PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2024/2
PROJETO	Caminhos para Prefeitura
COMPONENTES DO GRUPO	Guilherme Ryan Domingos Nunes Alex Sandro Silva Barbosa Matheus Seibt Thiago dos Santos Camargo

Breve descrição do problema

Buscamos entender o comportamento eleitoral e o engajamento dos eleitores em diferentes estados e faixa etárias, analisando as profissões, participação dos mesários e o se perguntado qual profissão seria ideal para ter o sucesso em uma eleição.

PERGUNTAS:

→CANDIDATOS:

- Quais estados têm apresentado maior crescimento no número de eleitores? Qual estado tem mais peso para eleger?

  Justificativa: O dataset de candidatos contém informações sobre o estado de origem dos candidatos e a quantidade de eleitores associados a cada estado.
  Com esses dados, é possível calcular a evolução do número de eleitores em cada estado ao longo do tempo, permitindo identificar aqueles que apresentam maior crescimento.

- O quanto as profissões influenciam para ter sucesso eleitoral?

  Justificativa: O dataset de candidatos inclui a profissão de cada candidato e os resultados das eleições.
  Ao cruzar a profissão com o sucesso nas urnas, será possível identificar correlações entre certas profissões e as chances de ser eleito

➔ RESULTADO:
Integração do DataBase CANDIDATOS para o RESULTADO, trazendo a perspectiva de quanto a profissão influência nos resultados da eleição.

Breve descrição da solução proposta

Análise detalhada dos dados dos Candidatos e dos Resultados, com o objetivo de responder às questões levantadas e fornecer insights valiosos para a organização parceira. O processo utilizará ferramentas de análise de dados e visualizações para identificar padrões e tendências relevantes.

Fases da Metodologia CRISP-DM

Fase	Tarefas	Conclusão
Compreensão do Negócio	- Identificar os objetivos do projeto e as necessidades da organização parceira - Definir perguntas principais	100%
Compreensão dos Dados	- Coletar e explorar os dados fornecidos - Identificar atributos relevantes para análise	100%
Preparação dos Dados	- Limpeza e formatação dos dados - Seleção das variáveis a serem analisadas	100%
Modelagem	- Gráficos gerados no Power BI, criação de Dashboards para a análise.	100%
Avaliação	- Decidimos fazer somente uma avaliação, juntando os aspectos de todos os integrantes da equipe.	100%
Implementação	- Trabalho finalizado	100%

Resumo do que foi concluído até o momento

Realizamos uma análise inicial das bases de dados dos Candidatos e dos Resultados, com foco na profissão dos prefeitos e na probabilidade de sucesso para cada perfil profissional. A partir dessa análise, definimos perguntas-chave para guiar nossa investigação e identificar os principais insights. Também discutimos os formatos ideais para gráficos e painéis, além das informações essenciais para uma visualização clara e objetiva. Nosso objetivo é que essas representações visuais comuniquem de forma eficaz os insights relevantes obtidos durante a análise.

Autocrítica

Até o momento, nosso grupo seguiu uma abordagem estruturada para a análise das bases de dados de Candidatos e Resultados, com o objetivo de gerar insights úteis para a organização parceira.

Verificamos as duas bases de dados, fizemos a compreensão, focamos mais nos atributos da base CANDIDATO e RESULTADO:

-NR_CANDIDATO(Numero de urna do candidato)

-NM_CANDIDATO(Nome de candidato pois somente NR não é o suficiente pois pode (e há)) repetição.

-CD_CARGO(Prefeito possui o código = 11)

CANDIDATO:

-CD_GRAU_INSTRUCAO(Caso seja util para mais perguntas)

-DS_OCUPACAO(Descrição para responder as profissões que influenciam)

-CD OCUPACAO(Existem de 2000 a 2020 possui 280 tipos de ocupações)

-CD_GENERO(Caso haja mais perguntas a serem feitas)

RESULTADO:

-DS_SIT_TOT_TURNO(De 2000 a 2020 o código de eleito mudou, então utilizamos DS ao invés de DC)

RELATÓRIO

1. Compreensão do Negócio

Nesta seção, apresentamos a compreensão do grupo sobre os objetivos e os requisitos do projeto a partir de uma perspectiva de negócio. O foco é identificar padrões que possam ajudar a prever o sucesso eleitoral, analisando a influência de diferentes profissões sobre as chances de um candidato se eleger, especificamente para o cargo de prefeito.

