# PROJETO EM CIÊNCIA DE DADOS

# SUMÁRIO

|  |  |
| --- | --- |
| SEMESTRE | 2024/2 |
| PROJETO | Análise de Dados Eleitorais |
| COMPONENTES DO GRUPO | Arthur Accorsi  Bruno Fernandes  Vitor Esposito |

## Breve descrição do problema

Investigar se existe uma correlação entre o nível de escolaridade dos eleitores e sua preferência por determinados partidos políticos.

## Breve descrição da solução proposta

Responder a pergunta da descrição do problema por meio de visualizações, agrupamento de dados e análise quantitativa.

## Fases da Metodologia CRISP-DM

| **Fase (Nome)** | **Porcentagem %** | **Tarefas** |
| --- | --- | --- |
| Compreens**ã**o do Neg**ó**cio | 90% | Definição do problema e das perguntas que queremos responder, entendimento do negócio/contexto e a identificação das variáveis relevantes/importantes. |
| Compreens**ã**o dos Dados | 70% | Coleta de dados, descrição e análise exploratória. |
| Preparação dos Dados | 80% | Limpeza dos dados, tratamento de valores nulos e/ou duplicados, transformações e agrupamentos necessários, e filtragem de variáveis relevantes. |

## Resumo do que foi concluído até o momento

Até o momento, realizamos a leitura dos datasets do TSE através da biblioteca do pandas, filtramos as colunas que entendemos fazer sentido para analisar o perfil dos eleitores, e fizemos alguns agrupamentos e agregações de dados com o intuito de qualificar a análise estatística. Além de que também avaliamos a presença de dados duplicados ou valores nulos. Por fim, geramos alguns gráficos scatter para visualizar a distribuição das variáveis e a relação entre elas no estado do Rio Grande do Sul.

## Autocrítica

No geral, pode-se afirmar que o grupo está no caminho certo, pois estamos seguindo e respeitando estritamente as fases da metodologia CRISP-DM. Cada etapa da metodologia é importante para que a próxima seja desenvolvida, além de que pela sua estrutura, ao encerrar a tarefa seguinte/posterior, aprendemos alguma coisa nova sobre a tarefa anterior, tornando-se um processo cíclico. A nota do grupo é 8,5, pois conforme foi citado anteriormente, estamos de fato utilizando as premissas da metodologia e estamos satisfeitos com os aprendizados, porém é a primeira vez que a utilizamos, e com toda certeza, há oportunidade para melhoria. Por fim, entendemos que sim, vamos conseguir cumprir 100% do escopo pretendido inicialmente.

-x-

# RELATÓRIO

# Compreensão dos Dados

## Coleta dos dados

As duas bases de dados foram encontradas no site do TSE.

## Descrição dos dados

## **Eleitorado 2020 - Perfil eleitor**: Uma base de dados que possui diversos campos, como faixa etária, escolaridade, estado civil, zona, seção, quantidade de eleitores daquele perfil.

Os campos de município, zona e seção são repetidas n vezes, sendo n o número de possíveis valores dos demais campos, ou seja: Porto Alegre, seção x, zona y, possui 30 eleitores da faixa etária de 30-40 anos, que são solteiros e possuem ensino superior completo. Haverá uma instância com todos os campos iguais, exceto a escolaridade, e outro com todos iguais, exceto o estado civil, e assim por diante. É uma base robusta de quase 1Gb.

Em sua maioria os campos são categóricos, exceto a quantidade de eleitores que é um dado numérico inteiro.

**Resultados 2020 - Votação em partido por município e zona**: Possui campos como a zona eleitoral, cargo de votação, partido, quantidade de votos nominais, anulados, votos para o partido, etc. É uma base mais enxuta, nos da o numero de votos para cada partido por zona eleitoral.

Em sua maioria os campos são categóricos, exceto a quantidade de votos que é um dado numérico inteiro.

## Análise exploratória dos dados

Desejamos verificar a correlação entre os graus de escolaridade dos eleitores gaúchos e suas preferências por determinados partidos políticos, e para isso vamos explorar mais a fundo as bases de dados selecionadas, “Eleitorado 2020” e “Resultados 2020” disponíveis no site do TSE. Realizamos a limpeza dos dados com um todo refletindo no encaixe do tema de pesquisa proposto, então selecionamos Features específicas para explorarmos cada base de dados:

Eleitorado 2020:  
Código do municipio, número de zona e grau de escolaridade foram os atributos selecionados, aproveitamos e agregamos somando a quantidade de eleitores (respectivamente, 'CD\_MUNICIPIO', 'NR\_ZONA', 'DS\_GRAU\_ESCOLARIDADE' e 'QT\_ELEITORES\_PERFIL').

