Origem dos dados:  
  
**Base dos dados:**  
  
https://basedosdados.org/dataset/562b56a3-0b01-4735-a049-eeac5681f056?table=2245875f-d1ef-490d-be29-4f8fb2191335

**Acesso via google big query**

A opção de acesso bia BQ é fornecida pela Plataforma Base dos Dados

**Sobre os dados**

SELECT

  column\_name,

  data\_type

FROM

  `basedosdados.br\_me\_caged.INFORMATION\_SCHEMA.COLUMNS`

WHERE

  table\_name = 'microdados\_movimentacao';

Tabela

Descrição gerada automaticamente com confiança média

**Interface gráfica do usuário, Aplicativo, Tabela

Descrição gerada automaticamente**

**Interface gráfica do usuário, Texto, Email

Descrição gerada automaticamente**

**Tabela

Descrição gerada automaticamente**

**Seleção dos dados**

Classificação CBO trabalhadores

As classificações oficiais de ocupação da **CBO 2002 (Classificação Brasileira de Ocupações)** podem ser encontradas diretamente no site oficial do **Ministério do Trabalho e Emprego (MTE).**

**Código CBO em:**[**https://www.mtecbo.gov.br/cbosite/pages/downloads.jsf**](https://www.mtecbo.gov.br/cbosite/pages/downloads.jsf)

**Grande Grupo 2 - ciências e artes**

**Resumo Hierárquico**

* **Grande Grupo 2**: Profissionais das Ciências e das Artes.
  + **Subgrupo Principal 21**: Profissionais das Ciências Exatas, Físicas e Engenharias.
    - **Subgrupo 212**: Profissionais da Tecnologia da Informação.
      * **Códigos 212xxx**: Engenheiros de computação, analistas de sistemas, programadores, técnicos de redes, entre outros.

Tabela

Descrição gerada automaticamente

SELECT \*

FROM `basedosdados.br\_me\_caged.microdados\_movimentacao`

WHERE cbo\_2002 LIKE '212%';

Query – DS a ser estudado

SELECT

    cbo\_2002,

    saldo\_movimentacao,

    categoria,

    grau\_instrucao,

    idade,

    raca\_cor,

    sexo,

    tipo\_movimentacao,

    salario\_mensal,

    sigla\_uf,

    ano,

    mes,

    horas\_contratuais

FROM

    `basedosdados.br\_me\_caged.microdados\_movimentacao`

WHERE

    cbo\_2002 LIKE '212%' AND

    categoria = '101' AND

    ano BETWEEN 2020 AND 2023 AND

    indicador\_aprendiz = '0' AND

    indicador\_trabalho\_parcial = '0' AND

    indicador\_trabalho\_intermitente = '0';

DS a ser estudado – Resumo

990705 – linhas

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

Criação de tabela para backup:  
CREATE OR REPLACE TABLE `rugged-airway-437623-s7.meuds.ds\_inicial`

AS

SELECT

    cbo\_2002,

    saldo\_movimentacao,

    grau\_instrucao,

    idade,

    raca\_cor,

    sexo,

    tipo\_movimentacao,

    salario\_mensal,

    sigla\_uf,

    ano,

    mes,

    horas\_contratuais

FROM

    `basedosdados.br\_me\_caged.microdados\_movimentacao`

WHERE

    cbo\_2002 LIKE '212%' AND

    categoria = '101' AND

    ano BETWEEN 2020 AND 2023 AND

    indicador\_aprendiz = '0' AND

    indicador\_trabalho\_parcial = '0' AND

    indicador\_trabalho\_intermitente = '0';

Tabela

Descrição gerada automaticamente

Dado o tamanho da tabela, escolhi baixá-la localmente após exportar para o google drive.

Interface gráfica do usuário, Tabela

Descrição gerada automaticamente

Metodologias viáveis

**Clusterização por K-Means**

**Decisão: Viável, com ajustes**

* K-Means pode ser útil para identificar **perfis** ou **grupos de trabalhadores** baseados nas variáveis demográficas (idade, gênero, raça, estado). Contudo, como o dataset não identifica indivíduos únicos, o objetivo do cluster deve ser apenas **identificar padrões gerais**, e não rastrear trajetórias de uma mesma pessoa.  
  **Pontos de Atenção:**
  + Certifique-se de padronizar as variáveis antes de aplicar o K-Means, especialmente se os valores numéricos (como idade) tiverem escalas diferentes das variáveis categóricas.
  + Avalie se faz sentido agrupar, considerando a qualidade dos clusters. Métricas como **Silhouette Score** podem ajudar.
  + Clusters poderiam revelar, por exemplo, perfis demográficos associados a maiores taxas de desligamento ou contratações em certas regiões.