Para alcançar esse objetivo, selecionamos duas bases de dados principais: Candidatos e Resultados. A base de Candidatos fornece informações detalhadas sobre o perfil de cada candidato, incluindo a profissão, enquanto a base de Resultados nos permite analisar os dados de vitória e derrota nas eleições. Nosso plano é cruzar essas informações para verificar se existe uma correlação significativa entre a profissão do candidato e a sua probabilidade de ser eleito.

Background

Este projeto está inserido na área de análise eleitoral e política, com o objetivo de explorar fatores que podem influenciar o sucesso de candidatos em eleições municipais, especificamente para o cargo de prefeito. Problema na quantidade de prefeitos por estado, tendo em média 0.6 a 0,92 de erros por estado durante os anos de 2000 a 2020.

A mineração de dados surge como uma solução adequada porque permite identificar padrões e correlações entre diversas variáveis presentes nas bases de dados. No caso deste projeto, a mineração de dados possibilita analisar o impacto da profissão dos candidatos, cruzando essa informação com os resultados das eleições para verificar se existem perfis profissionais que aumentam a probabilidade de vitória. Essa abordagem analítica é útil para revelar tendências que podem não ser evidentes à primeira vista e que podem ser aplicadas para entender melhor o comportamento eleitoral.

Objetivos de negócio e critérios de sucesso

O projeto visa identificar o impacto da profissão dos candidatos sobre suas chances de eleição para o cargo de prefeito, gerando insights que ajudem a prever padrões de sucesso eleitoral. Com esses insights, espera-se que partidos, consultores políticos e candidatos possam orientar suas estratégias de campanha de maneira mais eficiente, entendendo melhor o perfil profissional que tende a ser bem-sucedido nas urnas.

Com os dados auxiliam a identificar profissões específicas que estão mais associadas ao sucesso eleitoral, com o intuito de determinar se o perfil profissional do candidato impacta suas chances de vitória.

Critério de Sucesso: Estabelecer uma correlação estatisticamente significativa (nível de confiança de 95% ou mais) entre determinadas profissões e as probabilidades de eleição para prefeito.

Inventário de recursos

Foi utilizado a Base de Dados de Candidatos que contém informações com mais detalhes sobre os candidatos, incluindo profissão, escolaridade, estado civil, cor, entre outros. E na Base de Dados de Resultados que inclui o resultado das eleições passadas para o cargo de prefeito.

Por enquanto utilizamos somente Python para a organização dos dados e conhecer mais sobre cada base de dados.

Para a comunicação do Grupo, utilizamos o Microsoft Teams, onde decidimos as tarefas de cada integrante e discutimos sobre o insight a ser passado.

Requisitos, suposições e restrições

Elaboração de um relatório para o quanto a profissão do candidato influencia nas chances de ser eleito como prefeito.

Os dados são antigos, logo não terá alteração, são sólidos, porém existem alguns erros como a quantidade de prefeitos por estado, porém isso se verá no arquivo CSV gerado.

Terminologia

- Candidato: Pessoa que se inscreveu oficialmente para disputar uma posição em uma eleição. Neste projeto, estamos focando nos candidatos ao cargo de prefeito.
- Prefeito: Chefe do poder executivo de um município, responsável pela administração da cidade e pela implementação de políticas públicas locais.
- Profissão: Ocupação ou trabalho principal do candidato, que pode influenciar sua imagem pública e suas chances de eleição.
- Base de Dados de Candidatos: Conjunto de informações sobre os candidatos, incluindo profissão, idade, gênero e outros dados pessoais.
- Base de Dados de Resultados Eleitorais: Registro dos resultados das eleições passadas, indicando quais candidatos foram eleitos e quais não foram.

Objetivos de mineração e critérios de sucesso

O grupo teve como objetivo descobrir se existe uma relação significativa entre a profissão dos candidatos a prefeito e suas chances de vitória.
Ainda em processo a identificação em uma correlação significativa sobre a profissão e a probabilidade do sucesso eleitoral. Logo desenvolveremos gráficos para apresentar insights sobre a profissão e suas chances de vitória de maneira mais visual.

