# PROJETO EM CIÊNCIA DE DADOS

[Dispor esta página como capa do relatório do projeto do seu grupo. Apagar os comentários em vermelho]

# SUMÁRIO

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

## Breve descrição do problema

Descrever brevemente o problema proposto pela organização parceira – da forma como o grupo o compreendeu.

## Breve descrição da solução proposta

Descrever brevemente a solução proposta pelo grupo, enumerando claramente quais as entregas pretendidas\*.

## Fases da Metodologia CRISP-DM

Enumerar as fases do CRISP-DM e suas respectivas tarefas e indicar o percentual de conclusão de cada uma (listar pelo menos os itens da metodologia solicitados no enunciado do trabalho). Sugere-se utilizar uma tabela.

## Resumo do que foi concluído até o momento

Recuperar as entregas pretendidas\* e comentar o que foi concluído até o momento, destacando os principais desafios, as dificuldades enfrentadas e como foram/serão superadas, e o status de cada item conforme o seu planejamento.

## 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).

Dar-se (ao grupo) uma nota de 0,0 a 10,0, incluindo uma justificativa.

Por fim, indicar se o grupo acredita que cumprirá 100% do escopo pretendido ou não, incluindo uma justificativa.

-x-

# 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.

Os dados foram obtidos através do site do INEP.

Dados obtidos dos anos de: 2018,2019,2022,2023.

Fonte:

Métodos de aquisição: O próprio site do INEP forneceu os dados prontos em formato csv.

Os problemas encontrados foram os dados faltantes.

## 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%.