# **PROJETO EM CIÊNCIA DE DADOS**

# **Análise de Dados Educacionais Análise do Impacto da Pandemia de COVID-19 no Perfil dos Beneficiários do Prouni (2017–2020)**

# **SUMÁRIO**

|  |  |
| --- | --- |
| **SEMESTRE** | 2025/1 |
| **PROJETO** | Análise do Impacto da Pandemia de COVID-19 no Perfil dos Beneficiários do Prouni (2017–2020) |
| **COMPONENTES DO GRUPO** | Giovani Cancherini,  Eduardo Traunig,  Vinicius Quintian,  João Pedro Fossa |
| **LINK PARA O GITHUB** | <https://github.com/GiovaniCancherini/CPAD/tree/master/Trabalho> |

## **Breve descrição do problema**

O trabalho busca investigar os principais fatores que influenciam a distribuição de bolsas do Programa Universidade para Todos (Prouni) no Brasil. Para isso, serão analisados os dados abertos do Prouni referentes aos anos de 2017 a 2020, com foco na identificação de padrões e variações no perfil dos beneficiários ao longo do tempo. O objetivo é entender como características institucionais, geográficas e sociodemográficas impactam a concessão de bolsas, considerando tanto as modalidades presenciais quanto a distância (EAD), tipos de bolsa (integral/parcial) e perfil dos estudantes atendidos.

## **Breve descrição da solução proposta**

O grupo realizará uma análise exploratória dos dados educacionais do Prouni, buscando identificar padrões recorrentes e diferenciais na distribuição de bolsas em diferentes contextos regionais e institucionais. A pesquisa também avaliará como características como sexo, raça/cor, faixa etária e presença de deficiência influenciam o acesso ao programa. Com isso, pretende-se oferecer uma visão ampla sobre o funcionamento e alcance do Prouni no período analisado.

Entregas pretendidas:

1. Limpeza e tratamento dos dados entre 2017 e 2020.
2. Análise comparativa do perfil dos beneficiários por ano.
3. Visualizações que mostrem evolução de:

- Quantidade de bolsas por curso, região e modalidade.

- Relação entre perfil do aluno (sexo, raça/cor, faixa etária, PCD) e o tipo de bolsa.

1. Verificação de possíveis impactos em 2020 com relação aos anos anteriores.
2. Relatório final com conclusões, gráficos e resposta à pergunta inicial.

## **Fases da Metodologia CRISP-DM**

|  |  |  |  |
| --- | --- | --- | --- |
| **Fase** | **Tarefas Realizadas** | **Status** | **% Concluído** |
| 1. Entendimento do Negócio | Definição do problema, objetivos e perguntas auxiliares. | Concluído | 100% |
| 2. Entendimento dos Dados | Coleta dos dados do Prouni, identificação das colunas e variáveis disponíveis. | Concluído | 100% |
| 3. Preparação dos Dados | Limpeza de dados, tratamento de valores ausentes e padronização. | Concluído | 100% |
| 4. Modelagem | Não se aplica (projeto não envolve modelos preditivos). | Não aplicável | 0% |
| 5. Avaliação | Interpretação das análises e elaboração de respostas às perguntas. | Parcialmente Concluído | 70% |
| 6. Implantação | Geração do relatório final e visualizações. | Parcialmente Concluído | 70% |

## **Resumo do que foi concluído até o momento**

**Entregas pretendidas vs. andamento atual:**

1. **Coleta dos dados – Concluída.** Foram reunidos os dados do Prouni de 2017 a 2020 em formato .csv.
2. **Levantamento das colunas – Concluído**. As colunas foram identificadas e categorizadas entre dados demográficos, institucionais e acadêmicos.
3. **Limpeza dos dados – Concluída.** A limpeza dos dados foi devidamente concluída com a padronização dos nomes das colunas, remoção de campos sensíveis (como CPF), tratamento de valores ausentes e unificação das bases dos quatro anos.
4. **Análises iniciais – Parcialmente Concluída.** Com a finalização da etapa de preparação e limpeza dos dados, foi possível dar início à análise exploratória. O nosso grupo se concentrou em investigar aspectos estruturais e demográficos da distribuição de bolsas do Prouni entre 2017 e 2020.

As primeiras análises realizadas incluíram:

- A quantidade de bolsas concedidas por modalidade de ensino (Presencial vs. EAD), com destaque para a predominância do modelo presencial em todos os anos analisados;

- A distribuição das bolsas por sexo, evidenciando que a maioria dos beneficiários são do sexo feminino;

- A distribuição das bolsas por raça declarada, com maior presença dos grupos “Pardo” e “Branco”;

- A evolução do número de bolsas ao longo dos anos, com destaque para a queda significativa em 2020;

- A distribuição por estado e região, mostrando a concentração de bolsas no Sudeste, especialmente no estado de São Paulo.

