexo**:

    - Clusters 0 e 2 são homogêneos em gênero (valor 0, possivelmente masculino), enquanto o Cluster 3 é predominantemente do gênero oposto (média de 0,96, possivelmente feminino).
    - Clusters 1, 4 e 5 apresentam uma mistura de gêneros, com médias de 0,53, 0,57 e 0,48, respectivamente, indicando maior equilíbrio.

## Insights

- **Segmentação Demográfica**: O clustering revelou grupos distintos que podem ser úteis para segmentação de mercado ou políticas públicas. Por exemplo, o Cluster 5 (alta renda, educação universitária, mais velhos) pode ser um alvo para produtos/serviços premium, enquanto os Clusters 2 e 3 (baixa renda, desempregados, rurais) podem se beneficiar de programas de apoio social ou treinamento profissional.

- **Diferenças de Gênero**: Há uma clara separação de gênero entre os clusters de baixa renda: o Cluster 2 é predominantemente masculino (Sex=0), enquanto o Cluster 3 é majoritariamente feminino (Sex=0,96). Isso pode indicar desigualdades de gênero no acesso a empregos qualificados em áreas rurais, um aspecto que merece investigação mais aprofundada.

- **Educação como Fator Discriminante**: O Cluster 5 destaca-se pela alta proporção de indivíduos com educação universitária (85,5%), correlacionada com maior renda e cargos de liderança. Isso reforça a importância da educação para a mobilidade social no contexto deste dataset, enquanto os clusters com baixa renda (2 e 3) apresentam predominantemente ensino médio.

- **Sobreposição nos Clusters**: Gráficos de dispersão (idade vs. renda) e boxplots (distribuição de renda por cluster) revelaram sobreposição significativa entre os Clusters 0, 1, 2 e 4. Isso sugere que o KMeans pode não estar capturando todas as nuances dos dados, possivelmente devido à presença de outliers ou à natureza heterogênea das variáveis. Algoritmos baseados em densidade, como o DBSCAN, ou o tratamento de outliers podem melhorar a separação dos clusters.

Gráficos de dispersão (idade vs. renda) e boxplots (distribuição de renda por cluster) foram utilizados para visualizar os clusters. A sobreposição observada indica que os dados podem se beneficiar de outros algoritmos, como DBSCAN, para identificar clusters baseados em densidade e lidar com ruído.

# 7. Conclusão

A análise de clustering identificou cinco grupos distintos com base em variáveis socioeconômicas, revelando padrões claros relacionados à renda, idade, estado civil, educação, ocupação e localização geográfica. O KMeans foi eficaz para uma segmentação inicial, destacando grupos como profissionais de alta renda em áreas urbanas (Clusters 0, 4 e 5) e desempregados de baixa renda em áreas rurais (Clusters 2 e 3). No entanto, a sobreposição entre clusters e a possibilidade de o Cluster 5 estar vazio (necessitando verificação) sugerem que algoritmos baseados em densidade (como DBSCAN) ou tratamento de outliers podem melhorar os resultados. Esta análise pode ser aplicada em contextos de políticas públicas (ex.: suporte a grupos de baixa renda em áreas rurais) e estratégias de mercado (ex.: segmentação de consumidores urbanos de alta renda). A experiência reforça a importância da preparação dos dados, da análise exploratória e da avaliação de diferentes abordagens de clustering como etapas essenciais de um pipeline de aprendizado de máquina não supervisionado.

# 8. Referências


- Dataset Socioeconômico: Socioeconomic Factors and Income Dataset. Disponível em: https://www.kaggle.com/datasets/aldol07/socioeconomic-factors-and-income-dataset
- Scikit-learn Documentation: KMeans Clustering. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html