# 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 |

## Breve descrição do problema

O trabalho busca responder se a pandemia de COVID-19 alterou o perfil dos beneficiários do Programa Universidade para Todos (Prouni) ou impactou a distribuição de bolsas de estudo. Para isso, serão utilizados dados abertos do Prouni entre os anos de 2017 e 2020.

## Breve descrição da solução proposta

O grupo pretende realizar uma análise exploratória dos dados educacionais do Prouni, verificando padrões e alterações no perfil dos estudantes bolsistas, especialmente no recorte entre 2019 e 2020, ano de início da pandemia.

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. | Não iniciado | 0% |
| 6. Implantação | Geração do relatório final e visualizações. | Não iniciado | 0% |

## 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 – Ainda não iniciadas,** dependem da finalização da limpeza.
5. **Visualizações – Ainda não iniciadas.**

Principais desafios:

* Volume de dados e divergências nos nomes de cursos e instituições ao longo dos anos.
* Incerteza sobre como aplicar CRISP-DM corretamente no contexto do projeto.
* Curva de aprendizado com pandas/seaborn (caso a equipe esteja com menos experiência técnica).

Soluções adotadas/sugeridas:

* Divisão de tarefas técnicas e metodológicas.
* Solicitar apoio com automação do tratamento (ex: via script Python).
* Utilização de bibliotecas como pandas-profiling ou sweetviz para EDA mais rápida.

## Autocrítica

Até o momento, o projeto segue parcialmente aderente à metodologia CRISP-DM. A definição do problema está clara, e houve um bom início com o entendimento dos dados. No entanto, a fase de preparação dos dados está atrasada em relação ao ideal para este estágio do cronograma.

Lições aprendidas:

A importância de entender bem os dados antes de tentar gerar visualizações ou insights.

A necessidade de planejamento prévio para dividir o trabalho de forma equilibrada entre os membros.

A clareza de que nem todo projeto de dados exige modelagem, o que pode ser adaptado ao CRISP-DM.

Nota do grupo:

[ ]

Justificativa: ...

Cumpriremos 100% do escopo?

[Sim] — [ JUSTIFICAR ]

# 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

Redigir uma autocrítica do trabalho até então concluído. Destacar o grau de aderência à metodologia CRISP-DM e a satisfação geral do grupo com relação ao atual progresso. Destacar também as principais lições aprendidas (sejam técnicas ou sociocomportamentais).

**OBRIGATÓRIO**: Grupo deve fazer uma autoavaliação indicando uma nota de 0,0 a 10,0 e incluindo uma justificativa.

**QUANDO EM ANDAMENTO**: Por fim, indicar se o grupo acredita que cumprirá 100% do escopo pretendido ou não, incluindo uma justificativa. **SE JÁ FINALIZADO**: indicar se o grupo concluiu ou não os 100%.