PROJETO EXTENSIONISTA - COLETA, PREPARAÇÃO E ANÁLISE DE DADOS

SUMÁRIO

SEMESTRE	2024/2
PROJETO	Investigação sobre padrões de abstenção de votos nas eleições presidenciais mais recentes - 2018 e 2022
COMPONENTES DO GRUPO	Cristina Einsfield Heloysa Pelizon Samuel Morales

Breve descrição do problema

O tema escolhido para o projeto foi a exploração dos índices de comparecimento e abstenção das eleições presidenciais no Brasil, com foco em analisar características associadas aos eleitores que se abstiveram 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 uma filtragem de dados para apenas aqueles que representam o estado do Rio Grande do Sul, devido à grande massividade do dataset ao lidar com um escopo nacional. Além disso, para gerar um conjunto de exemplares 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	Concluído
Preparação dos dados	Concluído
Modelagem	Concluído
Avaliação	Concluído
Aplicação	A definir

Resumo do que foi concluído até o momento

Até o momento, o grupo realizou pequenas análises sobre o comportamento de indivíduos com relação ao comparecimento nas urnas das eleições de 2018 e 2022 de acordo com suas características sociais. Foram analisadas as frequências de cada categoria existente nas classes gênero, grau de escolaridade, faixa etária e estado civil para os datasets completos, a fim de auxiliar a tomada de decisão sobre o escopo do projeto, tendo em vista a necessidade de reduzir a quantidade de dados trabalhados.
Conforme o planejamento, a fim de melhorar a qualidade dos dados (que já se encontram em um estado considerado adequado), os próximos passos englobam retirar valores inadequados para as análises, realizar agrupamento de categorias para colunas de número muito grande de classificações e reduzir a dimensionalidade perante a remoção de variáveis não úteis para o projeto.
Ademais, posteriormente ao cenário de pré-processamento inicial dos dados, o objetivo do grupo é nichar ainda mais as observações, separando os municípios rio-grandenses pelo estado de presença na região metropolitana do RS. Desta forma, haverá a comparação de grupos sociais dos exemplares com comparecimento ou não nas eleições com destaque aos vieses de localidade urbana/não-urbana, permitindo uma análise mais precisa do escopo planejado do projeto.

Autocrítica

Um dos maiores desafios para o grupo foi lidar com conjuntos de dados tão robustos em questão de manipulação nas máquinas dos componentes. O grupo avalia a própria performance em um nível de 8,5/10. Além disso, a escolha de nicho para as análises também foi algo reflexivo. Por fim, a estrutura de grupo também demandou boa comunicação e organização dos integrantes.

-x-

RELATÓRIO

1. Compreensão dos Dados

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, assim como a diferença entre os atributos de cada um.

Descrição dos dados

Primeiramente, é necessário notar que os datasets estão distribuídos de forma diferente: a versão de 2018 é uma pasta que contém arquivos de cada Unidade Federativa, enquanto a de 2022 abrange todo o território nacional em um arquivo só. Originalmente, o dataset de 2018 era composto por 636.929 linhas e 35 colunas, enquanto o de 2022 continha 1.048.576 linhas e 23 colunas, ambos abrangendo diversos aspectos de grupos de eleitores de todas as zonas eleitorais brasileiras com eleitores. Vale notar que devido a anonimidade vetada a eleitores em relação a seus dados pessoais, os dados dos datasets escolhidos são agrupados por grupos demográficos; ou seja, cada linha representa um grupo de indivíduos com características similares. Ademais, o dataset de 2018 tem atributos que faltam no de 2022, então, para fins deste trabalho, estas serão removidas do nosso escopo, mantendo apenas as colunas em comum.

Assim, são descritas em ambos as que seguem:

