# PROJETO EXTENSIONISTA - COLETA, PREPARAÇÃO E ANÁLISE DE DADOS

# SUMÁRIO

| SEMESTRE | 2024/2 |
| --- | --- |
| PROJETO | Investigação sobre padrões de abstenção de votos nas eleições presidenciais. |
| COMPONENTES DO GRUPO | Cristina Einsfield  Heloysa Pelizon  Samuel Morales |

## Breve descrição do problema

O escopo escolhido para o projeto foi explorar os índices de comparecimento e abstenção das eleições presidenciais no Brasil, com foco em analisar padrões e motivos os quais levam eleitores a se abster do voto na ocasião. Para isso foram avaliados dois datasets referentes às eleições dos anos de 2018 e 2022 para serem modelados, preparados e, posteriormente, analisados.

## Breve descrição da solução proposta

Os objetivos para a solução do problema envolvem encapsular os dados disponíveis em um formato adequado para análise e descoberta de padrões. Para tanto, será feita a separação dos dados por regiões do Brasil, devido à grande quantidade de dados ao lidar com um escopo nacional. Além disso, para gerar um dataset mais conveniente para o objetivo de negócio, serão removidos dados não relevantes para o contexto e feita a conversão do restante para um formato facilmente analisável e visualizável.

## Fases da Metodologia CRISP-DM

| Fase CRISP-DM | Progresso |
| --- | --- |
| Compreensão do modelo | Concluído |
| Compreensão dos dados | Em andamento |
| Preparação dos dados | Em andamento |
| Modelagem | Pendente |
| Avaliação | Pendente |
| Aplicação | Pendente |

## 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

Os dados selecionados foram adquiridos do Portal de Dados Abertos do Tribunal Superior Eleitoral brasileiro, que é composto por coleções de diversos dados gerados ou custodiados pelo TSE. Esses dados são disponibilizados para uso livre por usuários, mantendo sempre os dados dos integrantes anônimos, e podem ser baixados localmente, assim como aplicados em redes de compartilhamento.

O maior desafio ao lidar com os datasets selecionados é a dimensão dos mesmos, uma vez que abrangem dados de escopo nacional.

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

## Descrição dos dados

Devido a essa anonimidade, assim como a dimensão nacional dos datasets escolhidos, - que faz com que eles tenham quantidades enormes de linhas - os dados dos datasets escolhidos são agrupados por grupos demográficos; ou seja,

## so um rascunho eu tinha escrito na outra secao mas acho melhor nessa depois conserto

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%.