# 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. | Em andamento | 70% |
| 3. Preparação dos Dados | Limpeza de dados, tratamento de valores ausentes e padronização. | Em andamento | [xx]% |
| 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 – Em andamento. Ainda está sendo feita a padronização de categorias e verificação de dados faltantes.
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 sua coleta inicial, passando por uma análise exploratória até uma avaliação de sua qualidade.

## Coleta dos dados

Descrição de como os dados foram adquiridos. Seu contexto, fontes, métodos de aquisição e problemas encontrados.

## Descrição dos dados

Descrição de cada base de dados, incluindo cada campo, tipo de dados, unidade utilizada, códigos etc.

## Análise exploratória dos dados

Descrição da exploração inicial dos dados, incluindo objetivos de exploração. Quais operações ou métodos foram realizados, quais padrões nos dados foram encontrados, sejam esperados ou não. Conclusões em relação aos objetivos de mineração e ao que deve ser realizado na etapa de preparação dos dados (limpeza, transformação, pré-processamento).

## Verificação de qualidade dos dados

Descrição das abordagens e critérios utilizados para avaliar a qualidade dos dados **originais** e o resultado encontrado com estas avaliações.

# Preparação dos Dados

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

## Limpeza dos dados

Apresenta o raciocínio para se incluir ou remover features no dataset final, bem como o que foi realizado com dados faltantes, quais transformações foram realizadas etc.

## Criação de atributos e registros

Caso novos atributos tenham sido criados, descrever o raciocínio por trás de sua criação, metodologia ou técnicas utilizadas etc. Também descrever, caso ocorra, a criação de novos registros (instâncias de dados) a serem inseridos no dataset.

## Integração de dados

Em casos em que mais de uma base de dados tenha sido utilizada, é necessário descrever como as bases diferentes foram integradas. Quais features foram reaproveitadas? Existem features ou registros redundantes? Como a redundância foi endereçada?

## Descrição do dataset final

Descrição do dataset no estado em que será utilizado na etapa de modelagem, após a realização do pré-processamento.

# 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%.