Plano de Projeto

1) Extrair dos dados das bases e criar um csv contendo apenas as informações dos candidatos que estão presente como eleitos no dataset resultado.
2) A partir do CSV criado, vamos relacionar os 280 tipos de profissões presentes no arquivo com cada estado e ano, a fim de responder à seguinte pergunta: a profissão influencia o sucesso nas eleições?

3) Com power bi vamos gerar graficos sobre o tema.

Avaliação inicial de técnicas e ferramentas

Python foi escolhido como a linguagem principal para manipulação e análise dos dados devido à sua flexibilidade e excelente bibliotecas de dados. Utilizamos especialmente a biblioteca Pandas para extrair, limpar e manipular as bases de dados, permitindo uma preparação eficiente e uma análise detalhada.

Para a criação de gráficos e dashboards, optamos pelo Power BI. Com o Power BI, pretendemos tornar os insights mais claros e acessíveis, possibilitando uma comunicação visual que destaque as correlações entre profissão e sucesso eleitoral. Essa ferramenta permitirá que as conclusões do projeto sejam apresentadas de forma informativa e precisa.

2. Compreensão dos Dados

Coleta dos dados

Os dados foram adquiridos a partir dos Datasets de candidatos e resultados eleitorais, abrangendo os anos de 2000 a 2020 em intervalos de 4 anos. A partir dessas bases, extraímos informações relevantes como número de urna, nome do candidato, situação de candidatura e código do cargo. Em seguida, criamos um Dataframe específico para identificar os vencedores.

Utilizamos um laço para filtrar os vencedores em cada ano (2000, 2004, 2008, 2012, 2016 e 2020), considerando as siglas dos municípios e os anos como parâmetros para buscar, dentro do Dataset de candidatos, aqueles que foram eleitos.

Descrição dos dados

1) De Candidato e Resultado extraimos:

   -NM_CANDIDATO(Somente número de urna não relevante, devemos comparar com nome pois o numero pode ser repetido entre candidatos)

   -NR_CANDIDATO(Numero de urna para criar relação entre ambos csv)

   CD_CARGO(Para filtrar somente prefeito que no caso é 11)

2) De Candidato extraimos:

   -DS_OCUPACAO(Para sabermos qual o seu trabalho)

   -DC_OCUPACAO(para faciliar a busca de ocupação)

-CD/DS_GRAU_INTRUCAO(Deixai para caso seja relevante no futuro saber seu nível escolar/academico)

-CD_GENERO(Deixei para caso precisemos, é sempre relevante ter informações desse tipo)

-DS_UE: município

3) De Resultado extraímos:

-DS_SIT_TOT_TURNO(Para sabermos se foi eleito. Poderiamos usar cd, porém de 2000 a 2020 os cd foram mudando os codigos então usamos descrição pois nunca muda.)

Análise exploratória dos dados

Criamos um método chamado **criar_csv**, que gera um arquivo meus-candidatos_vencedores.csv, onde armazenamos os dados dos candidatos vencedores com base no Dataset de resultados eleitorais. No entanto, identificamos uma discrepância na contagem de candidatos vencedores, que deveria corresponder ao número de municípios em cada estado. Essa diferença ficou, em média, dentro de uma variação de até 10 registros para mais ou para menos.

Essa variação se deve, em parte, às mudanças no número de municípios ao longo dos anos, dificultando o acesso a uma referência exata para o período analisado. Para monitorar esse desvio, desenvolvemos um método que calcula a média de inconsistências por ano. Observamos que apenas em 2016 a variação ultrapassou o limite aceitável de 0,90%, motivo pelo qual planejamos um tratamento específico para os dados deste ano.

Na preparação dos dados, também identificamos a necessidade de normalizar as ocupações dos vencedores por ano e estado, para facilitar a análise da relação entre a profissão e o sucesso eleitoral.

Verificação de qualidade dos dados

Considerando que se trata de dados eleitorais, a fidelidade e a qualidade dos dados são, em geral, elevadas e satisfatórias para análise. Embora tenhamos encontrado algumas divergências na situação das eleições nos resultados entre 2000 e 2020, não identificamos inconsistências graves que comprometam a integridade dos dados.

Em relação ao CSV criado com os candidatos vencedores, a qualidade é igualmente boa, com a necessidade apenas de uma limpeza adicional nos dados de 2016, devido à pequena variação observada. Não identificamos valores nulos no Dataset; isso foi confirmado com um código que percorreu todo o CSV para verificar a presença de valores null e #NULO, que são padronizações aplicadas pelo TSE.