Essas análises permitiram identificar padrões e possíveis indícios de impacto da pandemia, além de traçar o perfil dos principais grupos beneficiados pelas bolsas ao longo do período.

1. **Visualizações – Parcialmente Concluída.** Com base nas análises descritas, foram geradas visualizações gráficas que facilitaram a compreensão dos dados e reforçaram as conclusões extraídas.

Entre os principais gráficos produzidos, destacam-se:

- Gráficos de barras mostrando a evolução da quantidade de bolsas por ano e modalidade de ensino;

- Gráficos de setores (pizza) para ilustrar a distribuição percentual por sexo e raça;

- Gráficos de colunas comparando a quantidade de bolsas por UF e por região;

- Gráficos comparativos entre 2019 e 2020, que evidenciam uma queda na concessão de bolsas em diversas categorias — com maior impacto no número total de beneficiários e em algumas regiões específicas.

As visualizações geradas foram fundamentais para validar as hipóteses iniciais e demonstrar, de forma clara e objetiva, as variações no perfil dos beneficiários ao longo dos anos estudados.

**Principais desafios:**

- O grande volume de dados dificultou as operações de leitura e manipulação em notebooks com menos capacidade;

- A padronização de nomes de cursos e instituições foi trabalhosa, exigindo o uso de funções automatizadas para normalização textual;

- A aplicação da metodologia CRISP-DM em um projeto descritivo exigiu uma adaptação conceitual por parte do grupo;

- A curva de aprendizado com bibliotecas como pandas e seaborn representou um obstáculo inicial para alguns membros.

**Soluções adotadas/sugeridas:**

- Divisão clara de tarefas entre os integrantes, separando responsabilidades técnicas (limpeza e código) e analíticas (análises e interpretação);

- Criação de scripts Python reutilizáveis para padronização dos dados, tornando o processo mais rápido e seguro;

- Adoção de ferramentas como pandas-profiling e matplotlib/seaborn para facilitar a geração de gráficos e sumários estatísticos;

- Utilização de boas práticas de documentação e acompanhamento entre os membros do grupo para garantir que todos acompanhassem o progresso das tarefas.

# **RELATÓRIO**

# **Compreensão dos Dados**

## Esta seção descreve a compreensão inicial sobre os dados, desde a sua coleta, passando pela exploração e identificação de problemas, até a avaliação da sua qualidade para uso no projeto.

## **Coleta dos dados**

## Os dados utilizados neste projeto foram extraídos do portal de dados abertos do Ministério da Educação (MEC), mais especificamente da base de dados do Programa Universidade para Todos (Prouni), disponível publicamente em arquivos no formato .csv. Foram considerados os anos de 2017 a 2020, com o objetivo de permitir comparações antes e durante o início da pandemia de COVID-19. A aquisição foi realizada manualmente por meio de download direto dos arquivos disponibilizados no site oficial.

## **Descrição dos dados**

## Os dados analisados estão organizados em quatro datasets distintos, um para cada ano de concessão de bolsas (2017, 2018, 2019 e 2020). Cada linha da base representa um registro de bolsa concedida a um beneficiário, contendo as seguintes informações principais:

## **Dados geográficos**: região, UF, município do beneficiário.

## **Dados institucionais**: nome da instituição (IES), nome do curso, modalidade de ensino (Presencial ou EAD), turno e tipo de bolsa (Integral ou Parcial).

## **Perfil do beneficiário**: sexo, raça/cor, data de nascimento (utilizada para calcular a faixa etária), e se a pessoa é ou não portadora de deficiência.

## Essa estrutura permite a realização de análises temporais, comparativas e descritivas sobre o perfil dos bolsistas do Prouni ao longo dos anos analisados.

## **Análise exploratória dos dados**

## Durante a etapa inicial de análise exploratória, foi possível identificar os seguintes problemas recorrentes:

## **Valores ausentes** em colunas relevantes como raça, sexo e deficiência.

## **Presença de caracteres especiais**, principalmente acentos e símbolos mal codificados, que dificultavam a leitura automatizada e a padronização dos textos.

## **Inconsistência na grafia** de nomes de cursos, instituições e municípios entre diferentes anos.

## **Colunas com baixa relevância** analítica, como CPF do aluno ou identificadores internos, que não contribuíam para os objetivos da pesquisa.

