# PROJETO EM CIÊNCIA DE DADOS

# SUMÁRIO

| SEMESTRE | 2024/2 |
| --- | --- |
| PROJETO | Análise do corpo docente |
| COMPONENTES DO GRUPO | Arthur Dewes  Jonatan Giacomini  Pedro Sangali |

## Breve descrição do problema

análise do impacto profissional sofrido pelo corpo docente antes e depois da pandemia da COVID-19. O grupo compreendeu que o problema está relacionado às mudanças nas condições de trabalho, adaptação tecnológica e percepção profissional dos professores ao longo desse período. A proposta visa identificar padrões relevantes, mudanças significativas e possíveis correlações entre os dados coletados nesse intervalo de tempo.

## Breve descrição da solução proposta

O grupo propôs realizar uma análise comparativa entre os períodos mencionados, utilizando os seguintes conjuntos de dados: IRD, DSU, AFD e IED. A solução inclui:

1- Limpeza e pré-processamento dos dados

2- Integração das bases

3- Análise exploratória e visualização

4- Testes estatísticos para verificar impacto significativo

5- Elaboração de relatório final

## Fases da Metodologia CRISP-DM

| Fase | Tarefas Realizadas |
| --- | --- |
| Entendimento de um problema existente | Definição do objetivo e escopo da análise |
| Compreensão dos Dados | Coleta e descrição inicial dos datasets |
| Preparação dos Dados | Limpeza, transformação e integração de dados |
| Avaliação | Planejamento para aplicação de testes estatísticos |
| Implementação | Geração de insights e relatório final |

## Resumo do que foi concluído até o momento

Os dados foram coletados e organizados por período (pré e pós-pandemia).

A limpeza dos dados foi concluída.

A integração das bases será feita após a limpeza de cada conjunto individual.

Ainda não foram iniciadas as análises estatísticas nem a visualização de correlações.

**Desafios enfrentados:**

* Inconsistência nos formatos dos datasets.
* Necessidade de harmonizar anos e códigos entre bases.
* Alto volume de dados exigindo filtragem criteriosa.

**Ações para superar:**

* Padronização prévia dos campos e uso de scripts de automação para limpeza.
* Definição clara de indicadores e métricas a serem comparadas.

## Autocrítica

Até o momento, o grupo manteve boa aderência à metodologia CRISP-DM, principalmente nas etapas iniciais de compreensão e preparação dos dados. A principal dificuldade tem sido a heterogeneidade dos dados entre os diferentes anos e conjuntos, o que atrasou o avanço da etapa da análise estatística.

**Lições aprendidas:**

* Importância de planejamento e padronização ao lidar com múltiplas fontes de dados.
* Necessidade de dividir tarefas para aumentar a produtividade.

**Nota atribuída pelo grupo:** 8,9

**Justificativa:** O grupo avançou com consistência nas fases iniciais, mas está um pouco atrasado na preparação dos dados. O planejamento foi adequado, mas ajustes foram necessários durante a execução.

**Expectativa de cumprimento do escopo:** Sim, o grupo acredita que conseguirá cumprir 100% do escopo planejado, desde que o cronograma seja ajustado e o esforço coletivo se mantenha.

-x-

# RELATÓRIO

# Compreensão dos Dados

## Coleta dos dados

Os dados utilizados neste projeto foram obtidos a partir de fontes institucionais e governamentais que divulgam informações sobre a educação superior no Brasil. As fontes principais incluem microdados públicos disponibilizados por órgãos oficiais (como o INEP) e dados internos disponibilizados pela instituição de ensino parceira.

Foram utilizados quatro anos de dados: 2018 e 2019 (pré-pandemia) e 2022 e 2023 (pós-pandemia). Os dados de 2020 e 2021 foram propositalmente excluídos, pois correspondem ao período mais crítico da pandemia e apresentavam alta inconsistência.

A aquisição foi realizada por meio de download direto de arquivos no formato CSV e XLSX. Entre os problemas enfrentados, destacam-se:

* Dicionários de dados incompletos ou inexistentes em algumas bases;
* Codificações diferentes para campos equivalentes;
* Dados com formatos diferentes para anos distintos;
* Campos com valores faltantes ou fora do padrão esperado.