- **DT_GERACAO**: Data de geração do registro.
- **HR_GERACAO**: Hora de geração do registro.
- **ANO_ELEICAO**: Ano de ocorrência da eleição em questão.
- **NR_TURNO**: Número do turno - pode ser '1' ou '2'.
- **SG_UF**: Sigla do estado da federação da zona eleitoral.
- **CD_MUNICIPIO**: Código do município.
- **NM_MUNICIPIO**: Nome do município.
- **NR_ZONA**: Número da zona eleitoral.
- **CD_GENERO**: Código do gênero - 2 representa masculino, 4 feminino, e 0 não informado.
- **DS_GENERO**: Descrição do gênero - 'MASCULINO', 'FEMININO' ou 'NÃO INFORMADO'.
- **CD_ESTADO_CIVIL**: Código do estado civil - 1 representa solteiro, 3 casado, 5 viúvo, 7 separado judicialmente, 9 divorciado e, por fim, 0 representa não informado.
- **DS_ESTADO_CIVIL**: Descrição do estado civil - 'SOLTEIRO', 'CASADO', 'VIÚVO', 'SEPARADO JUDICIALMENTE', 'DIVORCIADO' ou 'NÃO INFORMADO'.
- **CD_FAIXA_ETARIA**: Código da faixa etária - o código é representado por um número de 4 dígitos contemplando a faixa etária do agrupamento em questão, ou seja, a faixa '21 a 24 anos' é representada pelo código 2124. Há no dataset também grupos que englobam apenas uma idade, que são dos eleitores entre 16 e 20 anos; nesse caso, a faixa de 18 anos tem o código 1800. A faixa etária de a partir de 100 anos de idade tem um código especial, que é 9999.
- **DS_FAIXA_ETARIA**: Descrição da faixa etária, são agrupados de cinco em cinco anos, exceto os eleitores de até 20 anos e com a partir de 100 anos - '16 anos', '17 anos', '18 anos', '19 anos', '20

anos', '21 a 24 anos', '25 a 29 anos', '30 a 34 anos', '35 a 39 anos', '40 a 44 anos', '45 a 49 anos', '50 a 54 anos', '55 a 59 anos', '60 a 64 anos', '65 a 69 anos', '70 a 74 anos', '75 a 79 anos', '90 a 94 anos', '95 a 99 anos', ou '100 anos ou mais'.

- **CD_GRAU_ESCOLARIDADE**: Código do grau de escolaridade - 1 representa analfabeto, 2 lê e escreve, 3 ensino fundamental incompleto, 4 ensino fundamental completo, 5 ensino médio incompleto, 6 ensino médio completo, 7 ensino superior incompleto, 8 ensino superior completo, e, por fim, 0 representa não informado.
- **DS_GRAU_ESCOLARIDADE**: Descrição do grau de escolaridade - 'ANALFABETO', 'LÊ E ESCREVE', 'ENSINO FUNDAMENTAL INCOMPLETO', 'ENSINO FUNDAMENTAL COMPLETO', 'ENSINO MÉDIO INCOMPLETO', 'ENSINO MÉDIO COMPLETO', 'ENSINO SUPERIOR INCOMPLETO', 'ENSINO SUPERIOR COMPLETO' ou 'NÃO INFORMADO'.
- **QT_APTOS**: Quantidade de pessoas aptas por agrupamento, ou seja, representa o total de pessoas no agrupamento.
- **QT_COMPARECIMENTO**: Quantidade de comparecimentos no agrupamento.
- **QT_ABSTENCAO**: Quantidade total de abstenções no agrupamento. O padrão é que esse campo e 'QT_COMPARECIMENTO' somados representem 'QT_APTOS'.
- **QT_COMPARECIMENTO_DEFICIENCIA**: Quantidade de comparecimentos de pessoas com deficiência.
- **QT_ABSTENCAO_DEFICIENCIA**: Quantidade de abstenções de pessoas com deficiência.
- **QT_COMPARECIMENTO_ITINERANTE**: Quantidade de comparecimentos itinerantes.
- **QT_ABSTENCAO_ITINERANTE**: Quantidade de abstenções itinerantes.

Análise exploratória dos dados

Nesta análise exploratória dos dados de abstenção de voto, planejamos identificar padrões demográficos e regionais que possam influenciar o comportamento eleitoral. A primeira etapa será observar as características principais dos eleitores, como faixa etária, gênero, grau de escolaridade e estado civil, para entender como cada grupo se comporta em termos de abstenção. Em seguida, analisaremos as diferenças entre áreas metropolitanas e não-metropolitanas para verificar se há variações na participação eleitoral conforme a localização.

Além disso, pretendemos avaliar as relações entre nível educacional e idade com as taxas de abstenção, partindo da hipótese de que eleitores mais velhos e com menor escolaridade tendem a se abster em maior número. Observaremos também se há discrepâncias entre gêneros e estados civis que possam influenciar o comportamento eleitoral.