## Com base nessas observações, foi elaborado um plano de tratamento de dados envolvendo a padronização de colunas, remoção de campos desnecessários, preenchimento de valores ausentes e normalização textual. As ações específicas realizadas serão detalhadas na seção seguinte (Preparação dos Dados).

## **Verificação de qualidade dos dados**

A verificação da qualidade dos dados foi conduzida com base nos seguintes critérios:

* **Completude**: Avaliação da proporção de valores ausentes em campos-chave (ex: sexo, raça, modalidade).
* **Consistência**: Verificação de divergências de grafia e formatação entre os anos, especialmente em campos categóricos.
* **Unicidade**: Identificação de duplicatas ou registros inconsistentes (nenhum encontrado).
* **Conformidade**: Conferência dos tipos de dados e formatos esperados (ex: datas válidas, UF com siglas corretas, campos categóricos padronizados).

Como resultado, foi possível validar que, apesar das inconsistências iniciais, os dados apresentavam qualidade suficiente para a análise proposta, desde que passassem pelas etapas de tratamento previamente descritas.

# **Preparação dos Dados**

Nesta seção, serão descritas as atividades realizadas para a construção do dataset final como limpeza, criação de atributos, inserção de registros, integração de bases etc. Ao final, há uma descrição do estado do dataset que será utilizado para a modelagem.

## **Limpeza dos dados** Os dados do Prouni foram obtidos em arquivos separados por ano (2017 a 2020), contendo informações sobre os beneficiários de bolsas. A etapa de preparação teve como foco padronizar e limpar os dados para posterior análise.

## As principais ações realizadas foram:

## Padronização dos nomes das colunas (letras minúsculas e sem espaços).

## Remoção da coluna de CPF por conter informações sensíveis e desnecessárias à análise.

## Tratamento de datas de nascimento com posterior cálculo da idade dos beneficiários.

## Preenchimento de valores ausentes nas colunas de raça, sexo e deficiência com a categoria “Não informado”.

## Uniformização de texto em colunas como nome do curso, instituição e modalidade de ensino, utilizando a biblioteca unidecode para remover acentuação e padronizar comparações.

## **Criação de atributos e registros** Foi criada uma coluna chamada idade, calculada com base no ano de concessão da bolsa e a data de nascimento do beneficiário.

## **Integração de dados**

## Os dados dos quatro anos foram unificados em um único Data Frame por meio de concatenação vertical (empilhamento de registros). Os nomes das colunas foram previamente padronizados para garantir a consistência. Também foi adicionada uma coluna explícita indicando o ano da concessão (ano\_concessao\_bolsa), útil para análises comparativas ao longo do tempo.

## **Descrição do dataset final**

Após a limpeza, o dataset final contém aproximadamente 885530 registros e 25 colunas, prontos para análise. Cada linha representa um beneficiário e suas características demográficas, institucionais e acadêmicas.

# **Autocrítica**

Até o momento, o nosso grupo apresenta um bom alinhamento com a metodologia CRISP-DM, tendo completado integralmente as etapas de entendimento do negócio, compreensão dos dados e preparação. Apesar de algumas dificuldades iniciais com a padronização dos dados e o volume de registros, foi possível superá-las com uma boa divisão de tarefas e apoio técnico no uso de bibliotecas de Python como pandas e unidecode.

A ausência de experiência prévia com esse tipo de análise foi compensada pela colaboração entre os membros e pelo foco na documentação clara das etapas. A aplicação da metodologia mostrou-se adequada, mesmo que parte da equipe tenha inicialmente tido dúvidas sobre como aplicá-la num projeto sem modelagem preditiva.

**Lições aprendidas:**

* A importância da preparação e padronização dos dados para garantir consistência na análise.
* O valor da organização em grupo e da comunicação constante entre os integrantes.
* A percepção de que a análise descritiva pode oferecer grandes insights mesmo sem o uso de modelos de machine learning.

**Nota do grupo:** 9,0

**Justificativa:** A nota reflete um trabalho sólido, com entregas completas e bem estruturadas até o momento. Ainda que o nosso grupo não tenha avançado significativamente nas etapas de avaliação e visualização, o processo de preparação dos dados foi executado com responsabilidade técnica, organização e alinhamento aos objetivos do projeto. A margem para melhoria está na antecipação da análise exploratória e geração de gráficos, o que será realizado nas próximas etapas.

**Cumpriremos 100% do escopo?**

[Sim] - O grupo está tecnicamente preparado e organizacionalmente alinhado para concluir todas as etapas previstas. O entendimento do problema, a preparação dos dados e o planejamento das análises foram executados de forma satisfatória. A etapa de visualizações e interpretações será iniciada a seguir e deve ser finalizada dentro do prazo estabelecido.