**1. Análise Exploratória de Dados (AED)**

**Objetivo:** Compreender padrões gerais e identificar insights iniciais.

* Distribuições de variáveis como:
  + **Idade**: Qual faixa etária é mais contratada ou desligada?
  + **Raça e Gênero**: Existem desigualdades ou tendências específicas?
  + **Estado**: Existem regiões com maior rotatividade?
* Relações entre variáveis:
  + Cruzar idade, gênero e estado com tipo de movimentação (admissão ou desligamento).
* **Tendências Temporais**:
  + Se o dataset inclui datas, analise sazonalidade ou padrões ao longo do tempo.

Métodos:

* Histogramas, boxplots e gráficos de dispersão.
* Tabelas de contingência e heatmaps para correlação entre variáveis categóricas.

**2. Clusterização (não supervisionada)**

**Objetivo:** Identificar grupos de trabalhadores com perfis semelhantes.

**a) K-Means**

* Agrupar perfis de trabalhadores com base em variáveis como:
  + Idade, gênero, raça e localização (estado).
* Exemplo de pergunta: Quais são os perfis dominantes de trabalhadores que mais contratam ou são desligados?

**b) DBSCAN**

* Útil para identificar **outliers** ou clusters de formatos arbitrários, como grupos que têm padrões incomuns de contratação ou desligamento.

**c) Hierarchical Clustering**

* Se a quantidade de clusters não for óbvia, a clusterização hierárquica pode ajudar a entender o agrupamento em diferentes níveis.

**3. Análise Preditiva (supervisionada)**

**Objetivo:** Estimar a probabilidade de eventos (ex.: desligamento, perfil de contratação).

**a) Random Forest ou Gradient Boosting (XGBoost):**

* Predição do tipo de movimentação (admissão ou desligamento) com base nas características do trabalhador.
* Pode ser usado para:
  + Identificar os **fatores mais importantes** para admissões/desligamentos.
  + Gerar insights como: "Idade e estado influenciam mais nas admissões/desligamentos?"

**b) Regressão Logística:**

* Modelar a probabilidade de um registro ser admissão/desligamento.
* Mais simples e interpretável que Random Forest, ideal para análises explicativas.

**c) Modelos Temporais (se houver datas):**

* **Séries Temporais (ARIMA, Prophet):**
  + Prever futuras contratações/desligamentos com base em dados históricos.
  + Analisar sazonalidades, como aumento de contratações em certos períodos do ano.

**4. Técnicas de Redução de Dimensionalidade**

**Objetivo:** Reduzir complexidade do dataset e melhorar a interpretabilidade.

**a) PCA (Análise de Componentes Principais):**

* Usado antes de técnicas como clusterização para reduzir o número de variáveis.
* Ajuda a identificar combinações lineares de variáveis que explicam melhor a variabilidade dos dados.

**b) t-SNE ou UMAP:**

* Visualização de dados em 2D ou 3D para identificar padrões ocultos em conjuntos de dados complexos.

**5. Abordagens Baseadas em Regras**

**Objetivo:** Explorar relações específicas entre variáveis categóricas.

**a) Análise de Regras de Associação (Apriori ou FP-Growth):**

* Exemplo: Encontrar padrões como "trabalhadores acima de 50 anos no estado X têm maior probabilidade de serem desligados."

**6. Técnicas para Balanceamento de Dados**

Se o dataset apresentar **desbalanceamento** (por exemplo, muito mais admissões que desligamentos ou vice-versa):

* Use técnicas como **SMOTE (Synthetic Minority Oversampling Technique)** ou **undersampling** antes de aplicar modelos preditivos.

**7. Métricas de Desempenho**

Certifique-se de escolher as métricas certas para avaliação, especialmente se usar modelos preditivos:

* Para classificação: **Acurácia, F1-Score, Precision, Recall, ROC-AUC**.
* Para clusters: **Silhouette Score, Davies-Bouldin Index**.

**Análises Relevantes para o Seu Dataset**

1. **Rotatividade por Perfil Demográfico**:
   * Existe um padrão de desligamento em certos grupos de idade, raça ou gênero?
2. **Impacto Regional**:
   * Quais estados têm mais contratações ou desligamentos?
   * Como o mercado de trabalho varia por região?
3. **Tendências Temporais**:
   * Existe sazonalidade nas admissões e desligamentos?
4. **Perfis de Trabalhadores**:
   * Quais são os grupos dominantes entre contratados/desligados (com base em clusters)?
5. **Fatores de Desligamento**:
   * Quais fatores mais influenciam o desligamento, segundo os modelos preditivos?