3. 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

Não houve dados faltantes nos filtros que fizemos em vencedores e candidatos(retornando cd_cargo=11), então limpeza não será necessária muito menos retirar dados repetidos pois usamos drop_duplicate para retira lós durante a filtragem. Adicionamos tudo em um novo CSV os vencedores com os dados de candidatos.

Já no power bi nós alteramos a data, para ficar no padrão, adicionamos localidade.

Criação de atributos e registros

Enquanto utilizavamos o método **criar_csv** fomos adicionando dados em duas colunas que criamos chamada ano e sigla para facilitar nossas consultas.

Integração de dados

Foi necessário combinar dados de duas bases: a base de **Resultado**, onde utilizamos as colunas de **cd_cargo** e situação de candidatura para identificar os candidatos eleitos para o cargo de prefeito (categoria 11) e distinguir os vencedores, e a base de **Candidato**, que contém informações detalhadas de todos os candidatos, incluindo a ocupação antes da eleição.

Inicialmente, filtramos os dados de **Resultado** para obter o número de urna e o nome dos candidatos eleitos. Em seguida, usamos essas informações para fazer a correspondência com a base de **Candidato**, retornando os dados relevantes, especialmente a ocupação dos vencedores.

Mantivemos também colunas adicionais, como gênero, escolaridade e cargo — sendo este último redundante, já que todos são prefeitos —, para oferecer maior flexibilidade no caso de novas perguntas ou análises exploratórias.

Descrição do dataset final

Após modelar as bases de dados selecionadas, criamos um CSV unificado contendo as principais colunas para nossa análise: NM_CANDIDATO (nome do candidato), NR_CANDIDATO (número de urna), CD_CARGO e DS_CARGO (código e descrição do cargo), CD_GRAU_INSTRUCAO e DS_GRAU_INSTRUCAO (código e descrição do nível de escolaridade), CD_OCUPACAO e DS_OCUPACAO (código e descrição da ocupação anterior), além da SIGLA e ANO para orientar a filtragem e busca. Esse arquivo consolidado permite uma análise integrada e facilita o cruzamento de informações para responder à nossa questão sobre a influência da profissão no sucesso eleitoral dos candidatos.

4. Modelagem

O objetivo desta etapa foi compreender se há uma relação significativa entre a profissão dos candidatos e suas chances de vitória nas eleições para prefeito, utilizando técnicas analíticas e visualizações dinâmicas para identificar padrões e tendências nos dados. A modelagem foi construída com foco em análises interativas no Power BI, permitindo a exploração iterativa e refinamento das hipóteses levantadas.

Técnicas e suposições de modelagem

Os dados indicaram que algumas profissões têm maior probabilidade de levar à vitória, especialmente em contextos específicos. No entanto, ficou evidente que o sucesso eleitoral é multifatorial, com forte influência de partido, estado e coligação.

Projeto de testes e experimentos

O objetivo dos testes e experimentos é avaliar se os modelos e análises desenvolvidos conseguem responder à questão principal: "Quão a profissão influencia nas chances de um candidato ser eleito prefeito?". Buscamos validar os resultados de forma estatística e prática, considerando a perspectiva de Ciência de Dados e o domínio eleitoral.

Descrição dos modelos

No decorrer do projeto, diferentes abordagens foram utilizadas para compreender a relação entre a profissão dos candidatos a prefeito e suas chances de vitória. Cada modelo foi projetado com o objetivo de explorar padrões, avaliar hipóteses e fornecer insights significativos, alinhados ao objetivo de mineração de dados: determinar o impacto da profissão no sucesso eleitoral.

Os modelos exploraram tanto análises descritivas quanto técnicas quantitativas. Algumas abordagens foram descartadas por não se adequarem ao escopo do trabalho, mas o aprendizado obtido contribuiu para decisões mais assertivas nos modelos finais.

Avaliação dos modelos

Os experimentos realizados buscaram validar as hipóteses levantadas sobre a influência da profissão nas chances de um candidato ser eleito prefeito. A seguir, apresentamos os principais resultados obtidos a partir das abordagens de análise descritiva, regressão logística e clusterização.