Para garantir a qualidade das análises, será essencial revisar a consistência dos dados, confirmando que o total de eleitores aptos corresponde à soma de comparecimentos e abstenções em cada registro. Com essa análise inicial, esperamos identificar padrões e tendências que servirão como base para a etapa de preparação de dados e modelagem preditiva, aprofundando o entendimento dos fatores que contribuem para a abstenção no processo eleitoral.

Verificação de qualidade dos dados

Na verificação de qualidade dos dados, foram observadas várias colunas com valores como #NE (não especificado) ou códigos de preenchimento como -3, indicando ausência de dados relevantes para as análises (especialmente em atributos como DS_IDIOMA_INDIGENA e DS_GRUPO_INDIGENA).

Além disso, algumas colunas, como QT_COMPARECIMENTO, QT_ABSTENCAO, QT_APTOS, apresentam dados numéricos que necessitam ser verificados para identificar valores atípicos (outliers) ou inconsistências que possam afetar a análise, como somas incorretas entre comparecimentos e abstenções. Registros com valores de QT_APTOS que somados não correspondem ao valor de QT_COMPARECIMENTO e QT_ABSTENCAO serão sinalizados para correção ou exclusão, conforme apropriado.

Por fim, é necessário verificar a presença de valores faltantes em colunas essenciais para a análise (como SG_UF, NM_MUNICIPIO, DS_GENERO, DS_ESTADO_CIVIL), assegurando que esses atributos estejam preenchidos para manter a integridade dos dados demográficos.

2. Preparação dos dados

Na preparação dos dados, algumas atividades essenciais foram realizadas para assegurar a integridade e usabilidade do conjunto de dados:

- **Remoção de colunas irrelevantes:** Colunas como DT_GERACAO, HH_GERACAO, QT_COMPARECIMENTO_DEFICIENCIA, QT_ABSTENCAO_DEFICIENCIA, QT_COMPARECIMENTO_ITINERANTE CD_COR_RACA, , DS_COR_RACA, CD_QUILOMBOLA, DS_QUILOMBOLA, CD_INTERPRETE_LIBRAS, DS_INTERPRETE_LIBRAS, CD_IDENTIDADE_GENERO, DS_IDENTIDADE_GENERO, CD_IDIOMA_INDIGENA, DS_IDIOMA_INDIGENA e QT_ABSTENCAO_ITINERANTE foram descartadas por não fornecerem informações diretamente relevantes para os objetivos da análise.
- **Tratamento de dados ausentes ou inconsistentes:** Linhas com valores #NE ou preenchimentos inválidos em atributos críticos foram removidas ou corrigidas, quando possível.
- **Conversão e normalização de valores:** Valores categóricos foram convertidos para códigos padronizados, e colunas numéricas foram revisadas para normalização e verificação de valores extremos ou atípicos.

Limpeza dos dados

A limpeza dos dados para a construção de um dataset finalista para as análises consistiu, após a avaliação, em redução de dimensionalidade a partir da exclusão de colunas consideradas pouco ou nada informativas para o objetivo do projeto e filtragem de dados com a retirada de exemplares compostos de dados inválidos. A escolha pela filtragem e não inserção foi a estratégia inicial para visualização do tamanho posterior do conjunto, que devido a sua escala não passou a adquirir poucos exemplares. Apesar do citado até então, a etapa de limpeza dos dados pode ser revista em prol das análises futuramente.

As colunas retiradas nos dois conjuntos de dados selecionados para o projeto são:

- DT_GERACAO;
- HR_GERACAO;

- QT_COMPARECIMENTO_DEFICIENCIA;
- QT_ABSTENCAO_DEFICIENCIA;
- QT_COMPARECIMENTO_ITINERANTE;
- QT_ABSTENCAO_ITINERANTE.

Os exemplares elegidos para remoção foram, por ora, foram os obtentores de valores 0,-3 ou -1 nas colunas de características socioculturais codificadas.

- CD_GENERO;
- DS_GENERO;
- CD_ESTADO_CIVIL;
- DS_ESTADO_CIVIL;
- CD_FAIXA_ETARIA;
- DS_FAIXA_ETARIA;
- CD_GRAU_ESCOLARIDADE;
- DS_GRAU_ESCOLARIDADE.