Resultados 2020:  
Nome do partido, Código do município, nome do município e número de zona foram os atributos selecionados, agregamos somando os atributos envolvendo a quantidade de votos para gerarmos o atributo geral 'QT\_TOTAL\_VOTOS' (respectivamente, 'NM\_PARTIDO','CD\_MUNICIPIO', 'NM\_MUNICIPIO', 'NR\_ZONA',QT\_VOTOS\_LEGENDA\_VALIDOS',QT\_VOTOS\_NOM\_CONVR\_LEG\_VALIDOS','QT\_TOTAL\_VOTOS\_LEG\_VALIDOS' e 'QT\_VOTOS\_NOMINAIS\_VALIDOS').  
Nos delimitados a escolher somente as instâncias de votos que pertenciam ao primeiro turno, filtrando a base de dados.

Juntamos as duas bases de dados, agora delimitadas, com respeito do atributo de código do município e a respectiva zona eleitoral para então obtermos uma base única com os seguintes atributos:  
 'CD\_MUNICIPIO', 'NR\_ZONA' 'DS\_GRAU\_ESCOLARIDADE', 'QT\_ELEITORES\_PERFIL', 'NM\_PARTIDO', 'NM\_MUNICIPIO' e 'QT\_TOTAL\_VOTOS'.

Plotamos um gráfico de barras horizontal demonstrando a quantidade de eleitores em cada um dos graus de escolaridade presentes na base de dados (em milhões) a fim de melhorarmos o nosso entendimento dos dados. Notamos que o perfil eleitoral predominante se caracteriza pela não completude do ensino fundamental, ultrapassando os 30 milhões. Seguido desse temos o eleitor que completou o ensino médio somente, 25 milhões. Por essa mesma predominância de perfil, decidimos procurar também os municípios com maior e menor número de eleitores com ensino fundamental incompleto. Município de Porto Alegre representando a cidade com menor índice desse perfil de eleitor, 19.09%, e Arroio do Padre, representando o maior índice, de 59.44%. Usando Porto Alegre de referência descobrimos também que, do outro lado desse grau de escolaridade, o maior índice de número de eleitores com ensino superior completo está localizado em Erechim, com 17.99%.

Durante essa primeira mineração percebemos de possível padrão que, dentre os municípios gaúchos em 2020, o **maior** **índice** de eleitores com **ensino superior completo** (Erechim, **17.99%**) **é menor** que o **menor** **índice** de eleitores com **ensino** **fundamental** **incompleto** (Porto Alegre, **19.09%**).

Nossa segunda tentativa de mineração em cima dos dados foi a de cruzar essa base de dados tratada por uma base dimensão que nos fornecia a latitude e longitude por município, a partir dela criamos a visualização em forma de mapa da distribuição de eleitores que completaram o ensino superior e também os que não completaram o fundamental.

## Verificação de qualidade dos dados

Após a pré-seleção das Features e junção das bases, iniciamos a verificação da qualidade dos dados, realizando a busca por dados nulos entre os atributos escolhidos, além da procura por dados duplicados após as agregações e criação de novas Features, também buscamos por alguma instância com total de votos negativos. Dentre as buscas feitas, não detectamos nenhum atributo com valor nulo, nem votos que resultavam em um número negativo, porém detectamos 17 duplicatas, causadas pois estávamos olhando para instâncias que anterior a agregação coincidiam em partido, região e quantidade de votos (só diferindo pelo tipo de voto, entre vereador e prefeito). Por conta disso decidimos realizar a análise a partir daí olhando somente para os casos em que o cargo concorrido é de vereador. Essas verificações foram importantes pois nos auxiliaram com a tomada de decisão sobre quais filtros e cortes deveríamos realizar em prol da análise do nosso trabalho.

Planejamento e conclusão quanto ao andamento do projeto até então:  
Reformular a exploração para a utilização dos dados na próxima etapa, visto que, mesmo que tenha sido possível a criação de várias análises exploratórias, a falta de uma conexão real ainda presente entre as bases de dados Eleitorado 2020 e Resultados 2020 ainda serve de obstáculo para uma conclusão mais direta quanto a preferência por determinados partidos.

# 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%.