5. Avaliação

O objetivo principal deste projeto foi responder à seguinte pergunta: *Quão relevante é a profissão de um candidato para suas chances de se eleger prefeito?* Para isso, buscamos identificar padrões significativos nos dados, utilizando técnicas estatísticas e de modelagem, com o propósito de gerar insights práticos sobre o impacto da profissão no sucesso eleitoral.

Os critérios de sucesso estabelecidos inicialmente incluíram:

1. Identificar profissões com maior probabilidade de vitória em eleições municipais.
2. Avaliar a relevância da profissão em relação a outros fatores, como partido político e região.
3. Comunicar os resultados de forma visual e acessível, com dashboards e gráficos no Power BI.

Avaliação dos resultados do projeto

Os resultados obtidos demonstraram que a profissão é um fator relevante no sucesso eleitoral, mas seu impacto é amplificado por outros fatores contextuais. O Power BI foi fundamental para a visualização e validação dos insights, garantindo uma análise interativa e de alto valor para o objetivo do projeto.

Revisão do processo e conclusões gerais

Apesar das limitações identificadas, o grupo conseguiu atingir plenamente os objetivos relacionados à criação de gráficos no Power BI, que apresentam os padrões de maneira clara e acessível. A pergunta inicial do projeto foi respondida parcialmente, evidenciando que a profissão possui relevância no sucesso eleitoral, mas sempre em interação com outros fatores.

O processo foi uma oportunidade valiosa de aprendizado, tanto em técnicas de modelagem quanto no uso de ferramentas como o Power BI. Para projetos futuros, recomendamos um planejamento mais detalhado e a ampliação da base de dados para incluir variáveis qualitativas e quantitativas adicionais.

6. Autocrítica

O grupo seguiu de forma consistente as etapas propostas pela metodologia CRISP-DM, o que nos ajudou a manter o foco no objetivo do trabalho e a estruturar bem nossas análises. A etapa de compreensão dos dados foi particularmente detalhada, considerando que trabalhamos com múltiplos datasets (candidatos e resultados). Durante a preparação dos dados, enfrentamos desafios com dados ausentes e inconsistências, mas conseguimos superá-los utilizando técnicas adequadas, como imputação e normalização.

Na fase de modelagem, aplicamos diferentes técnicas e iteramos sobre os modelos para garantir que obtivéssemos insights confiáveis. Apesar de alguns modelos não terem gerado resultados satisfatórios, eles contribuíram para o aprendizado técnico do grupo, fortalecendo as decisões para os modelos finais.

- Satisfação com o Progresso

O grupo está satisfeito com o progresso, principalmente pela clareza dos resultados obtidos. Conseguimos responder à pergunta inicial: "Quão relevante é a profissão para o sucesso eleitoral de um prefeito?". Nossos modelos e análises apontaram que a profissão tem, de fato, impacto nas chances de vitória, mas também está interligada com outros fatores, como partido, idade e localidade do candidato.

Entretanto, reconhecemos que algumas etapas poderiam ter sido mais otimizadas. Por exemplo, a fase de integração dos datasets demandou mais tempo do que esperávamos devido a problemas com a padronização dos dados. Além disso, a criação de métricas mais robustas logo no início do projeto teria acelerado algumas análises.

- Lições Aprendidas

Técnicas: Aprendemos a lidar com datasets complexos, realizar pré-processamentos como codificação de variáveis categóricas e interpretar diferentes métricas de desempenho de modelos.

Trabalho em equipe: O trabalho em grupo evidenciou a importância da comunicação clara e da divisão de tarefas. Percebemos que um planejamento inicial mais detalhado poderia ter evitado alguns retrabalhos.

- Autoatribuição de Nota

Atribuímos ao nosso grupo a nota 8,5/10.

Justificativa:

Pontos positivos: O grupo seguiu bem a metodologia, obteve respostas claras para a pergunta inicial e apresentou visualizações que reforçam os insights gerados.

Pontos a melhorar: Algumas etapas tomaram mais tempo do que o planejado, como a limpeza e preparação dos dados, o que reduziu o tempo disponível para explorar técnicas mais avançadas de análise.

- Cumprimento do Escopo

Acreditamos que cumpriremos 100% do escopo proposto. Os modelos já mostram que a profissão tem impacto nas chances de um candidato ser eleito, o que responde à nossa pergunta inicial. Além disso, estamos no caminho para apresentar visualizações claras e insights úteis que comprovam nossas descobertas.