Criação de atributos e registros

Um novo atributo categórico foi criado para indicar se o município pertence a uma região metropolitana ou não, visando uma análise comparativa de abstenções e comparecimentos entre essas áreas. Esse atributo foi incluído com base em dados adicionais de classificação dos municípios, o que permite uma segmentação mais refinada.

Durante a avaliação da proposta do projeto, foi decidido que o escopo deste mudaria de âmbito nacional para âmbito regional, mais especificamente para o estado do Rio Grande do Sul. Sendo assim, a fim de tornar a observação dos dados mais concisa e detalhada, foi escolhida também a temática de comparação de abstenções e comparecimentos nas eleições de 2018 e 2022 de indivíduos votantes em cidades de região metropolitana ou não-região metropolitana. Por este motivo, a inserção de uma coluna categórica em função de informar o estado da cidade perante o critério anterior foi considerada necessária. O raciocínio intrínseco

Integração de dados

Para integrar os datasets de 2018 e 2022, foram selecionadas apenas as colunas comuns a ambos os anos, descartando os atributos exclusivos de um único ano. Isso garante que as análises comparativas entre esses períodos sejam consistentes.

A integração foi realizada por meio de uma concatenação das tabelas, após assegurar a padronização dos nomes e tipos das colunas em ambas as bases.

Descrição do dataset final

O dataset final ideal para analisar e identificar padrões de abstenção de voto contém registros agregados por município, faixa etária, gênero, estado civil e grau de escolaridade, com colunas que incluem o total de eleitores aptos, o número de comparecimentos, e abstenções. Adicionalmente, contém uma coluna que identifica se o município pertence a uma região metropolitana, permitindo comparações entre áreas urbanas e rurais. O dataset foi filtrado para remover registros com dados ausentes ou inconsistentes, e

foram padronizadas as codificações para atributos categóricos, como gênero e estado civil. Esse conjunto de dados, limpo e padronizado, oferece uma base sólida para análise exploratória e modelagem, facilitando a descoberta de padrões demográficos e regionais que influenciam o comportamento de abstenção nas eleições, além de possibilitar insights sobre a influência de fatores como idade e escolaridade na decisão de comparecer às urnas.

Conclusões obtidas após análises

Após as análises finais, podemos concluir que os datasets seguiram o mesmo padrão de crescimento onde analfabetos continuam sendo o grupo que mais se absteve diminuindo a taxa de acordo com o grau de escolaridade,juntamente com os votantes com faixa etária mais elevada, porém, ao analisar a porcentagem de abstenção de cada dataset, vemos um aumento de quase 1,5% de 2018 para 2022, concluindo que a eleição de 2022 foi mais difícil para os eleitores votarem, ocasionando na abstenção. Inicialmente tínhamos a ideia de que progressivamente a taxa de abstenção iria diminuir, pois as campanhas governamentais e midiáticas influenciam a população a votar, foi visto que juntamente da faixa etária elevada, no campo de estado civil, os viúvos lideram como grupo que mais se absteve, podendo fazer relação com a faixa etária pois o número de viúvos aumenta progressivamente de acordo com a idade.

3. Autocrítica

O grupo enfrentou desafios técnicos e organizacionais significativos, especialmente no manuseio de grandes volumes de dados, o que exigiu uma adaptação das máquinas dos integrantes. Além disso, a definição do nicho para análise foi um processo reflexivo que envolveu um olhar crítico e escolhas cuidadosas. A estrutura do grupo, focada na comunicação e organização, foi essencial para o progresso.

Nas fases da metodologia CRISP-DM, o grupo começou pelo entendimento do negócio, na etapa 1 do projeto, definindo os objetivos e conhecendo as necessidades específicas para orientar o estudo. Em seguida, no entendimento dos dados, houve uma análise inicial para familiarizar-se com as características do conjunto de dados e identificar potenciais limitações e oportunidades. A fase de preparação de dados envolveu a limpeza e transformação, o que foi crucial para manter a qualidade e integridade dos dados antes da modelagem. Por fim, a equipe iniciou a fase de modelagem, explorando as técnicas mais apropriadas para o tipo de análise desejada, seguindo para a avaliação e ajustes conforme necessário.

Atribuímos uma nota de 8,5 ao trabalho atual, reconhecendo o esforço e o aprendizado técnico e comportamental. Acreditamos que o projeto eventualmente atingirá 100% do escopo inicial, com algumas adaptações para otimizar a execução final, mantendo o foco na qualidade das entregas.