## Descrição dos dados

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

| dataset | descrição | colunas | tipo |
| --- | --- | --- | --- |
| AFD | Adequação da formação do docente  Percentual de docentes com formação compatível por etapa e nível de ensino (g1–g5) | ano, unidade\_geografica, localizacao, dependencia\_administrativa,  EI\_g1, EI\_g2, EI\_g3, EI\_g4, EI\_g5,  EF\_total\_g1, …, EF\_total\_g5,  EF\_anos\_iniciais\_g1, …, EF\_anos\_iniciais\_g5,  EF\_anos\_finais\_g1, …, EF\_anos\_finais\_g5,  EM\_g1, …, EM\_g5,  EJAf\_g1, …, EJAf\_g5,  EJAm\_g1, …, EJAm\_g5 | **Categóricas**: unidade geográfica, localização, dependência administrativa  **Numéricas**: ano, todos os indicadores (%) |
| DSU | Percentual de docentes com curso superior por etapa de ensino | ano, unidade\_geografica, localizacao, dependencia\_administrativa,  EI\_total, EI\_creche, EI\_pre\_escola,  EF\_total, EF\_anos\_iniciais, EF\_anos\_finais,  EM, EP, EJA, ES | **Categóricas**: unidade geográfica, localização, dependência administrativa  **Numéricas**: ano, todos os percentuais (%) |
| IRD | Índice de regularidade do corpo docente, nas faixas de valor: 0–2, 2–3, 3–4 e 4–5. | ano,unidade\_geografica,localizacao,dependencia\_administrativa,baixa\_regularidade(0-|2),media\_baixa(2-|3),media\_alta(3-|4),alta(4-|5) | **Categóricas**: unidade geográfica, localização, dependência administrativa **Numéricas**: ano, valores de índice para cada faixa |
| IED | Esforço docente — percentual de docentes distribuído em seis níveis de esforço (1–6) para cada etapa de ensino. | ano,unidade\_geografica,localizacao,dependencia\_administrativa,EF\_total\_nvl1 … EF\_total\_nvl6,  EF\_anos\_iniciais\_nvl1 … EF\_anos\_iniciais\_nvl6 : EF\_anos\_finais\_nvl1 … EF\_anos\_finais\_nvl6 EM\_nvl1 … EM\_nvl6 | **Categóricas**: unidade geográfica, localização, dependência administrativa  **Numéricas**: ano, todos os percentuais em % |

## Análise exploratória dos dados

Nesta etapa inicial, exploramos quatro conjuntos de dados referentes a características do corpo docente: adequação da formação (afd), docentes com curso superior (dsu), regularidade do corpo docente (ird) e esforço docente (ied). O objetivo principal da análise foi compreender a distribuição e coerência dos dados por ano, etapa de ensino e nível geográfico, de modo a orientar o pré-processamento e as decisões de modelagem.

Foram aplicadas estatísticas descritivas, verificação de nulos e análise de distribuição.  
 Padrões observados:

* Alta concentração de docentes nos níveis intermediários de esforço e regularidade.
* Correlação positiva entre formação superior e adequação da formação.
* Diferenças regionais significativas.

Conclusão: foi necessário preparar os dados para garantir consistência e completude, com foco em limpeza e integração.

## Verificação de qualidade dos dados

Avaliamos os dados com base em:

**Completude**: registros com valores faltantes ou nulos.

**Validade**: verificar se os valores estão dentro de faixas esperadas (ex: percentuais entre 0 e 100).

**Consistência**: nomes e formatos padronizados.

**Resultado**: dados majoritariamente válidos, com necessidade de ajustes pontuais.

# 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**

* Remoção de colunas e registros com muitos nulos.
* Imputação por média regional.
* Padronização de rótulos categóricos.

### **Criação de Atributos**

* Criado índice composto de formação docente.
* Variáveis categóricas convertidas via one-hot encoding.

### **Integração**

* Bases unificadas por ano, unidade geográfica, localização e dependência.
* Redundâncias resolvidas e colunas alinhadas.

## Limpeza dos dados

Foram realizadas diversas etapas de limpeza para garantir a qualidade dos dados:

* **Remoção de colunas e registros com muitos valores nulos**, visando eliminar informações com baixa relevância e evitar distorções na análise.
* **Imputação de dados faltantes utilizando a média regional**, de forma a preservar a coerência estatística dentro de contextos geográficos semelhantes.
* **Padronização dos rótulos de variáveis categóricas**, com o objetivo de unificar representações e facilitar o tratamento computacional.

## Criação de atributos e registros

Durante o processo de preparação, novos atributos foram desenvolvidos com base em raciocínios analíticos:

* **Criação de um índice composto de formação docente**, com o intuito de agregar diversas variáveis relacionadas à qualificação dos professores em uma única métrica mais representativa.
* **Conversão de variáveis categóricas em variáveis binárias (one-hot encoding)**, permitindo que algoritmos de modelagem possam interpretá-las adequadamente.

Não foram inseridos novos registros no dataset; apenas atributos foram derivados a partir dos dados existentes.

## Integração de dados

Foram utilizadas diversas bases de dados, que passaram por um processo de integração para formar um conjunto unificado:

* **Unificação das bases por ano, unidade geográfica, localização e dependência administrativa**, assegurando a correspondência correta entre diferentes fontes.
* **Resolução de redundâncias entre colunas e registros**, com padronização de nomes de variáveis e eliminação de informações duplicadas.
* As features reaproveitadas foram aquelas que estavam presentes em múltiplas fontes e possuíam consistência semântica.

## Descrição do dataset final

Após o pré-processamento, o dataset final está estruturado da seguinte forma:

**Observações agregadas por unidade geográfica e ano**, com características associadas à infraestrutura, corpo docente, localização, dependência administrativa, entre outros.

**Atributos numéricos e categóricos devidamente tratados**, com dados faltantes imputados, rótulos padronizados e variáveis categóricas transformadas em formato compatível com algoritmos de modelagem.

**Sem registros ou colunas redundantes**, garantindo consistência e eficiência para a etapa de modelagem preditiva.

# Autocrítica

O trabalho foi bem conduzido, com aderência à metodologia CRISP-DM. Os principais desafios foram a integração das bases e o tratamento de inconsistências.

### **Nota do grupo: 9,0**

**Justificativa**: O processo foi tecnicamente sólido, mas ainda há espaço para testes de outras abordagens de pré-processamento.

**Situação atual**: O grupo acredita que concluirá 100% do escopo dentro do prazo.